

# Evaluating the Ethical and Social Implications of AI in Public Broadcasting

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieurin**

in

**Business Informatics**

by

**Laura Maria Vigl, BA BSc**

Registration Number 51834603


to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. Mag. Dr. Sabine T. Kőszegi

Assistance: Lara Schmalzer, BSc

Vienna, October 17, 2024



Laura Maria Vigl

Sabine T. Kőszegi



# Evaluierung der ethischen und sozialen Implikationen von KI im öffentlich-rechtlichen Rundfunk

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieurin**

im Rahmen des Studiums

**Wirtschaftsinformatik**

eingereicht von

**Laura Maria Vigl, BA BSc**

Matrikelnummer 51834603

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.-Prof. Mag. Dr. Sabine T. Köszegei

Mitwirkung: Lara Schmalzer, BSc

Wien, 17. Oktober 2024



Laura Maria Vigl

Sabine T. Köszegei



# Erklärung zur Verfassung der Arbeit

Laura Maria Vigl, BA BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 17. Oktober 2024

  
\_\_\_\_\_  
Laura Maria Vigl



# Overview of Generative AI Tools Used

**Grammarly, Inc. (Version: Pro)** <sup>1</sup> was used throughout the thesis to correct grammar and spelling and improve the clarity of the already existing text. The tool did not generate entire sentences but rather suggested better/correct wording of the parts of the already existing sentences.

**AiDitor: OpenAI, ChatGPT 4.0** <sup>2</sup> was used within the *AiDitor* environment sporadically throughout the thesis to correct grammar and spelling and improve the clarity of the already existing text. The tool did not generate entire sentences but rather suggested better/correct wording of the parts of the already existing sentences.

**AiDitor: WhisperTranscribe** <sup>3</sup> was used within the *AiDitor* environment to create an initial draft of the interviews, which then were overseen by the author by listening to the interview recordings and correcting the transcriptions.

---

<sup>1</sup><https://app.grammarly.com/>

<sup>2</sup><https://chatgpt.com/>

<sup>3</sup><https://www.whispertranscribe.com/>





# Kurzfassung

Diese Arbeit untersucht die ethischen und sozialen Implikationen von künstlicher Intelligenz (KI) im öffentlich-rechtlichen Rundfunk und beschreibt eine Teilstudie der diesjährigen *Public-Value-Studie (2024)*. Ziel dieser Untersuchung ist es, auf der Grundlage ausgewählter europäischer Best-Practice-Modelle der digitalen Transformation und deren wissenschaftlicher Evaluierung Perspektiven für die Weiterentwicklung des Grundmodells eines Public Network Value zu entwickeln. Damit soll ein adaptiver Analyserahmen für die digitale Weiterentwicklung öffentlich-rechtlicher Anbieter entworfen und perspektivisch europäische Kooperationsprojekte aufgezeigt werden. Die hier skizzierte Teilstudie befasst sich mit der Evaluierung eines KI-Tools namens *AiDitor*, ein mit vielfältigen Funktionen ausgestattetes Support-Tool für die Redaktionsarbeit. Ziel der Teilstudie ist es, Chancen und Herausforderungen zu bewerten, die durch den Einsatz dieses KI-Tools im redaktionellen Bereich entstehen. Dabei sollen die Anforderungen sowie die professionellen und berufsethischen Standards eines öffentlich-rechtlichen Rundfunks berücksichtigt werden.

Anhand einer Fallstudie mit qualitativen Interviews wurden unterschiedliche Perspektiven auf die Rolle von KI im öffentlich-rechtlichen Rundfunk erhoben. Darüber hinaus wird der *Z-Inspection* Prozess, ein IEEE-Standard, auf die Domäne des öffentlich-rechtlichen Rundfunks angewendet, einschließlich der Verwendung des von der Europäischen Union entwickelten sozio-technologischen Analyserahmens für vertrauenswürdige KI. Die Ergebnisse zeigen erhebliche Spannungen zwischen den Möglichkeiten von KI sowie sozialen und ethischen Erwägungen, was den Bedarf an robusten Governance-Rahmenwerken und Leitlinien unterstreicht. Auf der Grundlage einer Evaluierung der Studienergebnisse durch Experten werden Empfehlungen skizziert und verschiedene Handlungsfelder identifiziert, die den vertrauenswürdigen Einsatz von KI anleiten sollen.



# Abstract

This thesis examines the ethical and social implications of artificial intelligence (AI) in public broadcasting and describes a sub-study of this year's *Public Value Study (2024)*. The *Public Value Study* aims to develop perspectives for further development of the basic model of the Public Network Value based on selected European best practice models of digital transformation and their scientific evaluation. This is intended to design an adaptive analytical framework for the digital advancement of public service providers and to prospectively highlight European cooperation projects. The sub-study outlined here deals with the evaluation of an AI tool called *AiDitor*, a support tool for editorial work. The sub-study aims to assess the opportunities and challenges that arise from the use of this AI tool in the editorial field. The requirements of such an AI support tool as well as the professional and ethical standards of a public broadcaster will be considered.

A case study approach, including qualitative interviews, was employed to gather diverse perspectives on AI's role in public broadcasting. In addition, the *Z-Inspection* process published by IEEE is tailored to the domain of broadcasting, including the usage of a socio-technological analysis framework such as the framework for trustworthy AI developed by the European Union. The findings reveal significant tensions between technological advancements as well as social and ethical considerations, emphasizing the need for robust governance frameworks and guidelines. Based on a peer assessment conducted by the study author and experts in the field of AI, recommendations are outlined and different fields of action are identified aiming to guide trustworthy employment.



# Contents

<b>Overview of Generative AI Tools Used</b>	<b>vii</b>
<b>Kurzfassung</b>	<b>ix</b>
<b>Abstract</b>	<b>xi</b>
<b>Contents</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Statement . . . . .	4
1.3 Research Questions . . . . .	4
1.4 Aim of Work . . . . .	4
1.5 Structure of Thesis . . . . .	5
<b>2 Literature Review</b>	<b>7</b>
2.1 Terminology . . . . .	7
2.2 AI in Public Broadcasting: Key Areas and Importance . . . . .	14
2.3 Challenges and Implications of AI in Public Broadcasting . . . . .	16
2.4 Media Ethics . . . . .	18
<b>3 Evaluation Framework</b>	<b>21</b>
3.1 EU Framework for Trustworthy AI . . . . .	21
3.2 Discussion of the EU Framework . . . . .	24
3.3 The Z-inspection Process . . . . .	25
3.4 Reflections on the Z-inspection Process . . . . .	27
<b>4 Methodology</b>	<b>29</b>
<b>5 Research Design</b>	<b>33</b>
5.1 Set-Up Phase . . . . .	34
5.2 Assessment Phase . . . . .	37
<b>6 Results</b>	<b>45</b>
	<b>xiii</b>

6.1 Experts	45
6.2 Users	51
6.3 Stakeholders	55
6.4 Results with respect to the research question	60
6.5 Tensions between the ALTAI Requirements	63
<b>7 Evaluation and Recommendation</b>	<b>67</b>
7.1 Evaluation	67
7.2 ALTAI Recommendations	70
7.3 Fields of Action	75
<b>8 Discussion</b>	<b>79</b>
8.1 Limitations	79
8.2 Reflection on the Assessment Method	80
8.3 Conclusion	80
8.4 Future Work	81
<b>List of Figures</b>	<b>82</b>
<b>List of Tables</b>	<b>83</b>
<b>Bibliography</b>	<b>85</b>
<b>A Interview Questions</b>	<b>93</b>
A.1 Stakeholder	93
A.2 Experten	95
A.3 Users	98

# CHAPTER 1

## Introduction

### 1.1 Motivation

Public media companies are perceived as the fourth pillar of democracy due to their legal duty to provide education and information on political, social, economic, cultural and sports-related issues [Wie01]. For this reason, public media companies are driven by the demand for quality underlined by the ongoing quality discourse of journalistic content [PH10]. The ongoing discourse regarding quality expectations stems from the demands and critiques of society motivated by different scandals [Neu19]. Scandals such as the *Cambridge Analytica* scandal showed media manipulation on an entire nation, highlighting the importance of reliable information [HWJ20].

Based on the demand for quality and the special legal mandate of public broadcasting companies, opportunities and challenges arise from the latest technical innovations in Artificial Intelligence (AI). Driven by advantages such as efficiency [Can24], the public broadcasting company in Austria (ORF) developed an AI tool, the *AiDitor* for daily editorial routines. The *AiDitor* is a prototype where editorial teams can create and store customized prompts in their workspace, allowing them to produce different products from researched and verified content. Different technologies such as Amazon Web Services, OpenAI or Microsoft Azure define the core of the *AiDitor* and enable functions such as the translation of text, chat or the generation of various media content.

Unfortunately, AI's promising opportunities are accompanied by criticisms of its application and value in public broadcasting, which will be addressed in this year's (2024) *Public Value Study* by the ORF. This master thesis will be part of the *Public Value Study* by evaluating the implications arising from the usage of AI tools within the ethical and social standards in the given domain. The *Public Value Study* deals with democratic values, therefore discussing the activities and duties of the ORF. The public value report is released annually and is structured across five quality dimensions concerning the individual, societal, national, international and corporate value [KM23].

When dealing with AI, stumbling across challenges such as the struggle for truth, the creation of synthetic realities or the labeling of AI-based journalism is inevitable. The struggle for truth is underlined by the *World Economic Forum* by rating the threat of misinformation and disinformation as global risk number one as shown in Figure 1.1 [For21]. Limited explainability, often mentioned in the context of generative AI (genAI), is another concern. It refers to the ability to explain how an AI made certain decisions by tracing back an AI's decision-making process. Tracing back how certain decisions are made is crucial for accountability whenever decisions are made by the AI system independently of any domain. A further problem in the context of genAI is the so-called *hallucinations* which describe the problem of AI generating wrong answers in terms of facts and presenting them in full confidence [Can24].

FIGURE 1.3 Global risks ranked by severity over the short term (2 years)

\*Please estimate the likely impact (severity) of the following risks over a 2-year period.\*

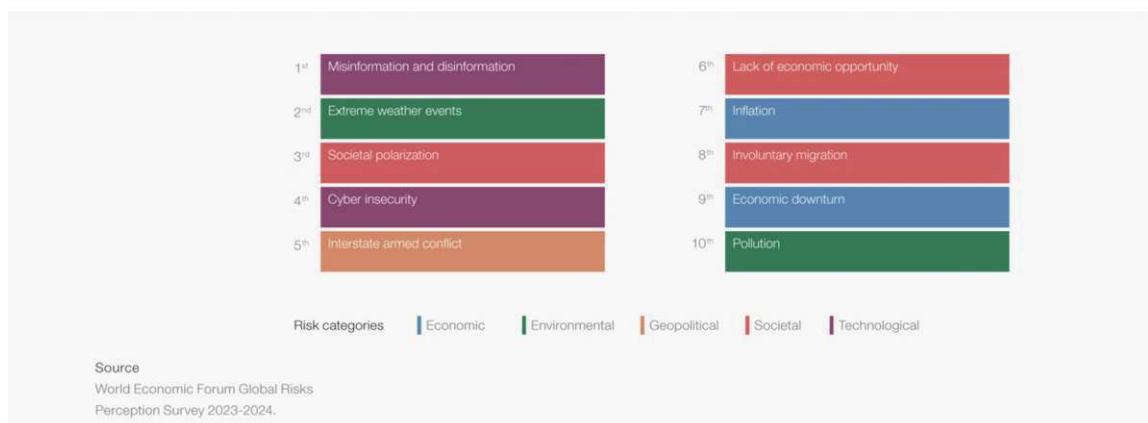


Figure 1.1: Classification of global risks rated by the *World Economy Forum* [For21].

These challenges are addressed by Leaver et al. in the paper *ChatGPT Isn't Magic* by discussing the hype referred to as "panic" [LS23] by the authors. The paper mentions open letters calling for a pause in AI development, such as the one by the institute *Future for Life* warning about the recent AI development in terms of describing AI as:

"...an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control" [LS23].

Besides, the hype of AI is accompanied by beneficial opportunities such as recognizing misinformation using AI tools or efficiently serving multidisciplinary markets such as the social media market by producing AI-generated content [Chr21].

A systematic study by Stahl and Eke compares the ethical challenges of genAI, using the Large Language Models (LLMs) developed by *OpenAI* as an example, to assess its opportunities and benefits [SE24]. The well-substantiated and comprehensive analysis utilizes



established methods of technology impact assessment, particularly for new technologies such as genAI. Figure 1.2 provides an overview of the ethical challenges identified by Stahl and Eke, where they assess that the negative consequences of genAI significantly overshadow the potential positive impacts.

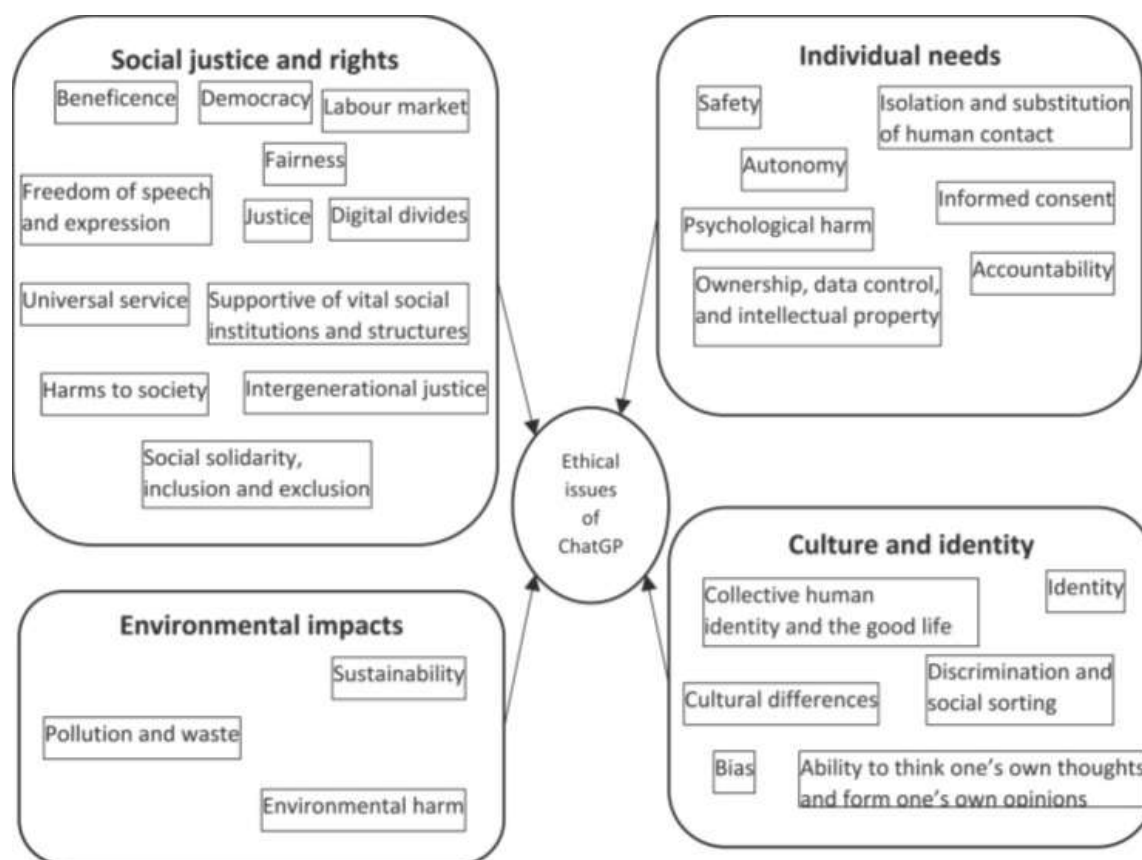


Figure 1.2: Ethical challenges with significant negative impact through *ChatGPT* [SE24].

This analysis, visualized in Figure 1.2 emphasizes that focusing on individual needs and rights is not sufficient to identify the risks and dangers of using genAI within a society and to take appropriate measures to mitigate the risks. Many of these challenges are directly linked to the public mandate of the ORF and should therefore be included in the awareness of its usage. On the one hand, the educational mandate becomes particularly relevant in this case because it should not only report on opportunities but also these societal challenges and dangers in an appropriate, diverse and accessible format for all the different experts and stakeholders. On the other hand, the ORF as a user of genAI assumes a role model function that needs to be fulfilled in a correspondingly responsible manner.

### 1.2 Problem Statement

The thesis aims to evaluate the opportunities and challenges arising from using genAI tools such as the *AiDitor* in the newsroom. The evaluation will take place against the background of the ethical and social standards of public service broadcasting (public value) and will provide guidelines for implementing the *AiDitor*. Hence, the thesis aims to improve the understanding of implications arising from using AI in the context of public broadcasting companies by assessing the opportunities and challenges concerning ethical and social consequences. The challenge lies in conducting a socio-technical analysis to understand the implications of technology within its context by involving different stakeholders.

### 1.3 Research Questions

The following research questions will be addressed in the master thesis:

- What are the ethical and social implications of using genAI tools within the newsroom of a public service broadcaster from a public value perspective?
- What are the opportunities and challenges of using AI tools in public service broadcasting from the perspectives of different stakeholders?
- How can the responsible integration and usage of genAI in the context of public service broadcasting be ensured?

### 1.4 Aim of Work

To address the problem and the research questions, this thesis will evaluate the ethical and social implications of using the *AiDitor* as an AI tool, focusing on the ethical values relevant to public service broadcasting. Hence, the literature review provides an assessment of ethical values in the context of public broadcasting. Moreover, the literature review covers relevant theoretical backgrounds, such as the current state-of-the-art regarding AI technology.

Furthermore, the thesis describes potential areas of usage such as research, production and distribution of journalistic content, aiming to highlight the importance and possibilities of AI technology. Moreover, the challenges and implications of AI are assessed within the given context. Different study participants will be identified, including experts and stakeholders, such as executives, users and developers of the *AiDitor*.

The development of the interview questions for the qualitative assessment aims to capture various perspectives from public broadcasting stakeholders regarding the employment of AI. This includes understanding their fears and expectations using a socio-technological assessment method. Moreover, recommendations on how trustworthy employment of AI technology can be ensured in public broadcasting companies are included.

The in-depth interviews will be analyzed to provide a comprehensive understanding of how AI usage affects the given context, considering the interests of the stakeholders. The aim is to gain new insights and develop theories. The results will be evaluated using the self-assessment tool of trustworthy AI developed by the European Union (EU). Moreover, the thesis includes an assessment of the potential limitations of the research and future work.

## 1.5 Structure of Thesis

This thesis is structured into eight chapters to comprehensively address the research questions and objectives:

Chapter **1** Introduction, describes the motivation for the study, the problem statement and the aim of the work. It introduces the key research questions that guide the investigation into AI's ethical and social implications in public broadcasting.

Chapter **2** Literature Review, provides the theoretical foundation for the thesis. It includes definitions and discussions of key terms such as AI and LLMs. This chapter also reviews the current state of AI in public broadcasting, exploring both the opportunities and challenges it presents. Additionally, it examines ethical considerations in media and AI, setting the foundation for further analysis.

Chapter **3** Evaluation Framework, describes the evaluation frameworks relevant to the study. Therefore comprehensively describes the EU framework for trustworthy AI and the *Z-inspection* process published by IEEE. Moreover, the Chapter discusses different strengths and critiques of the EU framework for trustworthy AI as well as the *Z-inspection* process.

Chapter **4** Methodology, outlines the research design and methods used to conduct the study. It describes the setup and assessment phases and details regarding the selection of study participants, the interview processes and the data analysis techniques.

Chapter **5** Research Design, provides a detailed description of the research design by tailoring different methods and approaches to the thesis research. It includes details on the setup phase, such as identifying relevant stakeholders and the development of the interview questions used within the study. The assessment phase focuses on how the collected data will be analyzed and interpreted.

Chapter **6** Results, presents and illustrates the findings from the research. It is divided into sections based on different groups of participants: experts, users and stakeholders. The results are discussed concerning the first research questions, highlighting key insights and patterns observed during the study.

Chapter **7** Evaluation and Recommendation, critically assesses the findings using state-of-the-art socio-technological assessment frameworks. It provides recommendations based on trustworthy AI and identifies fields of action for public broadcasting companies. This

chapter aims to offer practical solutions for integrating AI responsibly within the domain of public broadcasting.

The final Chapter [8](#) Discussion, outlines the implications of the findings. It addresses the study's limitations, reflects on the assessment methods used and suggests directions for future research. This chapter ensures a comprehensive understanding of the study's contributions and potential areas for further investigation.

# CHAPTER 2

## Literature Review

This chapter provides a comprehensive literature review on the role of AI in public broadcasting, focusing on the ethical and social implications. The chapter starts by defining key terminologies such as AI, LLMs, genAI, trustworthy AI and AI bias. It then explores how AI is transforming public broadcasting, highlighting key areas for instance research, production and content distribution. Challenges and ethical concerns, such as misinformation, bias and media ethics are discussed in-depth, setting the stage for a critical evaluation of AI's impact on democratic values and ethical journalism.

### 2.1 Terminology

#### 2.1.1 Artificial Intelligence

Providing a single definition for AI is challenging because the term serves as an umbrella term describing a wide collection of concepts, approaches and technologies. However, this section gives an overview of how AI was defined historically, including an investigation of its evolving interpretations in recent discourse.

AI is a combination of the term artificial, referring to an object or behavior that is not natural and therefore imitated by chemical or technical means [EB16]. Intelligence is the ability to learn, understand, make judgments or have opinions [Dica]. Derived from those two terminologies, artificial intelligence can be described as an unnatural object or behavior that can act and make decisions. However, the term depends on its context and the interpretation [EB16]. John McCarthy, described as one of the fathers of AI, defines AI as

"the science and engineering of making intelligent machines, especially intelligent computer programs" [MMRS06].

Around the same time in 1987, a similar definition, according to Haugeland, describes AI as

"the exciting new effort to make computers think" to create "machines with minds, in the full and literal sense" [Hau89].

Even though these two definitions give an idea of what artificial intelligence means, they do not give any further description of what is meant by *minds* or a *thinking computer*.

A more exhaustive definition in the *Encyclopedia Britannica*:

"AI is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" [Cop24].

In this definition, intelligent beings refer to human's higher intellectual processing capabilities. According to this definition, Ertel argues that all computers are AI systems because they can perform tasks typically associated with intelligent beings, such as calculations and text memorization. Other definitions describe AI as a machine that thinks and acts like a human, such as Kurzweil emphasizing the definition of AI as

"art of creating machines that perform functions that require intelligence when performed by people" [R90].

Aligned with this perspective is the definition of Tanimoto, who emphasizes that AI is

"a field of study that encompasses computational techniques for performing tasks that require intelligence when performed by humans" [Tan87].

The idea of AI has mostly stayed the same over the past few decades and comes down to what can be derived from the term AI itself. Hence, the thinking is done by objects such as computer programs or machines which are not natural. Attempting to imitate human behavior, which becomes evident by a more novel definition of Wang,

"AI systems mimic human intelligence by using available resources to solve problems through learning and adapting to environmental and contextual factors" [Wan19].

All these definitions share one thing: they all mention *intelligence* in some way, but the term is often not clearly explained. According to Ertel, this would imply an attempt to understand the human mind, which proves to be fascinatingly complex [EB16]. For this study, the definition of Wang will be used aiming to describe those environmental and contextual factors used within the problem-solving process.

### 2.1.2 Large Language Models

*ChatGPT* became incredibly popular by the end of 2022 for what reason Teubner et al. are describing the hype as "cultural sensation" [TFW<sup>+</sup>23]. The excitement reflects the ongoing progress in AI-powered developments such as chatbots and different language services for translation or content creation [O'L23]. Neural Language Processing (NLP) is part of AI and computational linguistics. Hence, the algorithms and concepts developed within the field of NLP build the foundation for LLMs [Ama23]. For this reason, this section will first introduce NLP, followed by a definition and description of Language Models (LMs) and LLMs.

The field of NLP incorporates the ability of machines to process, understand and generate natural language and therefore, can be described as one of the oldest sub-fields of AI. Initially introduced in the 1950s by Alan Turing in his paper *Computing Machinery and Intelligence* which deals with determining the characteristics of a machine having a human-like intelligence by the so-called *Turing Test*. The experiment involves humans, the evaluators, interacting with both a machine and a human without knowing which is which, aiming to determine the identities based on the responses. If the machine can convince the evaluators to be human, it successfully passes the *Turing Test*.

The experiment characterizes and measures different goals within the field of AI research and defines human-like machines as capable of understanding and generating natural language [Ama23]. The *Turing Test* serves as groundwork for NLP, which later developed into various research fields such as Natural Language Understanding, Natural Language Generation and Dialogue Management, which handles both input and output interactions [BGMMS21].

Furthermore, the research field of NLP is challenged with different tasks to close the gap between human and machine-based language. One challenge is the categorization of text called *sentiment analysis*, which involves assigning a label or category to a piece of text, such as categorizing emails as spam or identifying if the sentiment is positive, negative or neutral. Another example is the recognition of entities such as names, people or organizations within a given text. This field includes tasks like text generation, chatbots answering questions or summarizing text. Another notable task would be speech recognition by converting spoken language into written text [Ama23].

To fully understand how LLMs work, basic concepts of NLP such as tokenization, word embeddings, corpus and vocabulary need to be introduced. Tokenization is breaking down text into small units such as words or sub-words. Splitting text into small word chunks, the so-called tokens, is a crucial pre-processing step in NLP. The process results in a list containing all the tokens where all the text's punctuation is removed. Depending on the NLP tasks, different splitting methods can be used. A standard method for handling most English text and tasks is white space tokenization, where the text is split based on space, tabs or new lines. Another method is to split the text based on punctuation marks, such as commas, periods or exclamation marks [Ama23].

To increase the semantic meaning of a text and create a more expansive context window,



word embeddings are used in the field of NLP [MA24]. Word embeddings are vector representations of words that capture the semantic relationships between those words, aiming to understand the meaning of a specific word within a given context. Different word embedding methods, such as *Word2Vec*, enable significant performance improvement and established embeddings as a fundamental practice in NLP [Ama23].

The concept of corpus relates to an extensive collection of linguistic material in text documents or audio. It is used as the primary data source when performing different NLP tasks. A parallel corpus of texts in multiple languages is used for cross-lingual tasks and machine translation. Another example of a corpus would be the *Treebanks*, which are annotated with the sentence's grammatical structure and used in parsing and syntax-based machine learning. Multimodal corpora include text aligned with other linguistic material such as images, videos or audio used to generate information based on multiple modalities [Ama23].

The vocabulary describes a set of unique words or tokens within a text corpus and represents the set of words a system or model can use and understand. When processing text data, the first step usually involves tokenization, where the text is split into individual tokens. Followed by filtering and normalization, including tasks such as converting text to lowercase and removing punctuation or stop words. After these steps, the vocabulary is built by assigning a unique numerical value to each token. These numerical values are then used to create dense vectors, known as word embeddings. This enables processing textual as numerical data, which comes in handy in machine learning models, especially when processing large amounts of data [Ama23].

LMs were initially introduced in the early 1980s and are also known as Statistical Language Models (SLMs). SLMs aim to predict the likelihood of a token, in other words, the LM predicts the next meaningful character, word or string by considering the given context [BGMMS21]. A well-known example would be text prediction, also known as auto-completion, used in search engines, where the system predicts the next word or phrase based on the user's input and the ongoing typing activity. When, for instance, a user is typing into the Google search bar, the LM constantly predicts the following possible words and presents them as suggestions [MA24].

LLMs belong to the class of Neural Language Models, also advanced language models, because they learn statistical patterns and relationships between words in a large corpus of text using a neural network [Ama23]. The term *large* refers to the number of values or parameters the model can adjust during the training phase, reaching up to billions of parameters. Parameters are complex representations of the data, increasing the computational requirements within the training phase and therefore, representing the model's capacity to learn and express information.

Most of today's LLMs are Generative Pre-trained Transformers (GPT), the basic idea behind a GPT is the pre-training on large amounts of data before fine-tuning the model for specific tasks. Within the pre-training phase, the model gains a general understanding of the language, which then can be applied to specific tasks [Ama23].



With the limitation of only knowing up to the time the model is trained on [BGMMS21].

For instance, *GPT-3*, a conversational AI developed by OpenAI (2022), aims to generate text-based content and was trained on 175 billion parameters. Another example is the model *BERT* (2019), which aims to extract information and classify text. The model *T5* (2020) developed by Google is specialized in summarizing text and generating content. Additionally, essential factors such as the size of the training as well as the quality and diversity of the data influence the model's characteristics and performance. One big difference between these new Neural Language Models and the SLMs is how they handle words. Instead of just looking at single words, Neural Language Models use the concept of word embeddings, which creates a broader context window compared to SLMs [MA24].

### 2.1.3 Generative AI

The term generative AI describes a field within AI research concerned with generating new data, such as textual data, transformations of text, responses to prompts and translations using the capabilities of LLMs. The field also grew in developing models that can generate images, videos and music [MA24]. Therefore, the fields of application are extensive, comprising domains such as entertainment, healthcare, public administration or academia. [WPLK23] [SNLP22] [RGR<sup>+</sup>23].

### 2.1.4 Trustworthy AI

According to Lu et al., concerns about the impact of computers on society emerged almost simultaneously with the development of the first computers. When Tommy Flowers invented the first programmable computer, *Colossus*, in 1943, Norbert Wiener outlined new fields of academic research, now known as computer ethics. Within the field of computer ethics and the advances in the field of AI, terms such as responsible AI, ethical AI, trustworthy AI, AI for Good, Value-Driven AI and Digital Humanism all aim to create AI systems to benefit society [LZWX24]. Several governmental and private organizations have put effort into defining how the development and deployment of AI shall be done to ensure trust in AI systems and the benefit to society. Mariani et al. emphasize addressing risks associated with AI in terms of defining principles derived from human rights, ethical norms and legal properties [MRC<sup>+</sup>23].

An overview of different initiatives, described in Table 2.1, aims to determine recommendations, standards, principles and policies to support the development, deployment and usage of AI systems [BG21]. The initiative *The Partnership on AI* is a cooperation of six companies, Apple, Amazon, Google, Facebook, IBM and Microsoft, aiming to study and formulate best practices on AI technology. Moreover, the initiative states to be an open platform to discuss and engage on the influences of AI on people and society [Bin16] [Bel20].

Developed by the World Commission on the Ethics of Scientific Knowledge and Technology of UNESCO, *The UNESCO Initiative* was launched to study the ethics of AI, which resulted in the publication of a comprehensive study on the ethics of AI. The group

## 2. LITERATURE REVIEW

Initiative/Expert Group	Launched	Objectives
The Partnership on AI	2016	Study and Formulation of Best AI Practices
UNESCO Initiative	2017	Study AI Ethics Development of Principles/Recommendations
Future of Life	2017	Development of Principles
HLEG	2018	Development of Principles Assessment of AI Impact
IEEE EAD	2019	Development of Standard
OECD, AIGO	2022	AI Governance Assessment of AI Impact Development of Principles/Recommendations

Table 2.1: Summary of AI initiatives and their objectives.

consists of 24 experts from 24 different countries and backgrounds. The still ongoing study discusses issues related to AI, such as education, culture, science and peace, as well as the development of AI in less-favored countries. This resulted in eleven principles, including recommendations, such as human rights, inclusiveness, democracy and sustainability.

The European Commission originated 2018 the High-Level Expert Group (HLEG) on AI to develop a framework for trustworthy AI [CDGfCNT20]. The HLEG includes 52 multi-disciplinary experts emphasizing that trustworthy AI is lawful, ethical and robust. The framework is divided into three chapters. The first chapter describes the foundations of trustworthy AI by laying out four ethical principles: respect for human autonomy, prevention of harm, fairness and explicability of trustworthy AI. Chapter two addresses the realization of trustworthy AI by translating the ethical principles into key requirements and describing technical and non-technical methods that can be used during the implementation process.

The following seven requirements are defined by the HLEG in chapter two:

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-discrimination and Fairness

## 6. Environmental and Societal Well-Being

## 7. Accountability

The third chapter deals with assessing a specific AI application and therefore, consists of a semi-automated questionnaire, the ALTAI assessment list for trustworthy AI operationalizing the seven key requirements of trustworthy AI [Komb]. The questionnaire results are visualized with a spider diagram including specific recommendations [RGR<sup>+</sup>23].

Another noteworthy initiative is the Working Party on AI Governance (AIGO), introduced by the Economic Cooperation and Development (OECD). For overseeing and steering purposes regarding AI governance, a working party member is nominated by the OECD members primarily responsible for AI policies in their specific country. Hence, the different members of AIGO analyze the design, implementation, monitoring and evaluation of AI policies and action plans for different countries. For instance, various trends, such as investments in AI or demographics of AI professionals, are published per country on their website. Moreover, AIGO offers a catalog of tools and metrics to assess trustworthy AI, which helps AI actors build and deploy trustworthy AI systems. AIGO defines critical principles such as accountability and human rights, values and fairness, which can be compared to the principles defined by the HLEG. To guide the government's implementation of the AI principle, AIGO offers recommendations for AI ecosystems to benefit society [AIG24].

Initiatives such as the global initiative Ethically Aligned Design (EAD) founded by IEEE are creating standards to ensure the development and deployment of trustworthy autonomous and intelligent systems by respecting different values of ethics, philosophy and politics [Des16]. The *Future of Life*, a volunteer-led research and outreach organization, hosted a conference *Asilomar* in 2017 to discuss and formulate principles for useful AI and the mitigation of existential risks posed to humanity by AI. As a result, they developed principles such as safety, transparency, privacy and liberty [oL17].

The study will employ the framework of trustworthy AI of the HLEG, which will be described in more detail in chapter 3 Evaluation Framework.

### 2.1.5 Bias in AI

The term bias was initially used in social and natural sciences [ZK22]. With the application of AI across various domains, the term has also become common in computer science. However, a consensus in the literature on its precise definition still needs to be solved [GTC<sup>+</sup>23]. According to Ferrara, bias can be defined as a systematic error within a decision-making process that leads to unfair outcomes. Different sources, such as data collection, algorithm design and human interpretation, can cause a bias. In AI and machine learning, models trained on data containing different biases can reflect these biases, resulting in unfair or even discriminatory outcomes [Fer23]. Zhai et al. define three critical properties of bias shared across various literature domains: first, the deviation between an observation and the reality also referred to as *ground truth* in the literature.

Second, the authors highlight that a bias always occurs in a systematic rather than a random way. The final critical property is a tendency towards specific ideas or entities [ZK22].

The literature identifies various types of biases by different sources, such as data, algorithmic and user bias. Table 2.2 provides an overview and brief description of various AI biases. A sampling bias is a type of data bias caused by unrepresentative or incomplete data used to train machine learning models. Unrepresentative data is defined as data sources that already contain biased information. Incomplete data refers to data missing important information or containing errors. Algorithmic bias is caused by algorithms using biased assumptions or criteria to make decisions. A user bias such as, in Table 2.2 named interaction bias, is caused by people using AI systems, consciously or unconsciously reinforcing their own bias [Fer23].

Biases can raise challenges, such as the negative impact on individuals and society by enforcing inequalities and hence discrimination of underrepresented groups or individuals [Fer23]. A notable real-world example would be the COMPAS system used in the United States to support the criminal justice system, predicting the likelihood of a defendant re-offending. Where a study by ProPublica found that the system shows a bias against African Americans by predicting a higher likelihood, a high risk, of re-offending even if they had no prior convictions [ALMK22].

With the rise of genAI, harmful biases have emerged in text-to-image models such as StableDiffusion, OpenAI's DALL-E and Midjourney, which have produced stereotypical and racially biased outputs. For example, generated images of Chief Executive Officers (CEOs) primarily feature men, illustrating gender bias due to the underrepresentation of women in CEO positions. Similarly, images of criminals and terrorists generated by these models primarily pictured people of color. These examples highlight the importance of mitigation strategies, such as oversampling and selecting classifiers and methods based on group or individual fairness [Fer23].

### 2.2 AI in Public Broadcasting: Key Areas and Importance

The digital transformation within the public broadcasting industry can be described as transitioning from traditional broadcasting to modern public service platforms. Requirements such as the personalization of journalistic content, extensive data research and the production of texts written by AI are becoming increasingly important to remain competitive. Although AI has been used in traditional industries for quite some time, its usage in media companies has been considered critical. Due to the unique mandate, legal restrictions and other factors such as lack of financial resources, the adoption of AI in traditional media companies has been slow compared to other industries. However, given the advantages of AI, its innovation potential and significance must be analyzed and evaluated in detail [Chr21, Kre21].

Main areas of application are identified by Reinhard Christl: Production and distribution

Type of Bias	Description
Sampling Bias	Unrepresentative data skew model results
Algorithmic Bias	Algorithms reinforce existing prejudices
Representation Bias	Inadequate data diversity causes misrepresentation
Confirmation Bias	Model reinforces pre-existing beliefs
Measurement Bias	Flawed metrics distort data interpretation
Interaction Bias	User behavior influenced by AI responses
Generative Bias	AI generates biased outputs from training data

Table 2.2: Characterization of different types of AI bias.

of journalistic content [Chr21]. In research, AI offers a structured approach to analyze large amounts of data and therefore can be used to summarize text or search for specific keywords in large amounts of data. A notable example is the revelation of the Panama Papers, where the investigative research into the tax scandal involved analyzing three million documents. According to Süddeutsche Zeitung, an entire department was employed for analyzing, reporting and researching this data [Kre21]. Moreover, AI can be employed as supporting technology for journalistic routine tasks, such as transcribing interviews, talk shows or expert panels. [Chr21].

Furthermore, the ability of AI to self-learn and recognize patterns can lead to an entirely new approach of producing journalistic content. For instance, when creating articles, genAI can adapt to different writing styles and create creative content serving multiple platforms and audiences [Chr21]. The distribution of journalistic content involves using various recommendation systems to suggest personalized and regionalized content, enhancing the user experience. Additionally, AI can be used in various online media as an auto-moderation tool to pre-select or label comments for their suitability for publication. It can automatically analyze content and detect false information, improving the quality of content and enabling quick responses to potentially harmful content. Furthermore, the diversity of published contributions can be analyzed to ensure that a wide range of topics, opinions and perspectives are represented. This is particularly important in public broadcasting, as it supports the promotion of democratic discourse [Chr21].

## 2.3 Challenges and Implications of AI in Public Broadcasting

Even though the fields of application are extensive and essential, various scientists and experts in AI bring the challenges and risks associated with the technology to attention. For this reason, an open letter signed by, for instance, Tesla CEO Elon Musk or Apple co-founder Steve Wozniak, emphasizes a pause of AI developments to give AI companies and regulating institutes more time to define safeguards [Cla23]. Since the launch of *ChatGPT*, the hype has been accompanied by, according to Leaver et al., "fascination and panic" [LS23]. The author outlines statements from global economic analysts that the transformative effects of genAI will save labor costs and increase productivity. The downside mentioned by the author is the potential automation of up to 300 million jobs, leading to a remarkable disruption of the labor market. Moreover, the content being human-like is a great advantage for what reason communication barriers between humans and machines will break and significantly impact macroeconomic factors [LS23]. The change that AI will cause in various domains is undeniable. Thus, this section will investigate the challenges and implications of AI in the context of public broadcasting.

The advances of generative pre-trained transformers to produce human-like textual, visual and audio content are already used by leading media companies such as the Washington Post. Because the produced content is almost not distinguishable from human content, Longoni et al. investigated the influence AI has on the credibility of news. The study focuses on the perception of news accuracy, specifically to what extent news written by AI is accepted as true by humans, by performing experiments with 3,000 participants. The experiment included presenting and rating news items, specifically headlines tagged as written by AI, written by a human or both. Results showed that headlines written by AI are perceived as more inaccurate than headlines written by humans. People tend to rate headlines written by AI, even though they were true, as inaccurate [LFCP22].

Another study conducted a socio-ethical analysis by exploring the question of up to what extent humans can tell the difference between AI and human-created content. Using a playful approach, a game in which 2,590 participants rated content as either AI-generated or human-created. With an average score of 5 out of 10 correct answers, the experiment underscores the idea that people struggle to determine the difference. Humans have difficulty distinguishing between AI-generated content and content created by humans. The continuously advancing capabilities of AI raise concerns about the creation and distribution of deepfakes, particularly within the context of broadcasting companies. Scandals such as the *Cambridge Analytica* scandal already proved the potential harm to democratic societies by deepfakes [PSL20]. Furthermore, Ferrara highlights the danger of creating synthetic personas, where individuals with malicious intentions utilize AI technology to fabricate unreal identities, including animating these fake personas [Fer24b]. For this and many other reasons, OpenAI decided to hold back with the release of the *GPT-2* model because the company was concerned about malicious usage of the technology, such as the generation of misleading news articles, the creation of synthetic



personalities or impersonation of others, the creation of abusive or fake content [RWA<sup>+</sup>19].

Given that three billion people are expected to participate in political elections worldwide over the next two years, the World Economic Forum rates misinformation and disinformation as the number one global risk [For21]. Marcellino et al. describe the problem of misinformation and media manipulation and analyze how China is likely to employ AI technology to shape national and international conversations and opinions about China [MBMK<sup>+</sup>23]. On the other hand, AI can be employed as a tool for automated fact-checking to detect misinformation and mitigate the risk of manipulation. Choi et al. introduce *FACT-GPT*, a fine-tuned LLM that automatically checks facts and assesses the truthfulness of claims in social media content. The paper evaluates the ability of different LLMs to judge the textual relationship between social media posts and verified claims, demonstrating that LLMs can reliably assess these relationships with performance comparable to that of humans [CF24].

As mentioned in Section 2.1.5, different types of bias potentially threaten democratic values. LLMs containing bias can reinforce stereotypes and discriminate against various groups by reflecting the imperfections of our society [Fer24b]. Sun et al. explore the implications of gender biases in image generation by AI. The authors emphasize the thought that gender bias in media can harm women by either negatively affecting their self-perception or their cognitive and educational achievements. The author mentions a review of 33 experiments that found negative media content harms people's learning and thinking abilities in stereotyped groups.

In contrast, people not in these groups were either unaffected or even benefited from the biased content. Additionally, experiments show that TV commercials with gender stereotypes reduce women's performance in math and negatively influence their career choices [SWS<sup>+</sup>24]. Another notable phenomenon is the *butterfly effect*, a concept derived from the chaos theory, describing how small changes lead to remarkable and unpredictable outcomes in complex systems. The paper outlines that minor changes in the input data during the training process or algorithm parameters can disproportionately impact minorities and reinforce existing social inequities. Such as feedback loops and reinforcement learning, which describe the problem of small initial biases leading to significantly biased outputs by every learning iteration of the system [Fer24a].

Another notable challenge in the context of bias is the phenomenon of *hallucinating* genAI, where LLMs create realistic-sounding content in complete confidence even though the information is untrue or misleading. Based on hallucination and the spread of misinformation on the internet, Bali outlines a concerning long-term problem that future models might get trained on those inaccurate texts and images generated by AI. For example, the author mentions the website Stack Overflow, a question-and-answer forum developers use. Some Stack Overflow users created bots that aim to answer users' questions with the help of LLMs automatically. Even though some of the answers were of high quality, others were completely incorrect; for this reason, Stack Overflow updated its policy and prohibited the employment of LLMs to prevent users from struggling to distinguish between valuable and untrue information. The author mentions the Stack

Overflow problem to emphasize the danger, especially in research where various journals and funding agencies may end up with low-quality *junk science* created by LLMs [Bai24].

Furthermore, the implications on the labor market are, according to Eloundou et al., undeniable in terms of the general-purpose potential of LLMs and the effect on workers and their activities within the U.S. economy. The study revealed that considering current model capabilities, 80% of the U.S. workforce will have an impact by at least 10%. In addition, approximately 19% of jobs and activities will be affected by at least 50% with the introduction of LLMs [EMMR23]. Cantens explores the implications of genAI in public administrations and emphasizes the thought of a *risk of enslavement* and a high need to train humans with critical, creative and unconventional thinking skills. They are calling it one of the most exciting challenges of genAI transforming activities from pure execution to primarily monitoring and verification [Can24].

### 2.4 Media Ethics

The term ethical, defined by the Cambridge Dictionary, is described as morally correct or relating to what is right or wrong [Dicb]. Meanwhile, Ribino et al. describe ethics as a concern for the moral status of different entities, where morality is defined as the intentions, decisions and actions regarding what is right or wrong [RL<sup>+</sup>19]. Therefore, media ethics concerns what is right or wrong within communication and journalism. Gordon et al. take it one step further and describe media ethics as an "essential process" [GKM<sup>+</sup>12] constantly evolving within the media world. They highlight that there is no single definition of right or wrong, rather a gray area including dilemmas and conflicts of values [GKM<sup>+</sup>12].

To further evaluate these gray areas of conflict is important because the media strongly influences how people perceive the world. The information broadcasted through news, soaps or films affects the audiences' beliefs, values and fundamental commitments. Even though there are legal, sociological or psychological studies about peoples' preferences or how legal restrictions apply, most studies follow a philosophical nature. They aim to explore ethical issues, what responsibilities, rights and duties exist and how they might conflict by using a philosophical approach. Hence, the effects of, for instance, sexual or violent media content on humanity are assessed and evaluated within the studies of ethics. Even though there might be consensus about such matters, that programs including sexual or violent content should be prohibited, peoples' perceptions on those ethical issues differ. Defined by different moral standards, perceptions and beliefs, the study of right or wrong rarely results in consensus. It can be described as a mirror of society and its values [Kie02].

The study of media ethics includes the definition of principles, such as *truthtelling*, described as a fundamental principle of media ethics in literature. Christians et al. emphasize the thought that telling the truth under all conditions begins with the duty of everyone working in the media industry [CCKW16]. Even though the question of *What is truth?* asked by Pontius Pilate is a struggle to answer, Rosenstiel et al. define



*truthtelling* as seeking and reporting the truth as fully as possible, as well as considering truth as the most significant value and the primary duty of journalism. The principal was further described as accurate, honest, fair and courageous when gathering, interpreting and broadcasting information. Additionally, the principle aims to uncover the unseen and therefore, give a voice to the voiceless; Rosenstiel et al. even take it so far that the principle also includes holding the powerful accountable for their actions [RM13].

Transparency is another principle mentioned in the literature that aims to disclose information on how journalistic content is produced, which includes details about how the reporting was done and therefore, an explanation of the sources, evidence and decisions made. Transparency also means to reveal what is not known in case of mistakes or errors. The journalistic approach, the intent of the information, if, for instance, it is political or philosophical, should be straightforward for the receiver. Transparency also means revealing how a specific viewpoint impacts the published information [RM13]. The principle aims to hold and increase trust between the media companies and the stakeholders [CCKW16].

Additionally, the principle is described as *engaging in communities* or *minimizing harm*, which emphasizes protecting democracy and therefore, the common good of society. This means balancing the public need for information against the potential harm the information could cause to protect those affected by the disclosure of the information. Therefore, censor information and treat, for instance, subjects who are unable to give consent, victims of different crimes, inexperienced or minors with respect [Bro20]. The principle guides towards ethical diversity, which resists the temptation of using potentially harmful information to manipulate through fear or the desire to sensationalize. The principle aims to support the interests of a community respecting democracy by assuming that the community can participate in the discussion and contribute to the conversation [RM13].

These three principles are recognized in the literature as common ground regarding ethical journalism. Other principles are described, including acting independently and serving the public interest [Pla14]. Independence ensures that the journalistic content is free from bias and avoids conflicts of interest. Hence, activities, such as accepting gifts, money or political favors, which compromise the integrity of the journalistic content or damage the credibility of the broadcasting company, should be refused [Bro20]. Moreover, accountability and responsibility are crucial aspects of ethical journalism. Journalists must be accountable for their actions and content, acknowledging any mistakes and correcting them quickly [Bro20]. Respecting privacy is also mentioned and can be compared to the principle of *minimizing harm*, as it ensures that sensitive information is handled with respect and integrity [CCKW16]. Together, these principles form the ethical framework that guides journalists in researching, producing and distributing content.



# CHAPTER 3

## Evaluation Framework

This chapter provides an overview of the framework used to assess the ethical and social implications of AI in public broadcasting. It begins with a comprehensive description of the EU's framework for trustworthy AI, which outlines principles and requirements to ensure AI systems are lawful, ethical and robust. The chapter then introduces the *Z-Inspection* process, a practical approach to evaluating AI systems across different domains, focusing on ethical, technical and legal aspects. These frameworks will be employed in the study and therefore serve as the foundation for the evaluation of the *AiDitor*.

### 3.1 EU Framework for Trustworthy AI

The framework, visualized in Figure 3.1 for trustworthy AI introduced by the HLEG incorporates three components for developing and deploying trustworthy AI: lawful, ethical and robust. Lawful refers to AI complying with all relevant laws and regulations, ethical ensures compliance with ethical principles and values and robust refers to preventing unintended impact and harm from both technical and social perspectives. Although the HLEG emphasizes that these three components should ideally be implemented in harmony, they also acknowledge that some tension may arise between them. As shown in Figure 3.1 the framework of trustworthy AI is structured in three chapters, describing ethical principles as a foundation, the realization and the assessment of trustworthy AI [HLEGoAIEC19]. In the first chapter, *Foundations of trustworthy AI*, ethical principles concerning fundamental human rights are defined in the EU treaties, the EU charter and international human rights law. Based on these fundamental rights, the following ethical principles in the context of AI systems are defined:

- Respect for human autonomy

- Prevention of harm
- Fairness
- Explicability

Respect for human autonomy aims to design AI systems that complement and empower human cognitive, social and cultural skills rather than undermine or manipulate them. The principle of preventing harm states that AI systems should protect human dignity and physical integrity. Hence, they should be employed in a technically robust environment, protected against vulnerabilities and malicious use. Even though the authors are aware of many different interpretations and definitions of fairness, they still emphasize that AI systems should be free from unfair bias, discrimination and stigmatization, increasing social fairness. Additionally, this principle supports equal access to education, goods, services and technology. The principle of explicability is defined as directly communicating the purpose of the AI system to the users. In addition, this principle aims to enhance transparency by the ability to trace and explain how and why the AI system made certain decisions [CDGfCNT20].

Chapter two of the HLEG framework offers guidance for implementing and realizing trustworthy AI by defining seven key requirements, which will be outlined in more detail in the following paragraphs. The requirement, *Human Agency and Oversight*, aims to support people by making their own decisions and respecting their autonomy. This means AI systems should respect and enable a fair and democratic society, protect human rights and allow for human supervision. According to the HLEG, human dignity is an example of a human right that should be protected by ensuring non-discrimination, data and privacy protection in AI systems [CDGfCNT20]. According to Buruk et al., protection can be assured by different processes, such as *human-in-the-loop*, which refers to human intervention in all possible decision cycles of the AI system. The second possible process is the *human-on-the-loop*, which refers to the possibility of a human intervening during AI systems design and monitoring. The last possibility would be *human-in-command*, which is defined as a human overseeing the AI system's activities and deciding in what specific situation the AI system should be used. Hence, all the processes involving AI systems also include some human interaction in terms of control and oversight [BEA20].

The second requirement, *Technical robustness* is defined as being reliable; in other words, it means delivering a trusted service. This requirement defines AI as systems that consistently operate as intended while minimizing unintended and unforeseen harm and preventing unacceptable consequences. Also described as resilience, which refers to robustness when facing charges, they are developed with a preventative approach to risk. The ALTAI questions, introduced in Section 2.1.4, for this requirement are divided into security, safety, accuracy, reliability and fallback plans. The fallback plans include questions concerned with the reproducibility of the output. Reproducibility refers to the fact that two inputs result in the same output [CDGfCNT20].

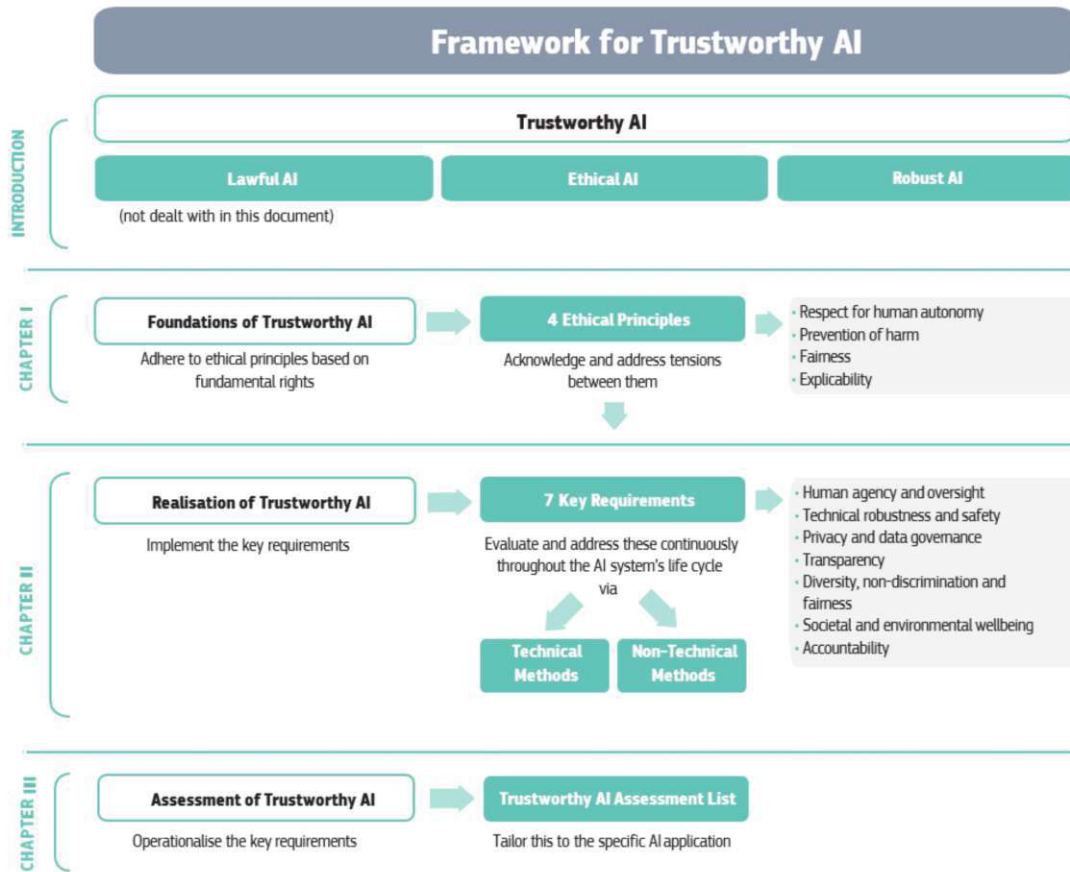


Figure 3.1: Ethics guideline of trustworthy AI introduced by the HLEG of the EU (2019) [HLEGoAIEC19].

Another requirement to ensure trustworthy AI is *Privacy and data governance*, which aims to protect the privacy and data of the users during the entire life cycle of usage, including the information initially provided by users and the data generated during interactions with the system. The data collection process should avoid leading to unlawful discrimination against individuals. Moreover, the requirement says that the quality and integrity of the data collected during the interaction with the AI system might contain biases, inaccuracies, errors and mistakes, which should be considered when generating outcomes and in future training phases. Access to data must be carefully regulated, with clear protocols defining who can access data and under what circumstances, ensuring that only qualified personnel with a legitimate need can access individual data [CDGfCNT20].

*Transparency* covers all relevant elements of an AI system, including data, the system itself and the business model. The sub-requirement traceability refers to the data and algorithms influencing the decision-making process of the AI system. The HLEG suggests that the data and algorithms should be documented employing the best possible standards.

Aiming to identify and prevent sources of errors. Explainability focuses on the AI system's ability to clarify the technical processes and human-related decisions. The AI system should be distinguishable from a human, ensuring users are aware they are interacting with AI and have the option for human interaction if needed. Considering the specific use case, the system's capabilities and limitations should be communicated effectively to the users [CDGfCNT20].

In addition, the requirement *Diversity, Non-discrimination and Fairness* underlines the importance of inclusion and diversity in an AI system in terms of avoidance of unfair bias, accessibility and universal design and stakeholder participation. The AI system should consider all types of biases to prevent harm to individuals and the economy. In case AI is deployed within any context, especially in a business-to-consumer process, the system should be accessible to all users regardless of their age, gender, characteristics or abilities. The requirement also includes stakeholder participation during the development process to ensure full consents during the whole process within an organization [CDGfCNT20].

In order to achieve trustworthy AI, *Societal and Environmental Well-being* must consider environmental impacts throughout the entire life cycle. As AI is employed in various domains such as education, work, care and entertainment, its impact on social relationships should be monitored. Moreover, the requirement addresses the sustainability and ecological responsibility of AI systems, including areas of global concern such as benefits for future generations. Additionally included in the definition is the consideration of the impact of an AI system on society and democracy, such as the influence of fake news and election processes [CDGfCNT20].

*Accountability* is the requirement to ensure responsibility regarding the AI system and its outcome during the entire life cycle. It is closely related to risk management, aiming to identify and mitigate risks as transparently as possible, including audits by third parties [CDGfCNT20].

## 3.2 Discussion of the EU Framework

Despite various strengths and critiques, the literature reaches a consensus that given the latest developments in AI and various experiences of its potential to be dangerous and harmful, strict guidelines, measures and principles need to be defined [RRW23, BEA20, KUD21]. Radclyffe outlines that the ALTAI tool is a significant advancement in implementing AI governance by opening the discussion on how different measures can be applied regarding processes, procedures and protocols. The author mentions successfully transforming the HLEG guidelines into measurable and quantified objectives as a second strength. Ethical values and definitions can vary significantly between individuals and groups, but governance reviews can be quantified through the ALTAI list, producing measurable numerical results. According to the author, another notable advancement of the ALTAI list is the comprehensive scope covering all relevant issues concerning trustworthy AI [RRW23].

Buruk et al. analyze ethical principles and values and offer a critical perspective on various evaluation frameworks. Moreover, the paper discusses efficiency using frameworks to identify ethical dilemmas in practice. One of the first limitations the authors mention is that countries such as Africa, South and Central America and Central Asia do not participate equally in developing the guidelines, which implies that countries actively participating in the development of the framework have a more significant influence in defining ethical dimensions of AI technology than others. When applying the ethical guide in practice, the authors argue there is no hierarchy among the ethical requirements. This means that even though the guidelines help identify the possible consequences, they do not help to evaluate and choose the most critical requirements. This issue leaves the evaluation of the requirements to the developers and users themselves, which, in case of conflicting requirements, makes an application of the guidelines in practice more difficult [BEA20].

Hickman and Petrin argue that within the domain of company law and corporate governance, the HLEG requirements promote many governance concepts but leave many questions unanswered due to their generic definitions. For instance, the requirements such as *Human Autonomy and Oversight* would require further definition in terms of what requirements should AI make value judgments or how and to what degree is ensured that individuals still learn tasks that might be delegated to machines to be able to oversee and control the AI system. The authors emphasize a more granular guidance for corporations. Until then, the guidelines serve more as a starting point to assess trustworthy AI [HP21]. Baldassarre et al. describe the problem as "high-level statements which are hard to translate into concrete implementation strategies" [BGKR24]. Emphasizing developing a more holistic approach to address the challenges of trustworthy AI in industrial cases, closing the gap in the theory [BGKR24].

### 3.3 The Z-inspection Process

To address the issue of applying the EU principles of trustworthy AI in practice, Zicari et al. developed a process, the *Z-Inspection* process, to assess trustworthy AI in different domains. Zicari et al. developed a holistic and dynamic approach to evaluate AI systems at every stage of the AI lifecycle, comprising the definition of different use cases, design and development. As visualized in Figure 3.2 the authors aim to create a process adaptable to any use case and various domains. Thus, the three main phases are designed to be tailored to any domain and team. In the assessment phase, the *Z-Inspection* process considers the seven key requirements of the EU framework for trustworthy AI by mapping the tensions to the requirements. The *Z-Inspection* process can be applied for auditing and performing ethical evaluations during the whole AI system lifecycle. The main idea of the *Z-Inspection* process is to orchestrate different teams of experts and assess the ethical, technical and legal implications of using AI systems. The process is divided into three main phases: the set-up, assessment and resolve phase [ZBB<sup>+</sup>21].



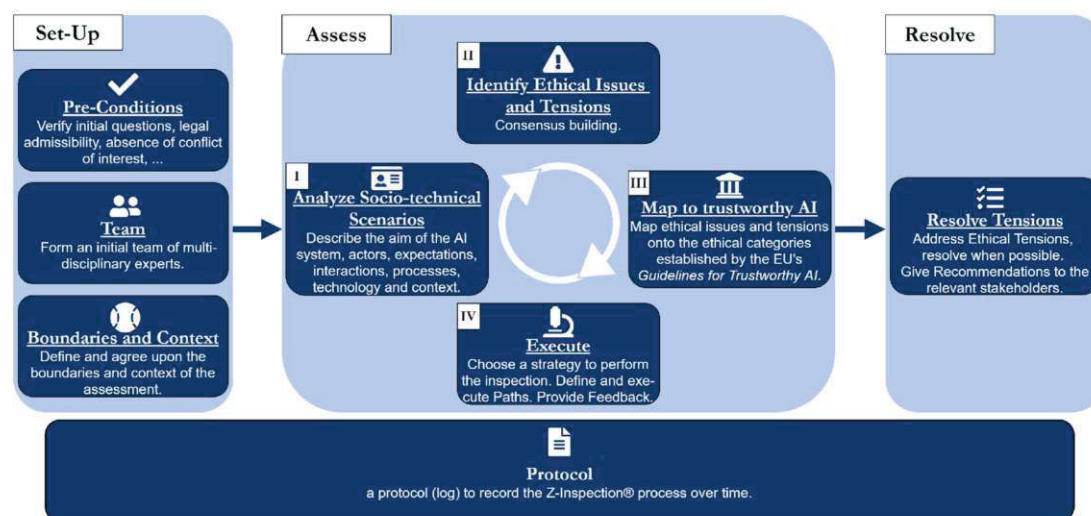


Figure 3.2: Overview of the three main phases of the *Z-Inspection* process [ZBB<sup>+</sup>21].

In the set-up phase, a team of multidisciplinary experts is defined by considering the different skills of team members and the resources willing to invest. The process, therefore, offers a catalog of questions to identify goals and potential conflicts of interest. The authors emphasize that AI systems operate within broader political and institutional contexts and are part of a more extensive scope of processes, products, services and interactions with people and data. For this reason, the set-up phase includes defining the boundaries, the context of the analysis and the time frame for the assessment [ZBB<sup>+</sup>21].

In the assessment phase, socio-technical scenarios are conducted to capture the goals and needs of all stakeholders, resulting in different values. Moreover, the scenarios identify possible tensions between those values by identifying ethical, technical and legal issues arising from using the AI system. A tension represents a conflict between different values, such as the quality of service, which might conflict with the principle of privacy. The output of the assessment phase is a list containing "flags" defining areas that need further investigation. Those issues are mapped to the requirements of the EU High-Level Experts Guidelines for trustworthy AI (HLEG).

Zicari uses the consensus method to validate the results, where the factor  $M$  represents a weight of relevance between ethical issues and flags for instance minimum relevance =  $X(0)$ ,  $X(1)$ ,  $X(2)$ ,  $X(3)$ , maximum relevance =  $X(4)$ . A group of experts assigns weights between the ethical issues and flags, followed by calculating the average score for each participant. When the difference between the scores exceeds a reasonable threshold, a consensus discussion is required. In the resolve phase, the level of trust and risk is quantified and visualized, including recommendations [ZBB<sup>+</sup>21].



### 3.4 Reflections on the Z-inspection Process

Vetter et al. applied the *Z-Inspection* process in healthcare and environmental monitoring, aiming to cover the technical aspects of AI systems and the complex socio-technical context in which the system operates within the assessment. The authors mention that the *Z-Inspection* process enabled to inclusion of various stakeholders from different backgrounds. Moreover, they described the process as a structured approach enabling the identification of ethical tensions and flags. They also outline the possible limitation of the high dependency on the stakeholder's knowledge. The assessment required all stakeholders to be familiar with the process and have a certain objectivity during the assessment. The authors mention the risks of subjectivity by overlooking important information due to gaps in knowledge or biases within the assessment team, such as the absence of experts. [\[VAB<sup>+</sup>23\]](#).



# CHAPTER 4

## Methodology

To address the problem and answer the research questions, a natural science approach will be conducted, focusing on the intersection of social-technical and behavioral domains. The study will employ a case study methodology, using in-depth interviews as the primary data source. A case study is defined as examining a phenomenon in a natural context and employing various methods for data collection to gather information for one or more entities [BGM87]. Benbasat et al. argue that a case study involves multiple research methods. Moreover, the author emphasizes the following four main steps when conducting a case study: defining the unit of analysis, for instance, if the study is focusing on individuals or groups; deciding whether it is a single-case or multiple-case design, data collection and analysis and exposition [BGM87]. The thesis employs a multiple case study design by analyzing the different perspectives of three groups of participants.

Moen et al. propose a guide on qualitative research methods, including participant observation, in-depth interviews and focus groups, data documentation and management. The paper argues that each method has its purpose. For example, in-depth interviews aim to collect data on an individual's history, perspectives and experiences, making them appropriate for the research conducted in this thesis. Moen et al. emphasize that interviews are interactional, meaning-making events where data is collected through an iterative process [MM15].

Qualitative data analysis is an ongoing, reflexive activity rather than a separate part of the research. Various types of analytical work, such as categorizing, searching for patterns, forming theories and relating empirical events to theoretical frameworks, are included in the analysis. The paper emphasizes the need to develop an *intimate* understanding of the data through different methods of comparing and exploring data. It also underscores the importance of analyzing metaphors and different figures of speech. Additionally, the paper discusses the connection between the research context and external models and theories, which helps summarize findings and insights. These findings form the basis for knowledge generalization and theory building. In addition, the paper addresses the

quality of qualitative research by highlighting epistemic values, such as transparency, which refers to the documentation and explanation of all research steps and decisions, the extent to which the researcher is positioned to capture and understand the context and whether the study is free from bias and encompasses multiple perspectives [MM15].

A more general view on the qualitative research methods in a sensitive context, such as addictions, is given by Neale et al., where the methodological steps are divided into planning, data collection and analysis [NAC05]. Where in the planning phase, the potential for ethical issues is addressed; therefore, the paper emphasizes ethical approval when it comes to extremely sensitive topics. Furthermore, access to the data within a specific field, such as persons to be interviewed, is part of the planning phase. The data collection is a summary of what is already described by Moen et al. and the same counts for the analysis part. Moreover, the limitations of qualitative research are outlined in terms of careful sampling of participants, which would result in a lack of reproducibility and generalizability. Another weakness is the potential for bias and misrepresentation and the high demand for resources and time needed to conduct qualitative research [NAC05].

The paper of Mack et al. provides a detailed introduction to qualitative methods, including suggestions for different tools used for the data management procedure, such as transcription protocols or data archive models [MWMG05]. Hence, a general example of what a transcription protocol might look like is given. For example, according to Mack et al., the interview transcript header consists of the participant's identification, interview category, location, date and time information. Also mentioned as necessary is the written consent of all the participants and brief information about the study and the problem statement [MWMG05].

The case study will thus be combined with qualitative methods and the basic concepts of the *Z-Inspection* process, introduced in Chapter 3, to assess the complex socio-technical implications of AI in practice. In addition, the study is divided into the phases of setup, assessment and resolution. Within the setup phase, the study aims to formulate a multidisciplinary team of study participants by respecting the participant's diverse interests, competencies and skills. A definition of the assessment context includes describing the technology in terms of functions, technical background and possible fields of application of the *AiDitor*. This information will tailor the ALTAI questions to the context and specific participant group by respecting the possible socio-technological scenarios.

In the assessment phase, the qualitative interviews are conducted, transcribed and analyzed, resulting in a list of potential ethical, technical and social issues, so-called flags, according to the seven requirements of the ALTAI self-assessment list. The analysis is done according to P. Mayring, who describes qualitative content analysis as a structured method for evaluating text-based data. Hence, the evaluation process is characterized by a rule-based, fixed procedure to enable a flexible evaluation of the data material, allowing different types of research questions to be answered depending on the research interest [MF19].

---

At the heart of the evaluation process is the qualitative interpretation of the data for what reason P. Mayring defines the following steps of the analysis process:

1. Determination of the material
2. Analysis of the situation and conditions the material was created
3. Formal characterization of the material
4. Determination of the direction of analysis (research questions)
5. Theoretical differentiation of the research question
6. Determination of the analysis technique
7. Definition of the units of analysis
8. Execution of the material analysis

Mayring further describes step six, the analysis technique of summarizing, as reducing and filtering the essential information from the material into a manageable corpus of linguistic data. Explication includes additional information, such as literature or encyclopedias, to explain parts of the material in more detail. In structuring content analysis, certain aspects are excluded from the material by a predefined system of representative categories. The categories can be either deductive, based on state-of-the-art research or inductive, based on the data, during the analysis process. Finally, Mayring defines quality criteria to ensure the validity of the research. Transparency is one quality criterion, where every step of the analysis is described and made understandable and accessible for third parties. Another criterion is the scope, which refers to the reproducibility of the content analysis process. Intersubjectivity is given when the data and results are discussed and critically reflected upon [MF19].

A similar approach of qualitative content analysis is proposed by Kuckartz, where text material is analyzed systematically and rule-based. The approach distinguishes between three primary forms: content-structuring, evaluative and type-forming. As a result, three different analysis options are offered simultaneously and it is possible to decide on one form or to combine the forms, depending on the research interest [Kuc18]. Both authors emphasize the thought that the approaches can be tailored, to a certain extent, to the context of the research questions.

In the resolve phase, the results are visualized and the categories are transformed into recommendations. The validation will be done according to Mayring by ensuring transparency, reproducibility and intersubjectivity by critically reflecting on the results. In addition, peer debriefing with the study participants in the field of artificial intelligence will be done, describing how believable and convincing the findings are by *respondent validation* according to Moen and Middelthon [MM15].



# Research Design

This chapter applies the methods introduced in Chapter 4 Methodology, to the context of the thesis by describing a tailored version of the *Z-Inspection* process. As visualized in Figure 5.1 the set-up phase describes the AI tool, including the current stage of development, functions and technical background. Moreover, a detailed description of the participant selection and the interview questions is described. The assessment phase defines the inductive and deductive categories, including a comprehensive description and differentiation between the categories used in the thesis. The resolve phase is covered in Chapter 6 Results.

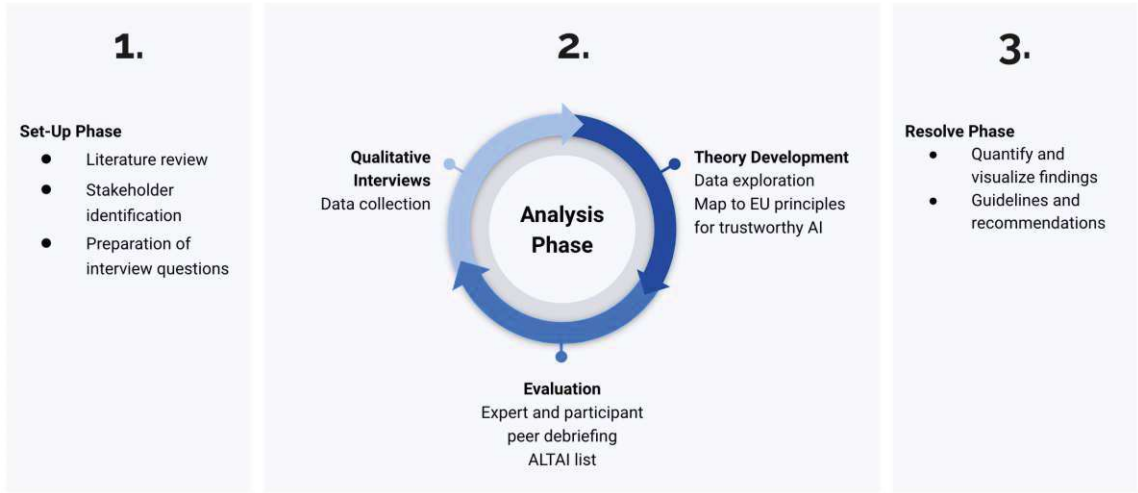


Figure 5.1: Visualization of the tailored *Z-Inspection* process that is used within thesis.

## 5.1 Set-Up Phase

### 5.1.1 Description of the AiDitor

The *AiDitor* developed by the ORF, is an AI tool designed for daily editorial routines, including various functions, such as generative, translate, transcribe and chat, as visualized in Figure 5.2. It serves as a platform that orchestrates different AI technologies into a single tool incorporated into the organization's existing infrastructure.

It allows editorial teams to create and store their customized prompts within their workspace, enabling them to produce various types of content from researched and verified sources. Therefore, the AI tool supports users in generating different journalistic content, including text, images, audio and social media posts. Even more tailored functions to the context are available, such as improving audio, which comes in handy when dealing with imperfect audio content such as background noise in interviews or interactive radio reports.

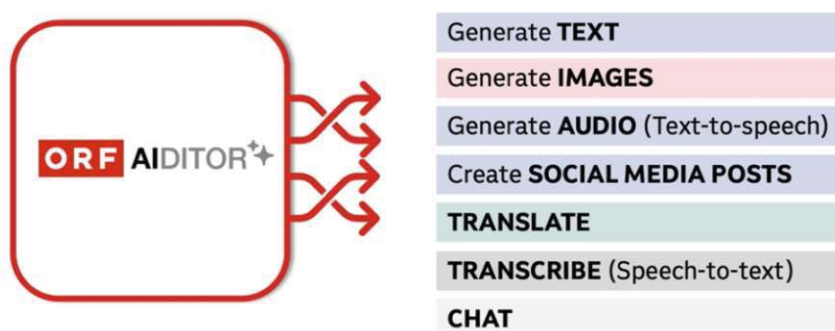


Figure 5.2: Different functions of the *AiDitor*.

The basic idea, illustrated in Figure 5.3, is that the input consists of researched content such as articles, radio reports or TV shows that can be processed into different journalistic formats. The ORF can use the results in various formats and distribute them on different platforms. For example, a journalist could take a verified article as input and generate a social media post or a two-second radio report.

Another use case would be to search for specific keywords in large amounts of data, such as finding exact phrases within an extensive TV discussion. Users can define custom prompts and provide explicit instructions to the *AiDitor* to influence the output of generative functions. Within the default workspace, a general prompt is defined to ensure alignment with the corporate values as well as the preferences and interests of the audience. The personal workspace allows the definition of custom prompts; for instance, if a radio channel's target audience falls within the age range of 18-30, custom prompts can be tailored to incorporate youth language, current trends and topics. Moreover, the tool offers the possibility of AI-based translations, transcribing (speech-to-text) and chatting with various models.



Concerning the technical background, it is noteworthy that the ORF uses third-party services, such as different LLMs, rather than developing and training them by themselves. The user can chat with models such as *Mistral (EU)*, *Claude (Opus)* or different versions of *ChatGPTs*. The same applies to the transcribing function, where the user can choose between *Whisper* or *Deepgram*. In addition, the organization's editorial database is connected with the different AI services used to generate content such as images. Security measures comprising authentication and authorization, including the access control schema of the *AiDitor*, are part of the ORF information technology (IT) infrastructure and are not in the scope of this thesis.

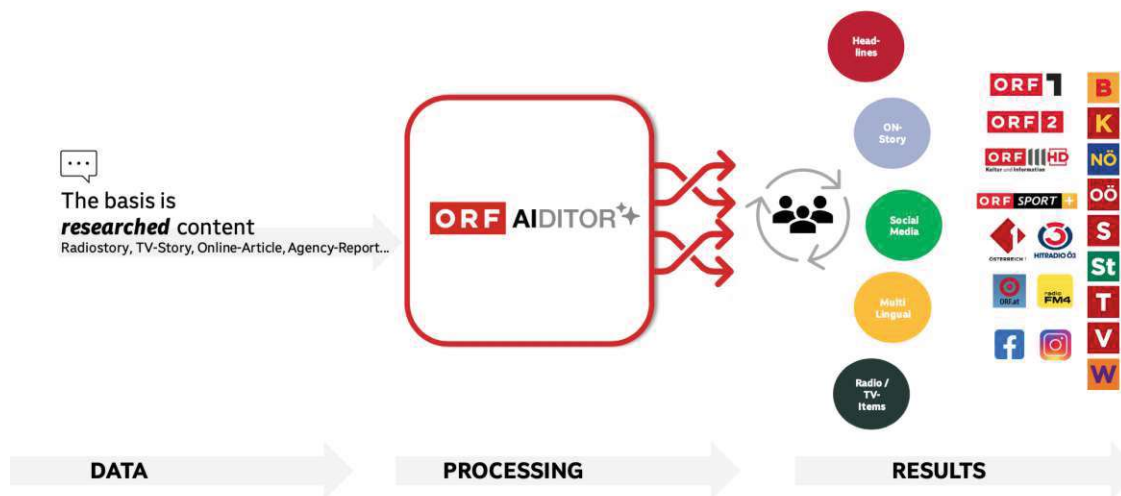


Figure 5.3: Illustration of the architecture, idea and possible distribution channels of the AI-based *AiDitor*.

### 5.1.2 Participants

Various stakeholder perspectives need to be considered to assess both the tangible and intangible opportunities as well as the challenges of employing the *AiDitor* in the editorial field. Hence, three relevant groups of study participants were identified: experts, users and stakeholders, resulting in a total of 13 participants. The experts provide extensive knowledge about the technical background, such as technical architecture, data privacy and governance topics. Since the *AiDitor* directly affects users and their daily routines, it provides another valuable perspective for a comprehensive assessment.

Additionally, stakeholders, such as the board of trustees, management and human resources were identified as relevant study participants. Table 5.1 illustrates descriptive information concerning all the study participants, including the study group and function. The different stakeholder's perspectives aim to assess the internal and external challenges and opportunities, therefore addressing questions such as whether the journalistic content is still believable when the ORF employs AI in their daily editorial processes. Participants with different levels of experience with the new technology and different interests are

Participant ID	Group	Function	Interview Duration
01	Expert	AI innovation management and development	58:47
02	Expert	Development	1:00:50
03	User	Editor-in-chief	16:07
04	User	Radio journalist and editor in the domestic policy department	26:03
05	User	Editor	32:48
06	User	Editor for online and social media	21:10
07	User	Editor	20:50
08	Stakeholder	Management radio	30:22
09	Stakeholder	Management regional	25:37
10	Stakeholder	Management radio	24:51
11	Stakeholder	Editorial board	18:23
12	Stakeholder	Technical director	23:32
13	Stakeholder	Management human resources	15:07
Average duration			28:48

Table 5.1: Information of the study participants and the interview duration.

selected to ensure the study's validity. Another crucial requirement, especially for the users, was that they had already interacted with *AiDitor* and, in the best case, had already integrated the tool into their daily routines.

### 5.1.3 Interview Questions

The interview questions in Appendix A are divided into two main groups: one for users and experts and the other for stakeholders. This was required to collect data according to the research questions and respect different levels of expertise, interests and perspectives. For the group consisting of experts and users, the questions of the ALTAI self-assessment list were tailored to the context of public broadcasting, ensuring that the questions offer comprehensive and insightful information within the assessment.

For instance, questions targeted to the experts were primarily concerned with the technical parts of the ALTAI list. In contrast, the questions for the users were more generic and concerned with topics such as societal and environmental well-being. Questions such as "Have you checked how the *AiDitor* behaves in unexpected situations and environments?" were designed for the experts. For example, "How is the user interaction with the *AiDitor* designed? Are there functions that support you in the decision-making process?" were tailored for the users.

The stakeholder assessment required a more general design of the questions based on the ALTAI list, aiming to capture the diverse opinions and beliefs of the stakeholders. Still, the stakeholder's perspective is critical in answering the first research question, which seeks to assess the opportunities and challenges of using AI tools in public broadcasting.

Another reason was their level of expertise, which is concerned with budget and planning rather than the technical implementation or the daily usage. Thus, collecting relevant data required questions such as, "What opportunities or challenges do you foresee for the ORF regarding *AiDitor*?" Additionally, questions targeted at human resources were designed to capture internal challenges, for instance, "Which skills among journalists do you think will become more important in the future? Which will become less important?"

## 5.2 Assessment Phase

The data collection process involved several key steps to ensure transparency. All the participants were first informed about the purpose of the study and how their data would be used and managed. Their consent, a prerequisite for participation in the study, was always obtained before conducting the interview. Depending on the availability of the participants, interviews were conducted either online or in person, providing flexibility to the participants' schedules. Each interview was always conducted between the thesis author and a single participant.

During the interviews, recordings were made, which were transcribed and anonymized to protect the participant's identities. The transcription of the interviews was performed using *AiDitor*, which was later replayed and proofread by the thesis author. On average, the expert and user interviews lasted 33 minutes and 48 seconds, whereas the stakeholder interviews were shorter, with an average of 22 minutes and 59 seconds.

The data collection process results in transcriptions, which Mayring characterizes as textual data. The transcriptions are formatted using a convention where speakers are identified with initials. Within the research process the initials *I*: denoting the thesis author speaking, while *B*: represents the participants' answers or statements. A single analysis unit corresponds to one interview per participant, where the smallest coding unit is defined by one word.

The analysis technique employed followed Mayring's structured content analysis method. Accordingly, the evaluation of the material used deductively defined categories, predetermined based on the EU framework for trustworthy AI and inductively defined categories

that emerged from the data during the process. Combining deductive and inductive categories allows for an assessment according to the ALTAI requirements and further exploring new perspectives and patterns during the analysis process. The following subsections describe the categories structured according to the two participant groups: (i) experts and users as one group and (ii) stakeholders as another group.

Category	Sub-Category
Human Agency and Oversight	Human Autonomy
	Human Oversight
	Resilience to Attack and Security
	General Safety and Fallback Plans
Technical Robustness and Safety	Accuracy
	Reliability and Reproducibility
	Privacy
	Data Governance
Privacy & Data Governance	Traceability
	Explainability
	Communication
Diversity, Non-discrimination and Fairness	Avoidance of Unfair Bias
	Accessibility and Universal Design
	Stakeholder Participation
Societal and Environmental Well-being	Environmental Well-being
	Impact on Work and Skills
	Impact on Society at Large or Democracy
Accountability	Auditability
	Risk Management
<i>AiDitor</i>	Description of <i>AiDitor</i>
	Functions of <i>AiDitor</i>

Table 5.2: Deductive code schema based on ALTAI list, including the sub-codes.

### 5.2.1 User and Experts

All the deductive categories used to assess the user and expert perspective are illustrated in Table 5.2, divided by the seven requirements of trustworthy AI, including the sub-categories. Based on the ALTAI list, the sub-categories required a further delimitation and are defined as follows:

- **Human Autonomy:** The AI system should support individuals in their decision-making process, being protected from unfair manipulation and automated decisions with significant impacts. Perception should be raised that decisions and interactions are with a system, not a human.
- **Human Oversight:** The AI system does not undermine human autonomy or cause harm, using approaches like *human-in-the-loop* to enable human intervention and control in every step of the system lifecycle.
- **Resilience to Attack and Security:** AI systems must be protected against vulnerabilities and attacks to ensure they are secure, dependable and resilient by preventing and mitigating possible risks and unintended behavior.
- **General Safety and Fallback Plans:** Processes to assess and manage AI system risks should be established, with safety measures appropriate to the risk level of the system. The AI system should have safeguards and fallback plans to ensure safe operation and minimize unintended consequences.
- **Accuracy:** The AI system must achieve high accuracy in judgments, predictions and decisions, significantly when impacting human lives, including mechanisms to indicate and manage potential errors.
- **Reliability and Reproducibility:** AI system must be reliable as a service and the output must be reproducible, functioning correctly across varied inputs and conditions.
- **Privacy:** The AI system must ensure privacy and data protection throughout the entire lifecycle, preventing unlawful or unfair discrimination based on collected data.
- **Data Governance:** The AI system must use unbiased data and ensure data integrity throughout the entire lifecycle. It also applies to systems that are not developed in-house. Strict access protocols ensure only qualified personnel can access individual data.
- **Traceability:** The AI system should ensure traceability and transparency by documenting the used data, processes and algorithms.
- **Explainability:** The decisions of the AI system and humans can be explained, especially when impacting people's lives.

- **Communication:** The AI system must be identifiable as such, inform users they are interacting with AI and decide against the interaction with the AI when human interaction is favored must be provided. At the same time, communicate their capabilities and limitations. Training material concerning correct usage and information on the AI system must be provided.
- **Avoidance of Unfair Bias:** Identifiable and discriminatory bias should be removed in the development and operation phases and oversight measures concerning bias should be installed.
- **Accessibility and Universal Design:** The AI system must prioritize user-centric design, incorporating accessibility features to ensure fair access for people of all ages, genders, abilities and characteristics.
- **Stakeholder Participation:** Throughout the entire lifecycle, regular stakeholder consultation should ensure worker's information, consultation and participation.
- **Environmental Well-being:** AI systems should prioritize environmentally friendly practices throughout their lifecycle, including resource usage and energy consumption considerations, to ensure sustainability.
- **Impact on Work and Skills:** The impact on social dynamics and well-being must be monitored throughout the entire lifecycle.
- **Impact on Society at Large or Democracy:** The AI system's impact on democracy and society should be carefully evaluated, especially in political and electoral contexts.
- **Auditability:** The AI system should be evaluated by internal and external auditors.
- **Risk Management:** Ensuring the ability to report and respond to adverse impacts of AI systems by all stakeholders.
- **Description of *AiDitor*:** Participant's point of view of the AI system.
- **Functions of the *AiDitor*:** Overview of the fields of application and functions used of the AI system.

During the assessment, the deductive categories were enhanced with the following inductive categories, describing the data collected in more detail:

- Human Oversight
  - Level of trust in output (full, medium, low trust)
  - Verifies/Checks the content (always, occasionally, never)
  - Checks content on specific information (names, numbers, facts)

- Decision of relevance
- Human Agency and Autonomy
  - *AiDitor* as an assistance tool
  - Level of awareness that the decisions/suggestions are based on algorithms
- General Safety and Fallback Plans
  - Behaviors in case of unintended behavior/error
- Accuracy
  - Assessed and monitored by humans
  - Experienced low accuracy during the usage
- Privacy
  - Uncertain about personal/sensitive data used by the *AiDitor*
  - AI uses personal data
  - Consent for the usage of personal data
- Data Governance
  - Trusts that the organization handles data governance/compliance
  - Avoids inputting sensible/personal data into the *AiDitor*
  - Aware that user data during the interaction with the *AiDitor* is collected
  - Aware of the authorization model
- Traceability
  - Trace back which data/sources were used to generate content
  - Trace back ORF data sources
  - Usage of different models
  - AI-generated content is not fully traceable
- Explainability
  - Researched content/information (ORF database) as input data
- Communication
  - Informed/trained about appropriate usage
  - Process to communicate the technical limitations and potential risks (e.g., bias)

- Training material/disclaimer provided
- Avoidance of Unfair Bias
  - User is sensitized/aware
  - AI reflects societal issues (mirrors human behaviors and values)
  - Mitigate through human oversight/responsibility
  - Not informed/aware of potential biases
- Environmental Well-being
  - Unconcerned about AI's environmental impact
  - Thinks the benefits of AI in solving global problems outweigh its CO2 impact
  - Concerned about the high cost of energy
  - Concerned about environmental impact
  - Did not think about it
- Impact on Work and Skills
  - Maintain competitive advantage
  - Potential for AI to replace tasks previously done by humans
- Impact on Society at Large or Democracy
  - AI can provide better, faster, more relevant information in the interests of the society
  - Risk of manipulation
  - Opportunity of distribution of mis- and disinformation
  - Issue to tell the difference between what is created by AI or human
- Risk Management
  - Creation of guidelines of the board management
- Opinion/Attitude
- Challenges
- Opportunities



### 5.2.2 Stakeholder

As the stakeholder interview consisted of different, more generic questions, the assessment consists of the following coding scheme:

- **Challenges:** This code comprises all issues that might occur by the usage of AI in the domain of public broadcasting. Therefore, it includes AI limitations, deepfakes, misinformation, misuse, data handling and homogenization of journalistic content.
- **Opportunities:** Describes all the advantages of employing AI and using the *AiDitor* in the journalistic context.
- **Fields of Application:** Including the distinction between tasks expected to be performed by humans and possible tasks that AI could do.
- **Usage of AI:** The code describes all the requirements to ensure trustworthy employment and usage of the *AiDitor* in the editorial context.
- **The *AiDitor*:** This code refers to the tool, including the participants' ideas, interface and experiences.
- **Expectations:** On groups such as management and employees.
- **Roll-out of the *AiDitor***
- **Cooperation:** between public broadcasting organizations
- **Guidelines**
- **Austria as Media Location**
- **Human Resources**



# CHAPTER 6

## Results

This chapter describes the results of the research according to the coding schema introduced in Chapter 5, starting with a comprehensive assessment answering the first research question 1.3: "What are the ethical and social implications of using generative AI tools within the newsroom of a public service broadcaster from a public value perspective?". The analysis illustrates various perspectives, similarities and contradictions in the participants' responses. A more condensed assessment concerning the ethical and social implications will be given in the Section 7.1 Evaluation. From which conclusions will be derived to assess the opportunities and challenges of AI in public broadcasting and answer the second research question. Based on the empirically collected study results, recommendations, described in Chapter 7, will be given to answer the third research question of how trustworthy AI can be employed in public broadcasting. The citations of the participant's statements are translations from German to English.

### 6.1 Experts

#### 6.1.1 Human Agency and Oversight

Regarding the principle of *Human Autonomy* results show that the *AiDitor* is explicitly designed as an assistance tool, providing suggestions rather than making decisions. This is underlined by statements such as, "Decisions in this sense are always made by the user". For instance, when creating and distributing social media content, responses align with the *AiDitor* suggesting content and never automatically distributing it. Statements, such as "It is not going to happen that the AI suddenly posts something, but rather it suggests a text that the user should review" capture the opinion of the experts.

Additionally, experts agree on implementing and deploying the *AiDitor* only as an assistance tool to enhance human capabilities rather than automating daily processes. However, the results reveal contradictions regarding the automation of posting on social

media. One expert mentioned that different functions can be restricted based on an underlying authorization model by stating, "Some people can do things that others cannot, for example, post automatically on Facebook pages".

Regarding that interaction with an AI system should be communicated to the users, experts also reached a consensus. They believe that no explicit communication is necessary. Results reveal that the user is not informed about the fact that the output is algorithm-based nor that they are interacting with an AI system. Instead, the experts outline the thought that the *AiDitor* is integrated into the existing ORF infrastructure, requiring users to explicitly request its usage by logging into an additional website or opening a separate window not requiring further information. Underlined by statements, such as "I assume that the user is aware. In other words, the user is using an AI tool and if a user is using an AI tool, they know that they are using an AI tool. It's separated from any ongoing processes and activities. So, they cannot accidentally generate something with AI, it has to be a very conscious choice".

Moreover, the experts agree that a *human-in-the-loop* is essential within the entire process. They refer to the ethical code of journalists as follows: "Journalists always work according to the journalistic code, which means they have to verify everything anyway". Results show that the experts have developed their definition regarding the level of automation, called *AI-in-the-loop*, which involves a human at the beginning, the AI tool in the middle and a human again at the end of the process. Hence, the experts outline the system's design and functionality in combination with journalistic values as their approach to covering the principle of *Human Autonomy*. However, results show a strong reliance on knowledge and correct human behavior.

Experts state that *AiDitor* informs the user before the first usage and during all processes to prevent over-reliance and to inform that the output can be faulty. However, concerning an explicit procedure or stop-button to safely abort an operation, the expert's responses align. In response to whether there is an option to stop a process, one expert responded: "You cannot, but it does not make sense because only texts are being created. It is not like something is being set in motion that would need to be stopped somehow". Another expert referred to standard mechanisms, such as pressing escape or reloading the browser.

### 6.1.2 Technical Robustness and Safety

Regarding the requirement *Resilience to Attack and Security*, results show that adversarial, critical or damaging effects on the internal organization and external stakeholders were not considered due to the early stage of development. Additionally, there is a strong reliance on the organization's infrastructure, such as firewalls as well as the organization's authentication mechanisms for protection against attacks. The same applies to measures ensuring integrity, robustness and overall security, where the experts refer to the general company IT security policies and guidelines.

Experts say that penetration testing of the AI system is planned before the roll-out. In addition, they state that they use standard encryption mechanisms for communication

between the services. However, the *AiDitor* is not certified for any cybersecurity certification scheme created by the *Cybersecurity Act in Europe* [Act24], nor is it compliant with any specific security standard. Additionally, the results show that due to the *AiDitor* orchestrating different services, the types of vulnerabilities and potential entry points for attacks, such as data poisoning through manipulation of training data, fully depend on external providers.

Results concerning the requirement *General Safety and Fallback Plans* establishment show no process to assess or mitigate potential risks is currently in place. Therefore, technical redundancies or parallel systems, as suggested by the HLEG, are not in place in case of any failure. In such an event, users would need to rely on their journalistic skills and resume work manually, as mentioned by one expert. The risk of malicious use, misuse or inappropriate usage of *AiDitor* is addressed through custom prompts as well as input and content filters, considering the journalistic context. Experts mention that standard filters are only partially applicable because journalists must report topics such as fatal traffic accidents or knife attacks which would be made impossible through standard filters. To assess the dependency of the *AiDitor* on stable and reliable behavior, the experts described an approach called *monkey testing* where the model's output is manually tested and assessed by inserting random and malicious content. The experts identify the risks of malicious input and output, unsatisfied users and the usage and publication of untrue or false information.

Regarding the requirement for *Accuracy*, results show that a feedback form is integrated within the *AiDitor* where different levels of accuracy and additional problems can be reported to the developers. Apart from that, no process or mechanism exists to monitor or document the *AiDitor's* accuracy. Moreover, the experts argue that a low level of accuracy only leads to critical, adversarial or damaging consequences if the human factor in the process fails. That is underlined by statements, such as "It could for sure happen if the generated products, the generated texts were not checked and if the journalists were not required to verify everything before publishing it". However, implementing a quality control mechanism is planned and incorporates the detection of errors in the generated content by another AI model. Based on the experience that especially numbers were generated incorrectly by older models, the feature would allow cross-checking against facts of the input data.

*Accuracy* also requires ensuring that the data used, especially the data to train the models should be up-to-date, of high quality, complete and representative of the environment where the system will be deployed. Since the ORF is not training any models but using third-party AI technology as a service, the responses align that the responsibility of clean training data does not fall on the ORF. Statements such as "Since we are not training any models for content generation ourselves, I believe this is something we do not need to consider because it simply has not affected us" and "We are not training any models yet, so this is not an area that concerns us" are mentioned in the context of model training and accuracy.

According to the experts, *Reliability and Reproducibility* in journalistic work require a certain degree of creativity, such as different outputs. They also point out that generative AI's nature is to produce varying results, making the measurement of *Reliability and Reproducibility* less critical. However, these aspects have been addressed in the development process by tweaking parameters. The experts highlight a high requirement for reproducibility when *AiDitor* is used for research, emphasizing the importance of models that document their sources. The responses show that mechanisms to measure *Reliability and Reproducibility* will be implemented within the automated quality control using different AI models. However, due to the current development phase, no procedure is in place for handling cases where the AI system yields results with low confidence.

### 6.1.3 Privacy and Data Governance

Regarding the requirement of *Privacy* within the process of logging interaction data, the experts' responses do not fully align. One response indicates that privacy is addressed in logging different information, where user data and statistics are collected and anonymized. For instance, the users' content and metadata of prompts are logged and anonymized until the user explicitly gives feedback on a certain issue. There is no possibility of tracing back details concerning the daily usage of *AiDitor* at the user level. Other responses mention that logs are not anonymized due to the current stage of development because it makes addressing errors more efficient. In addition to the second response, the participant mentioned that they plan to anonymize the data according to the organizational privacy guidelines before the roll-out.

Experts reached consent that either personal data or the advice of a data protection officer was involved in the development process. Thus, no information concerning the *General Data Protection Regulation (GDPR)* or a non-European equivalent could be collected during the study. However, the results reveal that no sensible or personal user data is employed or stored in databases during the process. Moreover, responses align regarding the authorization model to regulate access to only qualified personnel.

### 6.1.4 Transparency

As previously mentioned in the *Privacy and Data Governance* subsection, user interactions with the *AiDitor*, including errors and metadata related to user statistics, are logged. Experts argue that *Traceability* is largely ensured by *AiDitor's* design and functionality, as it typically requires human-researched and verified content as input, except for its conversational functions. Study results indicate that experts are aware that even though the input content may be human verified, the outputs from pre-trained models are always based on third-party data and algorithms and therefore, can never be fully traceable.

Regarding the requirement for *Explainability*, the consensus among the responses is that in the context of public broadcasting, explainability of the different outputs generated by the *AiDitor* is of low importance. This is because the decisions made do not significantly impact people's lives or cause harm to society. It is ruled out that the *AiDitor*, either now

or in the future, will fully automatically publish content requiring *Traceability*. Experts refer to guidelines currently being created within the organization concerning the labeling requirements of AI-based journalistic content.

Experts state that *AiDitor*'s capabilities and limitations have been communicated to users through explicit training sessions and inline information. Although there are no explicit training materials, the approach involves offering explanations and help directly within the *AiDitor*, combined with onboarding video material. The study shows that the focus is primarily on demonstrating the tool's opportunities and capabilities rather than teaching rules and proper usage, as it is assumed that journalists are already operating according to a specific code and must follow editorial rules.

### 6.1.5 Diversity, Non-discrimination and Fairness

The experts' responses on addressing unfair bias do not completely align. One opinion is that there was no direct strategy or process for addressing bias, relying instead on the input content, which is always based on human verification. This is highlighted by statements for instance "As far as I know, there is not a process except for the quality control that would, for example, say that the text or the statement is not correct in terms of content". Quality control refers to the previously mentioned approach where different AI models oversee and cross-check the output against facts, which is currently in development. Another opinion states that biases are directly addressed through the creation of custom prompts, which refer to the general guidelines and workflow of a public broadcasting company, which creates content for the public by respecting all critical, socio-critical and ethical issues.

The principles of *Accessibility and Universal Design* are currently addressed through standard browser mechanisms and the implementation of keyboard usage for the visually impaired. However, experts rank this as a very low priority because the *AiDitor* will be used exclusively within the organization by its employees and is, therefore, not planned to be employed within a business-to-customer context. Results show that there currently is no need or explicit use case to design the *AiDitor* to accommodate all potential users regardless of age, gender, abilities or other characteristics beyond the standard mechanisms already addressed.

Experts mention the incorporated feedback form within the tool to address the requirement of *Stakeholder Participation*, therefore involving users throughout the entire lifecycle of the *AiDitor*. Another planned feature is an AI-supported bot capable of answering user questions without requiring the user to fill out a form. Experts also collect feedback on errors, general questions and user satisfaction regarding the quality of the content through a thumbs-up or thumbs-down feature. The objective is to monitor *AiDitor*'s unexpected behavior, improve prompts and capture general user satisfaction.



### 6.1.6 Societal and Environmental Well-being

Results about *Environmental Well-being* show that potential negative impacts of the AI system on the environment were at the very bottom of the priority list of the experts, represented by statements such as "I do not want to say I do not care, but it is not part of my job, so to speak". One expert mentioned that they switched to GPT-4 mainly because of performance advantages rather than the advantage that everything is twice as fast and only generates half of the CO2 emissions. Statements such as "Yes, so the topic was not highlighted in the foreground, but of course, it is somewhere in the back of my mind" describe the study results. There are no mechanisms to evaluate the environmental impact of the *AiDitor*'s development, deployment and use, such as the amount of energy. The statement "Yes, you could see computing power as power consumption if you want, but I assume that overall, the benefits of AI are higher than the costs because a person who would do that also has to eat something. So, I think energy is a difficult topic. I would say there are no negative effects from computing power" was mentioned in the context of *Environmental Well-being*.

Regarding the *Impact on Work and Skills*, only a few short statements were made by the participants, addressing the requirement in a more general way. For instance, one mentioned that when using the *AiDitor*, a smaller workforce would be required because daily routine work could be done more efficiently. Another statement described that positive or negative implications could occur depending on the point of view but does not mention any further information. However, one expert outlines the risk that humans could rely too much on the system and therefore, require specific training to prevent the risk of de-skilling and acquiring new skills.

The impact on society or democracy is addressed in only a few brief statements. The experts agree that the *AiDitor* could negatively affect society or democracy only if the *human-in-the-loop* fails to fulfill their responsibilities. Statements like "You can only answer that with a yes if the person responsible does not fulfill their responsibilities" illustrate the perspective of the experts. Results illustrate that experts define the impact of *AiDitor* on society or democracy as primarily a result of human error. Consequently, one response emphasizes the importance of user awareness and education to minimize any negative impact.

### 6.1.7 Accountability

The research reveals that the current state of development is more focused on the functionalities and exploring the usage of the tool rather than on risk management. However, the organization has established an AI sounding board primarily concerned with creating guidelines, defining risks and developing mitigation strategies. The experts highlight the importance of having a centralized unit within the organization to manage potential risks, given the global advances and rapid technological changes. The AI sounding board also works in cooperation with the ethics board of the organization,



working on the "living object" and the AI guidelines. Additionally, specific training concerning potential risks will be developed and conducted before the roll-out.

The study indicates that *AiDitor* and its entire infrastructure are constantly audited by third parties, such as the network department. However, the results reveal that the coding still needs to be checked, as the plan is to provide and publish the code as open-source software to enhance trust, transparency and public value. Additionally, experts mention that the ORF has collaborated with various European public service companies, sharing experiences and opinions on AI.

## 6.2 Users

### 6.2.1 Human Agency and Oversight

Findings regarding the requirement of *Human Oversight* considering the level of trust in the *AiDitor*'s output range between full, medium or low confidence. Most respondents stated that they have either low or medium trust in *AiDitor*. Participants with low trust highlighted that the output is questionable, requires verification and is unsuitable for any decision-making process. Medium trust was described as the output being reliable, requiring some verification and suitable for non-critical decisions. Exactly one participant stated to have full confidence in the accuracy and reliability of the output as well as experiences the outcome to be suitable for critical decision-making. Results also show that users expect the output of AI tools to become more reliable in the future. They believe it is constantly improving and will undoubtedly be part of future journalistic daily routines.

One statement refers to an information bias: "So with the *AiDitor*, an ORF's tool, I have relatively high confidence because the *AiDitor*, as far as I know, is trained with our texts and therefore has a good baseline". Further describing if the *AiDitor* was trained on thousands of ORF texts, the tool successfully graduated the ORF school. Based on the current state of development, the ORF is not training any models but rather using pre-trained models as a service. The user also mentions that if the *AiDitor* had been trained on external texts, it would be more difficult to have confidence in the output because the data and parameters of the model are still being determined. Most users agree that the in-house development of the tool aligns with the ethical and journalistic principles outlined in Chapter [2.4](#).

The participants mentioned consistently checking and verifying the output of the *AiDitor*. Statements such as "Yes, always 100%, absolutely. It has never happened that I released the content without checking it. That just does not happen" reflect the study results. Users check on different types of information, such as names and numbers. The only issue identified in the assessment was about the ground truth. Results indicate that users do not always check and verify the output if the *AiDitor* is used for time-consuming tasks, such as transcribing interviews or summarizing large texts. One example mentioned was finding a specific phrase a German politician said within a discussion. In this case, the

*AiDitor* identified the phrase within seconds, but the user still listened to that specific part of the interview to be sure. However, when summarizing a large text corpus, users tend not to check the source text.

Study results show that *AiDitor* is described as an assistance tool that primarily aims to make suggestions rather than autonomous decisions among all participants. Multiple statements underscore the importance of determining what is relevant in journalistic content as a key element of journalistic thinking, which would only make sense to automate partially. In other words, the *AiDitor* is a tool that supports routine tasks and enhances human capabilities and creativity. Additionally, findings show that all users are aware that they are interacting with a non-human machine and understand that the output is based on algorithms.

### 6.2.2 Technical Robustness and Safety

Insights reveal that in cases of unintended behavior or errors, most users take action by directly contacting the developer or using the feedback form. Other's first attempt is to close or reload the browser application or consult with colleagues before reaching out to the development department. Only one user mentioned not taking any action by responding, "Nothing. I do not expect that this thing is infallible".

Most users mentioned that they had already experienced low accuracy with the *AiDitor* during usage. For instance, while using the chat function to proofread text, the tool found mistakes that were not inside the source text. Another example describes fully inadequate and unusable results when performing translation tasks with the *AiDitor*. Based on the evaluation, study results reveal that monitoring against accuracy is primarily done by humans. Some users mention phenomena, such as hallucinating AI and therefore prove awareness of certain phenomena and limitations of AI.

### 6.2.3 Privacy and Data Governance

Concerning the requirement of *Privacy and Data Governance*, the answers and levels of knowledge differ greatly and are described by different thoughts and experiences. Even though there was little data collected concerning privacy, the perspectives align and show that most of the users are uncertain about using personal or sensitive data by the *AiDitor*. Regarding the principle of data governance, no shared consent or contradictions could be found; rather, personal opinions and expectations were collected. One user is convinced that the organization will handle governance and compliance requirements. The statement underlines this, "I assume that once we have rolled out this tool, we will have spent an enormous amount of time and effort ensuring it is compliant, as there are huge compliance departments everywhere and they will have already taken care of it". Another user mentions avoiding inputting sensitive or personal data into the *AiDitor*, fearing it might be stored in a database. Other participants are aware of the authorization model, recognizing that they have more permissions than others.

### 6.2.4 Transparency

Regarding the requirements of *Explainability* and *Traceability*, user responses indicate that when ORF data is used as input, it can be traced back except for the data the AI is trained on. Moreover, the participants' opinions align with the idea that AI should not be used for research. Users agree on the importance of employing AI models, such as *Perplexity*, which indicates the sources used to generate the content. One participant mentions that it would be ideal to have the *Perplexity* functionality applied only on ORF data sources. Users are informed about the strengths and limitations of different models and the appropriate tasks for their use, acknowledging that AI-generated content is never fully traceable.

The results vary greatly concerning whether users were informed about the appropriate usage of the *AiDitor*. One participant mentioned not being informed about the appropriate usage, while others claimed to have learned it autodidactically by stating, "I taught myself. I experimented a lot and played around with it. I just spent a lot of time working on it. Of course, colleagues helped me at times, especially with transcriptions and such. But I actually learned it myself. It is just practice and spending time with it, I would say". Another user mentioned that there was an official offer for training but decided not to participate. Others describe it more as a feeling that the training was appropriate. One response was indifferent concerning the topic, with the user expressing confidence that regulation regarding the appropriate usage is handled by the organization, saying, "I hope so, yes. I am the kind of person who. . . It is not that important to me. I think many people are already dealing with all of this. So, I am not too worried about it".

Many responses could also be found concerning *Communication* about technical limitations and potential risks, ranging from yes to no. Two participants referred to a disclaimer at the beginning but could not exactly describe its content or mention instantly clicking it away. Other users mentioned being informed partially by saying, "About restrictions, yes, about risks, no. But that was implicit knowledge that was assumed", indicating their self-initiative. The same applies to responses concerning training material and disclaimers, with user responses varying greatly.

### 6.2.5 Diversity, Non-discrimination and Fairness

Responses indicate that no official information from the organization was given regarding avoiding unfair bias. However, most participants were aware of different biases of AI systems, except for one. Results show that users informed themselves and are interested in understanding the nature of AI and LLMs. Some users described examples where they experienced different biases while using the *AiDitor*. One example involved the generation of images, where winter and Christmas themes were presented to reflect a more Western cultural perspective.

Additionally, results reveal that users are confident in recognizing bias in AI image generation with minimal effort but find it more challenging in textual data. One user mentioned that recognizing bias in data is common sense for a journalist and can also be

found in non-AI-generated content. Another user referred to bias as a mirror of society, saying, "This is not a problem that AI has, but a problem that we have as a society. AI only reflects what we are".

### 6.2.6 Societal and Environmental Well-being

Different perspectives about *Environmental Well-being* were collected. For example, one user thinks that the benefits of AI in solving global problems outweigh its CO2 impact. In contrast, others are unconcerned about AI's environmental impact or have not yet thought about it. Other perspectives are defined by a concern of the high energy cost generated by AI training and usage. Regarding the *Impact on Work and Skills*, most participants outline the importance of gaining a competitive advantage by employing AI tools in the editorial field. One participant mentioned the potential risk for AI to replace tasks previously done by humans. The *Impact on Society at Large or Democracy* is addressed by outlining negative perspectives, such as the thread of AI used to efficiently create and distribute misinformation and disinformation, as well as the risk of manipulation or issues of humans to tell the difference between AI and human-generated content. One major concern is the accountability of journalistic content, which could be negatively influenced by people knowing the public service company employs AI to generate content. One participant mentions the advantage of AI's ability to provide better, faster and more relevant information in the interests of society. Only two responses were collected regarding the requirement of *Accountability*, outlining the importance of creating corporate guidelines and regulating the use of AI in public broadcasting.

### 6.2.7 Opinion/Attitude

Results outline different perspectives concerning the usage of AI in public broadcasting, from rather critical to open-minded, including that it is too early to have an opinion. One standpoint outlines the importance of staying competitive and fulfilling the requirement to serve multiple platforms and products most efficiently. Another perspective describes AI technology as part of a technological change process, emphasizing AI as a tool to increase productivity and quality of journalistic content. Others emphasize the creation of regulations and guidelines to ensure trustworthy AI and enhance the understanding of the nature of AI technology. Another opinion is that AI should be used as a support tool to enhance human capabilities rather than to replace them.

### 6.2.8 Opportunities

Users outline various fields of application, such as working with texts by utilizing AI for text correction and translations, summarizing information and increasing efficiency when handling large amounts of data. Another example in the context of efficiency mentioned was the significant reduction of time required to transcribe information. Additionally, opportunities for enhancing creativity were mentioned, such as using AI as a starting point for brainstorming or generating various headlines for inspiration. Furthermore, the

participants identified the potential for improving existing content, such as enhancing audio quality by reducing noise in recordings or interviews. Opportunities regarding the organization of metadata were also mentioned. One participant described the potential for a collaborative process, integrating human and AI capabilities and strengths.

### 6.2.9 Challenges

User responses highlighted several challenges of using AI in public broadcasting. While some users have no concerns about using AI tools, finding them invaluable for daily tasks, others mentioned different limitations. Participants find AI inappropriate for research purposes or acquiring new knowledge due to hallucination and data bias. Another challenge outlined was the problem of AI creating and distributing harmful content for society, described as emotionally charged fabricated information aiming only at sensational journalism rather than fulfilling objective reporting requirements. Besides the opportunity to spread mis- and disinformation, AI can also be used as a propaganda machine, especially if AI systematically makes the same mistakes, leading to systematic discrimination or bias, as mentioned by one participant. Concerns were also raised about the misuse of AI due to missing guidelines or rules, the improper labeling of journalistic content (particularly when only technical changes are made) and the relevance of information provided by AI. Furthermore, integrating AI into the existing IT infrastructure presents the threat of unverified AI-generated content being released automatically. In addition, users mention the lack of access to up-to-date information as a significant drawback, as well as the model's training, is based on random information of the internet.

## 6.3 Stakeholders

### 6.3.1 Challenges

Results show that the stakeholders are aware of AI limitations and therefore, highlight the importance of obtaining credibility, referred to as the ORF's greatest asset, as one of the biggest challenges. Participants also mention creating artificial realities by distributing misinformation and fake news. In addition, AI limitations such as hallucinating and the misuse of AI are described as threats to objective reporting by the participants. For instance, the problem of misuse is outlined by statements such as: "I am worried that it can also be used negatively and a certain AI bias steer reports in a certain direction".

Moreover, lacking *Human Oversight* is mentioned, such as prompts that could lead to faulty output that might get published. Another challenge described by the participants is data handling, including inputting sensible data into the *AiDitor* that would leak information that is not meant to be published. Finally, the thread of homogenization with regard to the output is underlined by statements such as: "Due to always inputting the same topics, some topics will be reinforced, while others are not perceived as positively or less frequent".

Other challenges about the use of AI are outlined by the concern that it could undermine the audience's trust, as people might become unsure about what is real and what is not. There are also worries about how employees will handle AI, with some fearing that not everyone will approach the technology positively or responsibly. Additionally, there is skepticism from the audience regarding the necessity of journalists with the introduction of AI, which needs to be addressed properly.

### 6.3.2 Opportunities

The stakeholders are convinced that introducing AI will benefit the organization and its employees. Journalists recognize that tools such as the *AiDitor* can make their work easier. The discussion includes ideas about AI taking over certain tasks, which will relieve journalists and free up resources, therefore maximizing output with the usage of AI. To address the employee's fears about AI responses, outline the strategy to promote AI within the company to handle the ten most tedious processes, thereby eliminating the most annoying repetitive tasks.

Another opportunity mentioned by the participants is the potential to serve the different ORF platforms by detecting fakes and publishing truthful information. One stakeholder expressed hope that AI could more easily support the production of verified content and distribute it across multiple platforms, contrasting with the misinformation that often circulates. Most of the stakeholders' responses outline the pressure to keep up with different platforms to be addressed by AI, making processes more efficient and easier.

Another benefit of employing AI in public broadcasting is cost savings. One participant mentioned that AI could allow ORF to produce content in multiple languages for different target audiences, which is currently not feasible due to cost constraints. Some responses outline the benefit of employing AI to gain a competitive advantage within the media landscape. The participants always mentioned that the competitive advantage must be connected to trustworthy AI employment, which was described as careful usage.

### 6.3.3 Fields of application

The fields of application are defined by the separation of tasks between humans and AI, as, for instance, humans should do creative tasks. Hence, the participants describe the role of the human as overseeing content and ensuring the output fulfills the ethical values and the quality requirements of a public broadcasting company where AI is intended to support journalists and relieve them of repetitive and time-consuming tasks. Described by statements such as "For example, I have a three-minute radio report. I can create a text from this report in just one step using speech-to-text technology. Additionally, I can search for related images, generate social media posts, translate the content and integrate chat functions. This allows me to complete usually time-consuming tasks that do not require significant journalistic expertise in a single step".

However, responses align that fully automating journalistic processes is not imaginable, underlined with statements such as "So this miracle thing, AI, does all of that automati-



cally. I have my doubts, though, whether in the future every ad text, every ad campaign, every image, every background element for the show *Zeit in Bild* will be automatically generated". Moreover, responses contradicted concerning the need to check the output of the *AiDitor*. One participant, for example, mentioned that it is not always necessary to check the automated text-to-speech traffic updates at night, while others insist that the output is always reviewed by humans.

#### 6.3.4 Usage of AI

To employ AI within the ORF while upholding journalistic standards, participants emphasized the importance of having a human in the loop. They believe that editorial teams, responsible for the published content, should always make the final decisions. By incorporating a *human-in-the-loop*, participants stated that the credibility of the published content can be ensured through human validation and fact-checking of the AI output. Participants agreed that public broadcasting companies have a special responsibility to society at large, which is why they emphasize the trustworthy employment of AI and outline the importance of obtaining public value. Participants also described different skills needed to work with AI, such as creating prompts.

#### 6.3.5 AiDitor

The technical director commissioned the *AiDitor* to promote AI literacy among employees and reduce fears connected to the new technology. Statements such as: "ORF and media companies are increasingly becoming technology-driven companies. We need to be open to innovation in both editorial and technical areas, which could prove to be a significant challenge" underline the importance of keeping up with current technological trends. Results reveal that the *AiDitor* was developed on a small budget within a small department, implementing different use cases of AI in the editorial field. The attitude towards the current state of development of the *AiDitor* is described as skeptical but exciting. Some responses attribute the different functions of the *AiDitor*, such as the automatic generation of online articles, to being skeptical, where suggestions for different headlines are considered helpful and as a good idea. Participants' attitude aligns towards the new technology, the functions and possible advantages of the *AiDitor* to be promising.

#### 6.3.6 Expectations

The study reveals expectations regarding AI to make editorial processes more efficient rather than automating them. This expectation is driven by curiosity towards the new technology, the interest in new developments and the possible opportunities to make the life of a journalist easier. One response describes that AI is not infallible, which also applies to humans and therefore requires a good balance of common sense while using the *AiDitor*. The employment of AI usage in journalism should grow gradually because, in public broadcasting, the only currency that matters is the trust of the audience. One expectation regarding AI in journalism is that people will understand and accept that AI

can still produce meaningful journalism. Once this acceptance is achieved, AI can be used for various tasks.

Participants expect the organization's management to engage with AI and promote the use of the *AiDitor*, encouraging employees to adopt it. Additionally, participants emphasize the importance of addressing employees' fears of being replaced by automation. Other observations indicate that the participants expect transparency in labeling AI-generated content. This is outlined by statements such as: "We have set the bar that everything presented to the public is genuine and not generated by AI. If something is AI-generated, then there must be absolute transparency about it".

Other responses outline that public broadcasting companies are seen as the fourth pillar of democracy with certain oversight powers and therefore must maintain these standards. Due to the funding by the Austrian households, participants mentioned the increasingly high external expectations and demands on the ORF. Therefore, the responsibility of AI lies within the editorial team, fulfilling the requirement of *Human Oversight*. Underlined by different stakeholder statements such as: "As a company, we need to consciously make decisions and acknowledge that we are dealing with a fantastic technology with a high potential to be misused. We must choose to stay on the right side and use AI tools that represent the best possible standard available".

In addition, the results describe the expectation of staying competitive and demonstrate ORF's engagement with modern technology by not losing essential journalistic skills. For instance, participants mention the importance of critical thinking skills and the ability to analyze and assess different outputs. AI should be used to detect fakes across various channels and counter misinformation bubbles. Furthermore, it should be ORF's responsibility to report and educate the public about media literacy. The content should be presented in a way that is easy for the audience to understand.

Employees are expected to be curious about AI and willing to experiment with it while ensuring control over the output. This introduces a new dimension to journalistic work, involving traditional research, critical questioning and verifying the edited content to ensure it reflects the intended meaning. Reinforced by the statement such as "I think this is a completely new skill being demanded of us. One aspect is journalistic research: thinking, asking the right questions and critically examining everything. But now we also have to thoroughly check the edited or shortened product to ensure it captures the essence of what we have discovered or created. I believe this represents a whole new dimension of our work".

Research indicates that participants expect a change in journalistic education. Therefore, responses highlight the possible change in journalistic work and their set of skills, as emphasized by the statement: "This also concerns a very important area of journalistic work, namely journalistic education. It will change drastically because journalists of my generation and many after have learned their craft by doing exactly what AI now accomplishes in a faster way for us". Responses outline that more technical skills will be needed to ensure trustworthy usage of the AI tools.



### 6.3.7 Roll-out

To ensure the responsible and trustworthy use of AI within the given domain, stakeholders emphasize the importance of creating guidelines to regulate its usage and potential fields of application. To fulfill the legal mandate while maintaining the public values described in statements such as "The public should be able to rely on ORF to handle AI responsibly and ultimately set guidelines for the Austrian media landscape". One suggestion is a traffic light system: green for acceptable uses, yellow for debatable ones and red for prohibited uses. These guidelines will be developed by employees from various departments, including technology, legal, training and editorial, to determine which competencies should not be replaced by AI.

Moreover, participants mention that training employees to use AI is essential. Regular workshops and training sessions should be implemented as well as a process to communicate the risks and limitations of AI. Some responses suggest that mandatory training may be required to ensure trustworthy engagement with the technology. Management is expected to support and drive this change, especially addressing employee resistance.

### 6.3.8 Collaboration with Other Broadcasters

Collaboration with other European public broadcasters is important and should be strengthened, especially regarding technological developments. This is facilitated by the *European Broadcasting Union (EBU)*, which promotes cooperation in programming and technology, including news where correspondents may not be present everywhere. European public media companies are not seen as competitors but as partners who can support each other with various tools, preventing the need for each to develop its own tools. The *AiDitor's* use in other media houses is an example of this practical cooperation.

### 6.3.9 Limits

Stakeholders also mention the limitations and risks of AI. For instance, AI-generated information that does not reflect the truth is described as a major threat to ORF's credibility. For this reason, some participants mentioned that AI should never be used to create synthetic realities such as avatars and artificially generated voices. Concerning avatars and synthetic realities, stakeholder responses are contradicting. Some see the creation of synthetic realities as an opportunity that the public will accept and others view it as a significant threat to ORF's credibility.

One statement highlights the concern and limitations of using synthetic realities: "The idea that a press release, say from the chancellor, could be read by an artificially generated voice of the chancellor just because it sounds better, I find that unacceptable. Or that avatars of moderators are recreated and appear on air so that the viewer does not know whether it is a real person or an avatar. These fabricated realities must be prevented". Furthermore, statements emphasize that the artificial use and virtualization of a human personality should only occur for individuals actively employed by ORF. Additionally,

participants mention the AI limitations in the context of human responsibility and oversight by mentioning that it should always be ensured that the overall assessment of the output lies with the editorial team and that the results are fact-checked.

Another limitation mentioned by the participants concerns the training data. Although some participants are aware of the limitations related to training data, there seems to be an information gap within the organization. Some responses are convinced that the *AiDitor* is trained on verified ORF content, therefore describing the tool as safe. This is reinforced by statements such as: "Well, the data that the ORF accesses with the *AiDitor* is not just any LLMs trained on random information, but rather data internally produced within our editorial text system. So, one can definitely rely on the validity of the data, ensuring that hallucinations will not occur".

### 6.3.10 Austria as Media Location

Observations show that the *AiDitor* aims to strengthen Austria as a media location by providing a significant competitive advantage. Highlighted by statements such as: "We are either on board or we will experience a significant competitive disadvantage". Additionally, participants mention that the Austrian media landscape should be strengthened by developing software that does not originate from American or Chinese corporations. This is particularly relevant due to the ORF's financing structure through the household levy. Furthermore, ORF should take on a pioneering role in the context of AI and set guidelines for the Austrian media landscape.

### 6.3.11 Human Resources

Few data could have been collected to understand the implications for human resources within the organization. However, participants expressed concerns about being replaced by AI and the human resources department currently identifies no additional issues or mitigation strategies.

## 6.4 Results with respect to the research question

In this section, the results will be used to address the research question [1.3](#): "What are the opportunities and challenges of using AI tools in public service broadcasting from the perspectives of different stakeholders?". The results are quantified and visualized in figure [6.2](#) and [6.1](#) highlighting the five most mentioned challenges and opportunities by the study participants.

Figure [6.2](#) illustrates the opportunities driven by the participants' high expectations to enhance efficiency in editorial routine tasks. Various experiences with the *AiDitor* support this opportunity. For instance, the ability to transcribe is highlighted as a significant time-saver in editorial routines. Another example mentioned by four out of five users is the increased efficiency in the inspiration and creative tasks of journalistic work, such as generating ideas or headlines. Furthermore, the potential to automate and improve

efficiency in various journalistic processes is associated with resource and cost advantages, an opportunity predominantly highlighted by management, reflecting the participants differing interests in employing AI. While these two opportunities are comparable, they capture different perspectives among the study participants. Users primarily describe the *AiDitor* as a tool to enhance efficiency within editorial tasks, whereas management emphasizes the potential for cost and resource savings through increased efficiency.

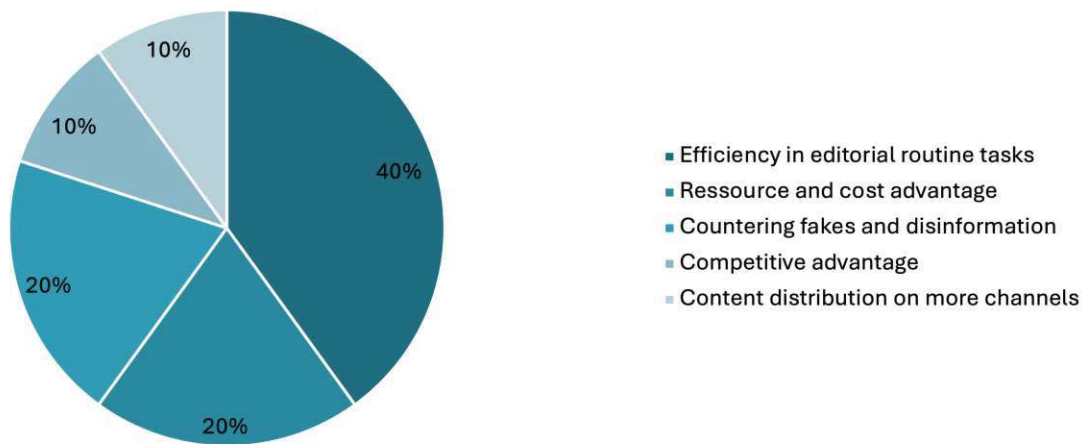


Figure 6.1: Opportunities of employing AI in the domain of public broadcasting.

Another opportunity mentioned by the participants is the potential to counter fakes and disinformation, which is also mentioned in the context of challenges. Participants do not provide further details or experiences but emphasize and are aware of the possibility of using AI to detect and counter fakes and disinformation. Additionally, the participants frequently mention the importance of gaining a competitive advantage through AI. This opportunity includes statements highlighting how employing trustworthy AI can strengthen the organization's market position by creating high-quality editorial content. Furthermore, participants are convinced that the organization, along with all employees and units, such as the established AI sounding board, aims to develop a trustworthy tool capable of providing the organization with a competitive advantage.

Finally, the participants mentioned the necessity of serving multiple platforms by creating and distributing content across various channels. Responses describe the shift in the media landscape, which has evolved from primarily broadcasting channels like radio and television as well as print media to the requirement of serving a wide range of social media and web channels. The *AiDitor* is seen as an opportunity to support these channels by offering functions such as creating a social media post from an article or producing a short radio report from an extensive television discussion.

Results illustrated in Figure 6.2 show that the greatest challenge of employing AI in public broadcasting is the thread of credibility to journalistic content mentioned by most of the participants. This challenge also comprises the possible mistrust in AI content and the issue of distinguishing between human and AI content of the consumers. Outlined by

statements such as "I can well imagine that the general public and, from my point of view, the public naturally finds it more difficult to differentiate between content. What comes from professional sources and what is simply generated? That is definitely a consequence that I am observing and fearing, that it will then become more difficult for the media to convey reliable content. Even more so than now, it is already, let us say, what Photoshop has started is now being reproduced in Dali". This challenge also includes the thread of different media organizations employing AI as a content slingshot, flooding different channels with unverified content, leading to a thread to the general credibility of journalistic work.

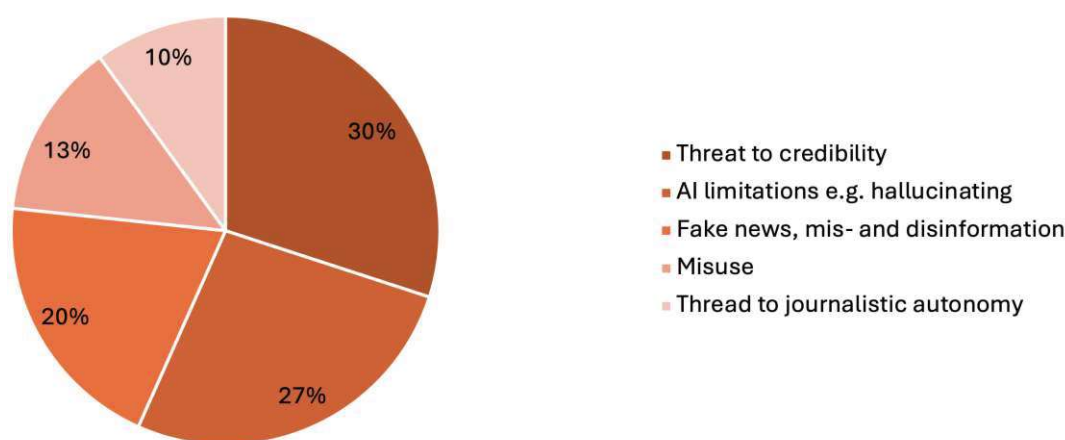


Figure 6.2: Challenges of employing AI in the domain of public broadcasting.

Predominantly mentioned by the users is the threat of AI limitations, such as various biases and hallucinations. These concerns are reinforced by user experiences, where they report witnessing the *AiDitor* getting facts wrong, such as mixing up names and misstating numbers. Another significant limitation of AI is the inability to trace how certain decisions are made and fully understand what data is used to generate the output. An additional challenge highlighted is the potential for AI to efficiently spread fake news, misinformation and disinformation. Given the public media's special mandate and responsibility in shaping public opinion and steering public discourse on different matters, participants view this as a significant threat to society.

The challenge of misuse is described as employing AI for journalistic tasks that are beyond the technology's limitations, such as conducting research or acquiring new knowledge on a topic. Study results show that this challenge is based on the novelty of the technology and missing guidelines, which should draw the line between what tasks the technology should and should not be used for. Another notable challenge is the thread to journalistic autonomy, defined by AI creating content and making relevant decisions. For instance, when condensing or summarizing content, the AI decides what is relevant by determining what information to include or discard, which may not align with a journalist's decisions.

## 6.5 Tensions between the ALTAI Requirements

Through the assessment, different tensions between the ALTAI requirements could have been identified. This section explores the tensions between the ALTAI requirements based on the study results.

### 6.5.1 Human Agency and Oversight and Technical Robustness

A tension occurs between the requirement of *Human Agency and Oversight* and *Technical Robustness*. *Human Agency and Oversight* is defined as implementing and employing AI systems that respect human control and decision-making capabilities. It involves designing AI systems that support human abilities, therefore supporting rather than replacing human judgment. The requirement is defined by a human maintaining agency, which means having the ability to make decisions, intervene in the process and understand the decision-making process of AI.

*Technical Robustness* refers to the ability of AI systems to perform reliable under various conditions. It involves designing AI tools to handle unexpected situations, minimize errors and operate without requiring constant human involvement. This also means that the AI tool must consistently perform the functions accurately and reliably and handle unexpected inputs or situations without failure.

The tension therefore arises because fulfilling *Technical Robustness* requires a high degree of automation. However, different levels of automation can conflict with the requirement of *Human Agency and Oversight*. An example would be condensing/summarizing content, as mentioned by one participant. For the function to perform reliably, the AI decides what information to include or discard and therefore conflicts with the requirement of *Human Agency and Oversight*.

To balance these two requirements, the development of the *AiDitor* should consider limiting the degree of automation when publishing and distributing journalistic content. This balance also involves designing the system to explain its decisions and actions to the users. Additionally, incorporating features that allow human intervention, such as an exit strategy or a stop button, is recommended to maintain *Human Agency and Oversight* without compromising technical robustness.

### 6.5.2 Privacy and Data Governance and Transparency

Another tension is identified between the requirement of *Privacy and Data Governance* and *Transparency*. The *Privacy and Data Governance* requirement aims to ensure that personal data is collected, stored and used responsibly. This means considering standards, such as the EU regulation of *GDPR*, ensuring trustworthy data management throughout the entire lifecycle of the AI tool. Key aspects of *Privacy and Data Governance* include data minimization, ensuring data is only used for its intended purposes and maintaining data security against breaches or unauthorized access. Data governance also involves establishing clear policies and procedures for data management, ensuring accountability

and providing mechanisms for data subjects to exercise their rights, such as deleting or access to their data.

The requirement of *Transparency* requires making the AI system's data sources, algorithms and decision-making processes visible and understandable to users and stakeholders. Moreover, the requirement involves explaining how AI systems work, the data used in their decision-making process and the potential impacts of their decisions. This is also crucial regarding accountability, hence holding developers and users responsible for their actions.

The tension between the two requirements arises because fulfilling *Transparency* often means revealing information about the data and processes used by AI systems, which can conflict with the *Privacy and Data Governance* requirement. For instance, providing detailed explanations of AI decision-making processes may require disclosing information about the data used, including personal data. This disclosure can risk compromising the privacy of individuals whose data is involved.

Balancing these requirements involves developing strategies that ensure *Transparency* while upholding *Privacy and Data Governance* standards. One approach mentioned by the participants is the anonymization of logs. As the tool is employed in a professional organizational setting, it is recommended to keep the possibility of tracing interactions with the *AiDitor* to a user even when the tool is fully rolled out. However, to successfully manage the tension, it is recommended to create guidelines for data access, usage and sharing.

### 6.5.3 Societal and Environmental Well-being and Diversity, Non-discrimination and Fairness

Between the requirement of *Societal and Environmental Well-being and Diversity, Non-discrimination and Fairness* a tension could be identified through the study. The requirement of *Societal and Environmental Well-being* aligns with the ethical principles described in Section 2.4 Media Ethics as *engaging in communities* and *minimizing harm*, where the focus lies on the broader impact of broadcasting content on society and democracy. This includes protecting democracy and the common good by raising awareness about social and environmental issues and protecting the vulnerable. Public broadcasters are responsible for producing content that informs and educates the public about political, social, economic, cultural and sports-related issues.

However, the requirement of *Diversity, Non-discrimination and Fairness* in public broadcasting means that AI systems should be designed to create inclusive content that represents a wide variety of voices and perspectives. This involves ensuring that no group is unfairly misrepresented or excluded and that journalistic content reflects the diversity of society, including different cultures, genders, ages and socio-economic backgrounds. It involves giving voice to underrepresented groups, ensuring equitable coverage of issues and avoiding biases in the broadcasted content.



The tension between promoting *Societal and Environmental Well-being* and ensuring *Diversity, Non-discrimination and Fairness* arises during the development by, for instance, implementing certain input/output filters, another example would be the usage of different prompts. Employing such mechanisms can lead to unintentionally misrepresenting or excluding certain groups or perspectives.

For example, a prompt aiming to prioritize content that highlights the urgency of climate action, which is essential for *Societal and Environmental Well-being*. However, if this content mostly reflects perspectives from certain groups, it may fail to represent the full spectrum of voices and experiences related to environmental issues. This can lead to perceptions of bias and exclusion, particularly if underrepresented communities feel their specific concerns and contributions to environmental sustainability are overlooked. Balancing these two requirements could be done by communicating the limitations of AI by for instance outlining the effect of prompt engineering.





# Evaluation and Recommendation

In this chapter, the study results are evaluated with a peer-assessment method by applying the ALTAI self-assessment list among experts in the field of AI. Based on the peer-assessment recommendations are provided as well as different key areas for action are identified. Moreover, the focus lies on analyzing the findings from the study and providing recommendations based on the results presented in Chapter 6.

## 7.1 Evaluation

The study results introduced in Chapter 6 are validated through peer-assessment. Six students of the *Technischen Universität Wien* were invited to discuss and evaluate the results. All six peers, visualized in Table 7.1, are either graduates or currently enrolled in a master's program, majoring in *Software Engineering and Internet Computing* or *Business Informatics*. The evaluation process comprised an initial workshop where the

Evaluator ID	Gender	Age	Major
01	M	24	Software Engineering and Internet Computing
02	M	24	Software Engineering and Internet Computing
03	M	26	Software Engineering and Internet Computing
04	F	27	Business Informatics
05	F	26	Business Informatics
06	M	29	Business Informatics

Table 7.1: Overview of study evaluators.

results were presented, followed by a question and answer session involving the author, the study assistant and the peers.

Afterward, the evaluators answered the ALTAI questions using the automated self-assessment tool developed by the HLEG. The evaluators were provided with Chapter 5 Research Design and Chapter 6 Results. As shown in Table 7.2 some results indicate a higher deviation than others, therefore a follow-up discussion was required to fully understand the evaluation of the peers. The evaluation results are accumulated and illustrated in Figure 7.1 and will be described comprehensively in this section.

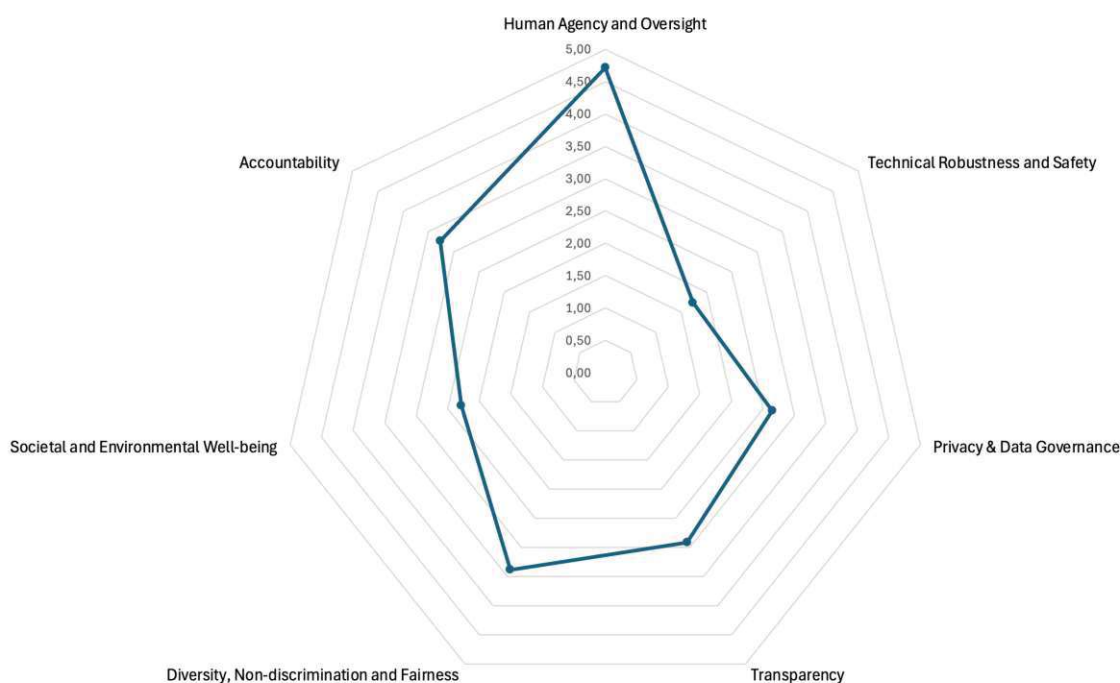


Figure 7.1: Accumulated results of the ALTAI self-assessment tool.

The results showed that the responses deviated regarding the requirement of *Transparency*. This deviation required a discussion during the peer debriefing to fully understand the evaluation result. The discussion revealed that the deviations were based on differing opinions regarding the use of third-party AI products. One peer claimed that using third-party AI providers makes the AI's output and decisions never fully traceable. In addition, they mentioned the risk of a possible bias in the training data, which would be reflected in the output of the *AiDitor*. Contrarily, others were convinced that the *Transparency* requirement is adequately fulfilled by employing prompt engineering combined with the *AiDitors* architecture, which includes the ORF's researched content as input data.

Eval. ID	Human Agency and Oversight	Technical Robustness and Safety	Privacy and Data Governance	Transparency	Diversity, Non-discrimination and Fairness	Societal and Environmental Well-being	Accountability
01	4,8	2,4	2	3,1	2,9	2,5	3,8
02	3,9	1,4	3	3,4	3,9	1,2	2,5
03	5	1,6	2,6	2,1	3	2,5	2,1
04	5	1,2	2,4	2,4	3,6	2,6	3,4
05	4,6	1,2	2,8	2,6	3,7	2,4	3,8
06	5	2,6	3,1	3,9	3,2	2,5	4
Average	4,72	1,73	2,65	2,92	3,38	2,28	3,27

Table 7.2: Results of the peer-assessment including the average score per ALTAI requirement.

A comparable range of deviations was found in the evaluation results of *Diversity, Non-discrimination and Fairness*, ranging from 2.9 to 3.9. The peer debriefing highlighted the same concern of using third-party AI providers, making it challenging to fully avoid reinforcing unfair biases due to limitations in understanding the algorithmic design and the parameters, data and performance of the AI models. Additionally, the peers mentioned that considering diversity in the training data, as suggested by the HLEG, depends on third-party AI providers. Others argued that a higher rating was justified given the current state of development and the group of users educated according to the journalistic code of ethics.

The current stage of development, along with reliance on organizational IT infrastructure, company policies and guidelines, led to higher deviations, ranging from 1.2 to 2.4, in evaluating the requirement for *Technical Robustness and Safety*. During the peer debriefing, some peers described reliance on organizational authorization and authentication as fully inadequate, while others found it sufficient. The majority highlighted their lack of knowledge about the organizational IT infrastructure, policies and guidelines, making it difficult to rate during the evaluation process.

Requirements with very low deviations, such as *Human Agency and Oversight* and *Societal and Environmental Well-being*, were only discussed briefly in the peer debriefing. The brief discussion was used to derive recommendations beyond the results of the ALTAI self-assessment list and therefore, considering the current stage of development and the

domain of public broadcasting. Some peers mentioned that the requirement for *Privacy and Data Governance* was also difficult to assess due to missing information in the study or the difficulty of mapping the ALTAI questions to the given context/information.

### 7.2 ALTAI Recommendations

The research question regarding how responsible integration and usage of genAI in the context of public service broadcasting can be ensured will be answered in this section by considering the empirically collected results and the peer evaluation. The recommendations are structured according to the seven requirements of the EU framework for trustworthy AI and refer to the recommendations given by the ALTAI self-assessment tool respecting the context of public broadcasting. Moreover, recommendations for the peer discussion are incorporated into the following section.

#### 7.2.1 Human Agency and Oversight

Recommendations to fulfill the requirement of *Human Agency and Oversight* include the establishment of a procedure to avoid end-users over-relying on the AI system, which can be addressed directly through the user and supported by technical and organizational measures. Technical measures aim to limit how AI tools can automate processes, ensuring critical decisions, such as publishing information, always require human intervention. In contrast organizational measures include the definition of roles and responsibilities of journalists and editors, which should be reflected in the authorization model of the *AiDitor*. Regular audits of the AI processes established in the organization should ensure compliance with guidelines and identify instances of over-reliance.

Based on the participants' responses, stating that they ignore the inline information, the recommendation would be to implement mandatory training during the entire lifecycle of the *AiDitor*. Furthermore, the assessment shows that the current exit strategy for processes is recommended to go beyond standard mechanisms such as the escape key or reloading the browser by, for instance, explicitly implementing the possibility to stop a process.

#### 7.2.2 Technical Robustness and Safety

Regarding *Technical Robustness and Safety*, it is recommended that a risk management process is implemented, including the assessment of risks, the definition of risk metrics and the determination of risk levels for the AI system in each specific use case. Additionally, it is recommended that possible threats to the AI system be identified, such as design faults, technical faults and environmental threats. Moreover, it is required to assess the dependency of the *AiDitor*, ensuring its stable and reliable behavior. This involves defining the technical or environmental factors that could lead to unstable and unreliable outputs. When the *AiDitor* is used in productive operations, fault tolerance should be considered, such as duplicate systems or another parallel system (either AI-based or

conventional). Additionally, creating a fall-back plan of the models and user profiles that can be restored in an emergency is recommended.

Another recommendation is to employ measures to ensure the quality of the data including the training data. Therefore, we recommend regularly updating third-party models and services to address the issue of data quality as well as possible system vulnerabilities. Furthermore, the risks associated with third-party AI technology should be assessed initially and with every update. An initial risk assessment as well as assessments during the entire system life-cycle must consider the different AI limitations, such as biases or hallucinations, including the creation of mitigation strategies.

Moreover, it is recommended to implement a process to monitor the *AiDitor's* output beyond *Human Oversight*. Responses indicate that the ORF plans to oversee the output with different AI models, which is one possible approach. Additionally, implementing procedures to calculate and handle different levels of the AI system's confidence score, which represents the likelihood that the output of an AI model is correct and will satisfy a user's request, is advisable. In simple terms, this score indicates how well the AI model *understood* the user's intent and managed to provide a satisfying response [Ton21].

### 7.2.3 Privacy and Data Governance

Responses indicate that even though no sensitive data is used during the development process, the rights to physical, mental and moral integrity and the right to data protection are currently addressed with pseudonymization of the logs. These measures should be enhanced to include mechanisms concerning the right to object and the right to be forgotten in the *AiDitor*. This can be achieved by allowing users to delete all or parts of their historical interactions. Additionally, it is recommended that users are informed about the system's data processing activities and their purpose, such as how and what data is collected, used, stored and shared. If user data is collected explicit consent should be obtained before collecting, processing or storing data. This consent must be freely given and revocable at any time.

It is recommended that the system is designed to empower users and give users control over their data, allowing them to manage their privacy settings and data-sharing preferences. This includes the option to deny the sharing of interaction data with the *AiDitor*. User training and awareness should include best practices for data protection and privacy. Moreover, consulting a data protection officer during the development process is recommended to align the *AiDitor* with relevant standards, such as ISO or IEEE or broadly adopted protocols for daily data management and governance. The data protection officer's responsibilities include managing regular audits to ensure compliance with data protection laws and internal policies.

### 7.2.4 Transparency

The study reveals that the ORF data, such as articles, radio reports or TV discussions, used as input for the *AiDitor* are explainable and traceable but the data used to train the third-party AI models is not. Therefore, it is recommended to develop guidelines that regulate the usage, by for instance describing different best practices. Even though the results show that journalists operate under a certain code of ethics, defining their routine regarding the usage of the tool, is essential. It is particularly important to inform users about the appropriate usage of the *AiDitor*, especially given the significant variations found in responses.

Developing a comprehensive training program for all stakeholders and potential users is recommended. This should include the technical aspects of using AI tools and best practices, as well as potential risks and limitations. Ensure that these training sessions are well-documented and easily accessible. While autodidact learning is valuable, it should be supplemented with official training to ensure all users have a consistent and in-depth understanding of the *AiDitor*. Moreover, it is recommended to explicitly communicate to the users they are interacting with a machine and that the decisions and suggestions are based on an algorithmic nature. Continuously survey the users to determine whether they understand and are satisfied with the decisions and suggestions of the *AiDitor*.

### 7.2.5 Diversity, Non-discrimination and Fairness

Since the tool orchestrates different AI services, establishing a strategy or a set of procedures to avoid creating or reinforcing unfair bias is limited and dependent on third-party technology providers. The recommendation is to consider the diversity and representativeness of all concerned stakeholders in the training data. For this reason, the recommendation would be to assess the third-party providers concerning diversity and non-discrimination, employ only models and develop functions that fulfill the ethical standard of public broadcasting. In addition, it is recommended that processes be assessed and put in place to test and monitor for potential biases during the entire lifecycle of the *AiDitor*. The thumbs-up and down feature within the tool is an already existing example of such a process using human feedback.

Moreover, the usage of different awareness and debiasing tools is recommended. One example would be the extensible open-source toolkit for detecting, understanding and mitigating algorithmic biases, *AI Fairness 360 (AIF360)*<sup>1</sup>, developed by IBM [BDH<sup>+</sup>19]. The *AIF360* provides state-of-the-art bias mitigation techniques, enabling developers to detect and reduce bias. The original *AIF360* Python package includes techniques from eight published papers within the broader algorithmic fairness community. It features over 71 metrics for detecting bias, nine algorithms for mitigating bias and a unique, comprehensive method for explaining metrics to help users understand the significance of the bias detection results.

---

<sup>1</sup><https://aif360.res.ibm.com/>

Additionally, the *AIF360* offers several realistic tutorial examples and notebooks that demonstrate the main features for industrial use and can be adapted to any domain. *AIF360* is the first system to combine bias metrics, bias mitigation algorithms, explanations of bias metrics and industrial usability in an open-source toolkit. A significant practical limitation of *AIF360* is that its algorithms for detecting and mitigating bias are designed specifically for binary classification problems and need to be adapted for multi-class and regression problems [BDH<sup>+</sup>19].

Other libraries, such as *Aequitas*<sup>2</sup> and *Local Interpretable Model-agnostic Explanations (LIME)*<sup>3</sup> offer effective metrics for more complex models but only detect bias without correcting it. *Aequitas* mainly aims to detect bias and visualize the results based on demographic groups. *LIME* aims to measure feature importance; in other words, it can identify which features are driving a particular prediction of the model [vRH22].

Despite the ALTAI recommendations emphasizing the importance of assessing whether AI system interfaces are designed with consideration for individuals with special needs or disabilities to mitigate the risk of exclusion, the author and evaluation participants decided on a different approach. The conclusion is that adhering strictly to standard mechanisms was sufficiently adequate within the specific domain and the organizational context to meet the requirement of fairness in system design. This decision was driven by the belief that standard mechanisms could effectively ensure inclusivity and accessibility without using mechanisms beyond the already established practices in the given context.

### 7.2.6 Societal and Environmental Well-being

Concerning the societal impact, it is strongly recommended that the potential positive and negative impacts of the *AiDitor* on the daily routine tasks of journalists be considered. Results reveal that the impact on work and skills is rarely addressed by the three participant groups, which can be based on the interview questions or indicate the missing awareness of the participants. Only a few responses from experts mentioned the possible effect of deskilling the workforce when fully employing the AI tool. An explanation for the few data collected concerning human resources can be based on the sampling of participants or the interview questions but could also indicate a lack of knowledge of the implications of AI on humans within the organization.

Hence, it is recommended that all stakeholders be assessed and informed of the possible impact of AI tools on work and skills. This can be done explicitly through initial training and information sessions while using the *AiDitor* as a supportive tool rather than automating journalistic tasks. To ensure that content is overseen by the user it is crucial to make publishing content not too convenient and easy. Furthermore, constantly assessing and developing *new* skills, such as prompt engineering and critical thinking, is recommended. Although the empirical study did not capture any fear among the workforce about being replaced by AI, internal employee development and training could

<sup>2</sup><https://www.aequitas-project.eu/>

<sup>3</sup><https://interpret.ml/docs/lime.html>



help address potential concerns. Define measures to ensure that the work impacts of the AI system are understood within the work processes, context and the whole socio-technical system.

Regarding the impact on *Environmental Well-being*, results show little to no concern among the participants about the *AiDitor*'s negative environmental impact. Hence, it is recommended to create an understanding of the environmental impact of AI in public broadcasting. The ALTAI recommendations go beyond creating an understanding and state that implementing mechanisms to measure and reduce the environmental impact of the AI tool throughout its entire lifecycle is essential. Due to the increasing interest in assessing and measuring the environmental impact, various approaches and methods have been introduced in the literature.

Dodge et al. have introduced a comprehensive method for measuring the carbon intensity of AI operations within cloud instances. Their framework is designed to assess operational carbon emissions by leveraging detailed, location-based and time-specific marginal emission data for each unit of energy consumed. This approach considers the variations in carbon emissions from using different energy sources and grid efficiencies across various geographical locations and times. This method shows the environmental impact of AI and helps optimize energy use and reduce emissions in cloud computing. It highlights the importance of considering location and time when considering energy consumption to make AI technologies more sustainable [DPTdC<sup>+</sup>22].

Another method to quantify the carbon emission of, for instance, the third-party services of machine learning is introduced by Lacoste et al., considering factors such as the location of the server, the energy grid it uses, the length of the training process as well as the hardware architecture used for the training. Even though many sources are available concerning the environmental impact, such as the carbon footprint or the water consumption of machine learning, there currently is no established standard to assess the environmental impact in practice. Also argued by different authors in literature is a lack of transparency concerning available data of organizations training the models [LLSD19].

Regarding the impact on society and democracy, the inductively developed categories introduced in Chapter 5 reveal an understanding of the potential effects of AI usage. However, it is recommended that the expected impact of AI within the organization be explicitly assessed and communicated to all stakeholders. This proactive approach will ensure that the users of the *AiDitor* understand the implications on society and democracy. Moreover, the author and evaluators of the study recommended actively engaging with the public to educate them about the *AiDitor*. Information broadcasted via articles, videos and social media posts can reach a wide audience and foster the public's ongoing education. This transparency builds trust between the public and the ORF using AI, ensuring that the technology is developed and deployed with the aim of aligning with social values and ethical standards. It also empowers individuals to engage in meaningful discussions and make informed decisions about AI, contributing to a more democratic discourse.



### 7.2.7 Accountability

Recommendations concerning *Accountability* would include establishing a risk management process for identifying, assessing, documenting and mitigating potential negative impacts of the *AiDitor*. This includes defining a risk management officer and training staff on risk management by creating a shared language about potential risks. Provide appropriate training for all individuals involved in the *AiDitor*'s development, deployment and oversight. This training should cover both the ethical principles and the legal frameworks applicable to the AI system, ensuring that all stakeholders know their responsibilities and the broader implications of AI technology, including creating a channel for all the stakeholders to report vulnerabilities, risks or bias in the *AiDitor*.

*Accountability* in the context of AI usage within public broadcasting involves establishing mechanisms to ensure that AI systems are developed with a preventative approach to risk, are reliable and behave as intended while minimizing unexpected harm. Hence, implementing a process involving regular audits is recommended to achieve *Accountability*. These audits should be conducted both internally and by independent external parties to evaluate the AI system's performance, development process and adherence to ethical guidelines.

External audits are crucial to ensure alignment with compliance standards, regulations and laws and provide an unbiased assessment of the AI tool's integrity and effectiveness. Consulting with various external partners, including AI developers, ethics boards and legal experts, can offer valuable insights and diverse perspectives, contributing to more robust accountability measures. These partners can help identify potential biases, risks and vulnerabilities in the AI system, ensuring it remains aligned with ethical standards and legal requirements. Therefore, the recommendation is to foster partnerships across borders with European public broadcasters.

## 7.3 Fields of Action

The ORF must comprehensively address the topic of AI in the near future and on various levels. The current developments in the field of genAI open up the possibility of automating cognitive routine tasks, thus creating significant potential for gaining efficiency in terms of time and costs. Competition with private competitors and the use of public funds are necessary to realize the potential for rationalization. In addition, the new technology also offers exciting perspectives concerning the individualization and regionalization of journalistic products and increasing accessibility and the quality of services. The ORF continues its path toward becoming a digital technology and media organization that aligns with ethical and legal standards. These potentials are counterbalanced by the risks associated with the use of genAI, which require careful risk assessment and the development of appropriate measures and strategies. Based on the ALTAI recommendations, three fields of action could have been identified by considering the AI tool and the context of public broadcasting.

### 7.3.1 Quality vs. Efficiency

The first field of action is finding the balance between quality and efficiency. On one side, high-quality standards for journalistic content are currently well-secured by professional, ethical standards. The high quality of journalistic content ensures and justifies society's trust in public service broadcasters. The results show that genAI could significantly contribute to expanding a high-quality offering by personalizing and regionalizing content and increasing accessibility. Moreover, by developing its genAI tool, the ORF has the opportunity to position itself as a role model for developing and using trustworthy AI in the given domain.

However, the ORF is not only competing with private media houses but also has to account to the public, with the introduction of a general household fee regarding the economic use of those fees and public funds. GenAI could contribute to the desired increase in efficiency and competitiveness by automating routine processes. Previous studies suggest efficiency gains can be realized, particularly by automating cognitive routine tasks and compensating for qualifications and experience using AI systems. This allows for personnel savings in terms of quality, for instance, with the fact that even less experienced and less qualified employees can take on qualified tasks with the help of appropriate AI tools and quantity, such as employees handling more tasks through the automation of routine tasks.

The central question will be whether the *AiDitor* can meet the quality standards and the desire for increased efficiency. Currently, the trade-off between quality and efficiency is not sufficiently considered. The risk assessment of the *AiDitor* shows a problematic dependence on external providers, which poses significant risks regarding the quality of generated content (bias, hallucination, etc.). It is assumed that users will check the quality of genAI outputs, placing the quality risk only on the users. Whether the expected savings can be realized this way is questionable. It is also unclear whether the current (pilot) version of the *AiDitor* would comply with the European AI regulation once it comes into effect.

For this reason, we recommend that strategic investment in a certifiable, self-developed AI system, in collaboration with other public service media houses, is essential. Training proprietary foundation models with their own data, on which high-quality media-specific applications can be developed, would be an important strategic investment in (data) authority, independence and quality.

A strategic collaboration at the European level with other public service media houses, particularly within the European Broadcasting Union, promises significant synergies and support from the European Commission. Developing an appropriate risk management system could partially resolve the trade-off between quality and automation, as more tasks can be delegated to high-quality AI without extensive post-checks.

### 7.3.2 Professional Ethical Standards and Digital Literacy

The second field of action is between professional, ethical standards and digital literacy regarding trust. Society's trust in public service broadcasting organizations is built on the high professional and ethical norms and standards of journalists. Socialization of these norms and values through journalistic training within the organization is of great importance. Along with formal norms, laws and organizational regulations, standards and guidelines form the social and ethical compass of journalistic work.

For this reason, we concluded that with AI tools, it will be necessary to develop these standards appropriately and create new standards for dealing with AI technology. The involvement of professional associations and educational institutions is crucial. A broad bottom-up process and adapting training curricula in digital literacy are required to accomplish this transformation.

Ultimately, the responsibility remains with humans, as shown by the assessment of the ALTAI requirements of for instance *Accountability*. Journalists and all the other users of the *AiDitor* can only fulfill this responsibility if they gain the necessary competencies and their actions align with their principles and values. Additionally, it will be necessary to mitigate or proactively counteract possible negative impacts of automation. This includes addressing issues such as automation bias and inattentiveness.

### 7.3.3 Trust Building and Educational Mandate

The role and trust of society in public broadcasters are central to our democratic and free coexistence. The AI challenges and limitations described in the literature as well as highlighted by various responses regarding the usage of genAI for individuals and society will pose significant challenges for society and organizations in the coming years. The ORF plays a crucial role in addressing these issues, particularly in terms of professional ethical standards and digital literacy.

Consequently, large-scale initiatives could strengthen digital literacy among the population through various formats and levels of reporting and media offerings for different population groups. The ORF can act as a role model, setting ethical guidelines in dealing with new technology and thereby continuing to strengthen public trust.



# Discussion

This chapter aims to discuss the study's limitations, such as participant sampling and reproducibility. Moreover, it describes the experiences of applying the *Z-Inspection* process combined with the HLEG trustworthy AI framework. This is followed by a brief conclusion, which sums up the most important findings of the case study. The last section outlines possible future research.

## 8.1 Limitations

One notable limitation of this study is the participant sampling. The backgrounds, experiences and opinions of the participants influence the results. The study aims to assess the challenges and opportunities of AI tools in public broadcasting. It is crucial to identify diverse participants to provide a comprehensive understanding of the phenomena and capture various perspectives. Due to the tool's novelty and the AI technology being employed in an organizational setting for daily editorial tasks, participants were required to have some experience with the tool and basic knowledge of AI technology. This requirement made finding enough participants challenging. For instance, the expert group comprised only two participants because only two individuals had developed the tool and could answer the interview questions. Additionally, sampling can be subjective and influenced by the participant's interests, which may lead to withholding information or ignoring results to avoid adverse outcomes. This risk can compromise the integrity of the research, leading to biased or incomplete conclusions that may misguide future studies and the recommendations provided in Chapter 7.

Furthermore, the risk of subjectivity may arise if participants provide responses they believe are expected rather than their true thoughts or behaviors. Future studies could incorporate additional data, such as evaluating interaction data from the *AiDitor* or analyzing error logs. Another limitation of this study is its reproducibility. Since the research concerns a specific AI tool tailored to the needs of ORF, the results may not be

generalizable to other public broadcasters. Future research could consider incorporating multiple tools employed by different media organizations to enhance the reproducibility of the findings.

## 8.2 Reflection on the Assessment Method

Regarding the questions on the self-assessment list for trustworthy AI by the HLEG of the EU, it was found that not all questions could be tailored to the domain of public broadcasting and the AI tool. The primary issue with tailoring the question considering the *AiDitor* was that the ORF is not training any models but rather using AI as a service. Hence, all the questions concerning the training of AI models, for instance, the training data, could not have been answered within the assessment. Based on that experience, the *Z-Inspection* process might apply in practice by offering well-defined steps, but fully mapping the findings to the ALTAI questions is impossible.

Another observation is that the recommendations in the self-assessment list are incredibly generic and focus more on what should be done rather than how to do it. This is exemplified by recommendations such as "Assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system" and regarding *Societal and Environmental Well-being*, "Define measures to reduce the environmental impact of your AI system's lifecycle and participate in competitions for the development of AI solutions that tackle this problem" [Koma]. These recommendations provide a broad idea of what employing AI in a trustworthy manner should look like, but they need more detailed guidance on practical implementation. We came to the conclusion that the assessment list requires experts to translate these recommendations into a specific action plan to ensure the trustworthy deployment of AI in a particular context.

Apart from the generic characteristics, which also make the assessment method applicable to all different domains, it enables the assessment of the implications and setting future goals of applying AI in the given domain. Therefore, it successfully derives opportunities and challenges regarding AI and guides trustworthy employment. We observed a higher initial effort by defining the participants and collecting and analyzing the data. The other steps, such as mapping the ALTAI question's findings using the online self-assessment tool, take less time and require less resource effort. The same applies to the evaluation of results and follow-up panel discussions. For this reason, we concluded that assessing trustworthy AI in practice involves a significant initial effort. However, the following steps require relatively low effort and resource intensity once the data is collected and analyzed.

## 8.3 Conclusion

The case study consists of qualitative interviews, according to Mayring, followed by a comprehensive assessment, the *Z-Inspection* process, to assess trustworthy AI in practice. The results are evaluated by a peer-assessment process, including follow-up

discussion, which served as the foundation for the recommendations according to the ALTAI requirements. Based on the study results, we identified opportunities and risks as well as derived recommendations for action in three areas: (i) quality vs. efficiency, (ii) professional ethical standards and digital literacy and (iii) trust-building and educational mandate.

The assessment shows that guidelines are essential to successfully managing disruptive technologies such as genAI. The key pillars for trustworthy AI usage include (i) European regulation, which establishes a normative framework for the use and dissemination of this technology, (ii) the development and adherence to technical standards (technical implementation) and (iii) the enhancement of professional ethical standards (based on self-regulation and voluntary commitment) to ensure the safe and ethical application of genAI technologies. While the first two pillars have already received considerable attention, the equally important third is often neglected. Therefore, the recommendations for action mainly focus on this third pillar.

## 8.4 Future Work

Further research could include scaling in size, incorporating additional broadcasting companies in the public and private sectors and using different AI tools, enhancing the study's reproducibility and theory development. Furthermore, developing a broadcasting standard method to assess the implications of AI in the given domain could be based on this thesis. The standard could employ the *Z-Inspection* process and the EU principles of trustworthy AI requirements only tailored to the broadcasting domain. The questions consider whether the tool orchestrates different services or if the organization is training AI models by itself. The research could also translate the more generic requirements into a practical, applicable list of actions for fully achieving trustworthy employment of AI.

Another area worth investigating is the long-term impact of AI tools such as the *AiDitor* on editorial staff's workflow and job satisfaction. This could involve a multi-year study tracking changes in productivity, job roles and employee satisfaction as AI becomes more integrated into their daily tasks. Understanding these dynamics would help in designing AI tools that support rather than automate tasks and hinder human autonomy. This would aim to ensure that technological advancements are employed in a trustworthy manner, especially respecting the requirement of *Societal and Environmental Well-being*.

# List of Figures

1.1	Classification of global risks rated by the <i>World Economy Forum</i> [For21].	2
1.2	Ethical challenges with significant negative impact through <i>ChatGPT</i> [SE24].	3
3.1	Ethics guideline of trustworthy AI introduced by the HLEG of the EU (2019) [HLEGoAIEC19]. . . . .	23
3.2	Overview of the three main phases of the <i>Z-Inspection</i> process [ZBB+21].	26
5.1	Visualization of the tailored <i>Z-Inspection</i> process that is used within thesis.	33
5.2	Different functions of the <i>AiDitor</i> . . . . .	34
5.3	Illustration of the architecture, idea and possible distribution channels of the AI-based <i>AiDitor</i> . . . . .	35
6.1	Opportunities of employing AI in the domain of public broadcasting. . . .	61
6.2	Challenges of employing AI in the domain of public broadcasting. . . . .	62
7.1	Accumulated results of the ALTAI self-assessment tool. . . . .	68



# List of Tables

2.1	Summary of AI initiatives and their objectives.	12
2.2	Characterization of different types of AI bias.	15
5.1	Information of the study participants and the interview duration.	36
5.2	Deductive code schema based on ALTAI list, including the sub-codes.	38
7.1	Overview of study evaluators.	67
7.2	Results of the peer-assessment including the average score per ALTAI require-	
	ment.	69



# Bibliography

- [Act24] EU Cybersecurity Act. The eu cybersecurity act, 2024. Access: 2024/06/18.
- [AIG24] AIGO. Oecd working party on artificial intelligence governance (aigo), 2024. Access: 2024/05/24.
- [ALMK22] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.
- [Ama23] Thimira Amaratunga. *Understanding Large Language Models : Learning Their Underlying Concepts and Technologies*. Apress Media LLC, Berkeley, CA, first edition. edition, 2023.
- [Bai24] Christopher A. Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.
- [BDH<sup>+</sup>19] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Nate-san Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.
- [BEA20] Banu Buruk, Perihan Elif Ekmekci, and Berna Arda. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, 23(3):387–399, 2020.
- [Bel20] Haydn Belfield. Activism by the ai community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 15–21, 2020.
- [BG21] Bertrand Braunschweig and Malik Ghallab. Trustworthy ai. In *Reflections on Artificial Intelligence for Humanity*, volume 12600 of *Lecture Notes in Computer Science*, pages 13–39. Springer International Publishing AG, Switzerland, 2021.

- [BGKR24] Maria Teresa Baldassarre, Domenico Gigante, Marcos Kalinowski, and Azzurra Ragone. Polaris: A framework to guide the development of trustworthy ai systems. *arXiv preprint arXiv:2402.05340*, 2024.
- [BGM87] Izak Benbasat, David K Goldstein, and Melissa Mead. The case research strategy in studies of information systems. *MIS quarterly*, pages 369–386, 1987.
- [BGMMS21] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- [Bin16] Tas Bindi. Amazon, google, facebook, ibm, and microsoft form ai non-profit, 2016. Access: 2024/05/24.
- [Bro20] Fred Brown. *Media ethics: A guide for professional conduct*. Society of Professional Journalists Indianapolis, IN, 2020.
- [Can24] Thomas Cantens. How will the state think with chatgpt? the challenges of generative artificial intelligence for public administrations. *AI & SOCIETY*, pages 1–12, 2024.
- [CCKW16] Kathy Brittain Richardson Clifford Christians, Mark Fackler, Peggy Kreshel, and Robert Woods. *Media Ethics: Cases and Moral Reasoning, Ninth Edition*. Routledge, 2016.
- [CDGfCNT20] European Commission, Content Directorate-General for Communications Networks, and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office, 2020.
- [CF24] Eun Cheol Choi and Emilio Ferrara. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1441–1449, 2024.
- [Chr21] FH-Prof. Dr. Reinhard Christl. Digitale transformation, vom broadcaster zum qualitätsnetzwerk. 1. Auflage, *Österreichischer Rundfunk, ORF*, 2021.
- [Cla23] Laurie Clarke. Call for ai pause highlights potential dangers. *Science*, 380(6641):120–121, 2023.
- [Cop24] B.J. Copeland. artificial intelligence, 2024. Access: 2024/01/18.
- [Des16] Ethically Aligned Design. Ieee global initiative on ethics of autonomous and intelligent systems, 2016. Access: 2024/05/24.

- [Dica] Cambridge Dictionary. intelligence. Access: 2024/05/08.
- [Dicb] Cambridge Dictionary. Meaning of ethical. Access: 2024/04/28.
- [DPTdC<sup>+</sup>22] Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. Measuring the carbon intensity of ai in cloud instances. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1877–1894, 2022.
- [EB16] Wolfgang Ertel and Nathanael T Black. *Grundkurs künstliche intelligenz*, volume 4. Springer, 2016.
- [EMMR23] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.
- [Fer23] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.
- [Fer24a] Emilio Ferrara. The butterfly effect in artificial intelligence systems: Implications for ai bias and fairness. *Machine Learning with Applications*, 15:100525, 2024.
- [Fer24b] Emilio Ferrara. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, pages 1–21, 2024.
- [For21] World Economic Forum. The global risk report 2024, 19th edition, insight report, 2021. Access: 2024/01/18.
- [GKM<sup>+</sup>12] A David Gordon, John Michael Kittross, John C Merrill, William Babcock, and Michael Dorsher. *Controversies in media ethics*. Routledge, 2012.
- [GTC<sup>+</sup>23] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *The British Journal of Radiology*, 96(1150):20230023, 2023.
- [Hau89] John Haugeland. *Artificial intelligence: The very idea*. MIT press, 1989.
- [HLEGoAIEC19] B-1049 Brussels High-Level Expert Group on Artificial Intelligence European Commission. Ethics guidelines for trustworthy ai, 2019. Access: 2024/07/20.

- [HP21] Eleanore Hickman and Martin Petrin. Trustworthy ai and corporate governance: the eu's ethics guidelines for trustworthy artificial intelligence from a company law perspective. *European Business Organization Law Review*, 22:593–625, 2021.
- [HWJ20] Joanne Hinds, Emma J Williams, and Adam N Joinson. “it wouldn't happen to me”: Privacy concerns and perspectives following the cambridge analytica scandal. *International Journal of Human-Computer Studies*, 143, 2020.
- [Kie02] Matthew Kieran. *Media ethics*. Routledge, 2002.
- [KM23] Österreichischer Rundfunk Konrad Mitschka. Quality dimensions and 18 performance categories frame, how the orf fulfills its public service mandate., 2023. Access: 2024/05/07.
- [Koma] Europäische Kommission. The assessment list for trustworthy artificial intelligence. Access: 2024/05/24.
- [Komb] Europäische Kommission. Ethics guidelines for trustworthy ai. Access: 2024/05/24.
- [Kre21] Andrian Kreye. Künstliche intelligenz - die rote linie, 2021. Access: 2024/01/18.
- [Kuc18] Udo Kuckartz. *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung*. Beltz Juventa, Weinheim/Basel, 2018.
- [KUD21] Davinder Kaur, Suleyman Uslu, and Arjan Duresi. Requirements for trustworthy artificial intelligence – a review. In Leonard Barolli, Kin Fun Li, Tomoya Enokido, and Makoto Takizawa, editors, *Advances in Networked-Based Information Systems*, pages 105–115, Cham, 2021. Springer International Publishing.
- [LFCP22] Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 97–106, 2022.
- [LLSD19] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.
- [LS23] Tama Leaver and Suzanne Srdarov. Chatgpt isn't magic: The hype and hypocrisy of generative artificial intelligence (ai) rhetoric. *M/C Journal*, 26(5), 2023.

- [LZWX24] Qinghua Lu, Liming Zhu, Jon Whittle, and Xiwei Xu. *Responsible AI : best practices for creating trustworthy AI systems*. Addison-Wesley, Boston, [first edition]. edition, 2024.
- [MA24] Michael McTear and Marina Ashurkina. *Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents*. Apress L. P, Berkeley, CA, 1 edition, 2024.
- [MBMK<sup>+</sup>23] William Marcellino, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, and Jackson Smith. *The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0: Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI*. RAND Corporation, Santa Monica, CA, 2023.
- [MF19] Philipp Mayring and Thomas Fenzl. *Qualitative inhaltsanalyse*. Springer, 2019.
- [MM15] Kåre Moen and Anne-Lise Middelthon. Chapter 10 - qualitative research methods. In Petter Laake, Haakon Breien Benestad, and Bjorn Reino Olsen, editors, *Research in Medical and Biological Sciences (Second Edition)*, pages 321–378. Academic Press, Amsterdam, second edition edition, 2015.
- [MMRS06] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12–12, 2006.
- [MRC<sup>+</sup>23] Riccardo Mariani, Francesca Rossi, Rita Cucchiara, Marco Pavone, Barnaby Simkin, Ansgar Koene, and Jochen Papenbrock. Trustworthy ai—part 1. *Computer*, 56(2):14–18, 2023.
- [MWMG05] Natasha Mack, Cynthia Woodsong, Kathleen M MacQueen, and Greg Guest. *Qualitative research methods*. Family Health International, 2005.
- [NAC05] Joanne Neale, Debby Allen, and Lindsey Coombes. Qualitative research methods within the addictions. *Addiction (Abingdon, England)*, 100(11):1584–1593, 2005.
- [Neu19] Christoph Neuberger. Öffentlich-rechtlicher rundfunk und qualitätsdiskurs. *Media Perspektiven, ARD-MEDIA GmbH*, 2019.
- [oL17] Future of Life. Asilomar ai principles, 2017. Access: 2024/05/24.
- [O’L23] Daniel E O’Leary. Enterprise large language models: Knowledge characteristics, risks, and organizational activities, 2023.

- [PH10] Sigit Pranawa and Raheili Humsona. Die vierte säule der demokratie: Rolle der medien in der sozialen bewegung. *südostasien-Zeitschrift für Politik • Kultur • Dialog*, 26(1), 2010.
- [Pla14] Patrick Lee Plaisance. Media ethics. *Los Angeles: Sage. Preston, C.(2005). Advertising to children and social responsibility. Young Consumers: Insight and Ideas for Responsible Marketers*, 6:61–67, 2014.
- [PSL20] Reza Arkan Partadiredja, Carlos Entrena Serrano, and Davor Ljubenkovic. Ai or human: the socio-ethical implications of ai-generated media content. In *2020 13th CMI Conference on Cybersecurity and Privacy (CMI)-Digital Transformation-Potentials and Challenges (51275)*, pages 1–6. IEEE, 2020.
- [R90] Kurzweil R. *The age of intelligent machines*. MIT press Cambridge, 1990.
- [RGR<sup>+</sup>23] Jyri Rajamäki, Fotios Gioulekas, Pedro Alfonso Lebre Rocha, Xavier del Toro Garcia, Paulinus Ofem, and Jaakko Tyni. Altai tool for assessing ai-based technologies: Lessons learned and recommendations from shapes pilots. *Healthcare*, 11(10), 2023.
- [RL<sup>+</sup>19] Patrizia Ribino, Carmelo Lodato, et al. A norm compliance approach for open and goal-directed intelligent systems. *Complexity*, 2019.
- [RM13] Tom Rosenstiel and Kelly McBride. The new ethics of journalism: Principles for the 21st century. *Los Angeles, Estados Unidos da América: The Poynter Institute*, 2013.
- [RRW23] Charles Radclyffe, Mafalda Ribeiro, and Robert H. Wortham. The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*, 6, 2023.
- [RWA<sup>+</sup>19] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. Better language models and their implications. *OpenAI blog*, 1(2), 2019.
- [SE24] Bernd Carsten Stahl and Damian Eke. The ethics of chatgpt – exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74:102700, 2024.
- [SNLP22] Anton Sigfrids, Mika Nieminen, Jaana Leikas, and Pietari Pikkuaaho. How should public administrations foster the ethical development and use of artificial intelligence? a review of proposals for developing governance of ai. *Frontiers in Human Dynamics*, 4, 2022.



- [SWS<sup>+</sup>24] Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. Smiling women pitching down: auditing representational and presentational gender biases in image-generative ai. *Journal of Computer-Mediated Communication*, 29(1):zmad045, 2024.
- [Tan87] Steven L Tanimoto. *The elements of artificial intelligence: Using common LISP. Principles of computer science series*. New York: Computer Science Press, Inc., 1987.
- [TFW<sup>+</sup>23] Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101, 2023.
- [Ton21] Guy Tonye. Machine learning confidence scores — all you need to know as a conversation designer, 2021. Access: 2024/07/20.
- [VAB<sup>+</sup>23] Dennis Vetter, Julia Amann, Frédérick Bruneault, Megan Coffee, Boris Düdder, Alessio Gallucci, Thomas Krendl Gilbert, Thilo Hagendorff, Irmhild van Halem, Eleanore Hickman, et al. Lessons learned from assessing trustworthy ai in practice. *Digital Society*, 2(3):35, 2023.
- [vRH22] W. van Zetten, G.J. Ramackers, and H.H. Hoos. Increasing trust and fairness in machine learning applications within the mortgage industry. *Machine Learning with Applications*, 10:100406, 2022.
- [Wan19] Pei Wang. On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37, 2019.
- [Wie01] Verlagspostamt 1030 Wien. Bundesgesetzblatt für die republik Österreich. 83. Bundesgesetz, mit dem das Bundesgesetz über die Aufgaben und die Einrichtung des Österreichischen Rundfunks (Rundfunkgesetz – RFG) und das Arbeitsverfassungsgesetz 1974 geändert werden, 2001.
- [WPLK23] Richard Watermeyer, Lawrie Phipps, Donna Lanclos, and Cathryn Knight. Generative ai and the automating of academia. *Postdigital Science and Education*, pages 1–21, 2023.
- [ZBB<sup>+</sup>21] Roberto V. Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, Naveed Mushtaq, Gemma Roig, Norman Stürtz, Karsten Tolle, Jesmin Jahan Tithi, Irmhild van Halem, and Magnus Westerlund. Z-inspection®: A process to assess trustworthy ai. *IEEE Transactions on Technology and Society*, 2(2):83–97, 2021.

- [ZK22] Xiaoming Zhai and Joseph Krajcik. Pseudo ai bias. *arXiv preprint arXiv:2210.08141*, 2022.

## Interview Questions

Die Fragen soll die Einstellung und das Befinden gegenüber KI-Tools im öffentlich-rechtlichen Rundfunk berücksichtigen. Das Ziel ist es, die Standpunkte, Einstellungen, Meinungen, Ängste und Chancen der Teilnehmer zu erfassen. Dabei können die folgenden Fragen bei der Analyse behilflich sein:

1. Nutzen Sie KI privat? Was denken Sie über KI?
2. Haben Sie davon gehört, dass der ORF eine KI entwickelt?
3. Wie würden Sie Ihre Einstellung gegenüber der Verwendung von KI-Tools im öffentlich-rechtlichen Rundfunk beschreiben?
4. Haben Sie Ängste und Bedenken in Bezug auf KI im öffentlich-rechtlichen Rundfunk?

### A.1 Stakeholder

#### A.1.1 Allgemeine Fragen

Der ORF, als öffentlich-rechtliches Medienunternehmen, entwickelt derzeit seine eigene generative KI, den AiDitor. Haben Sie von diesem Projekt gehört?

Welche Erwartungen haben Sie an den AiDitor?

Was würden Sie sich vom ORF und seinen Mitarbeiter\_innen in Bezug auf den Einsatz von KI-Systemen wünschen?

Welche Chancen sehen Sie, wenn der ORF in Zukunft KI einsetzt?

Sehen Sie Gefahren bzw. haben Sie Bedenken, falls der ORF in Zukunft auf KI setzt? Falls ja, welche?

Worauf müssen/sollen Journalisten/Journalistinnen in Zukunft achten, wenn sie mit der KI arbeiten?

### A.1.2 Publikum (Publikumsrat)

Fühlen Sie sich ausreichend informiert zum Thema KI durch den ORF?

Wie schätzen Sie die Rolle des ORF ein? Auch im Vergleich zu anderen Medienunternehmen?

Glauben Sie, dass KI die Medienlandschaft verändern wird? Die Rolle und Aufgaben des ORF verändern wird? Wenn ja wie?

Was ist Ihnen als Kund\_in wichtig? Welche Hoffnungen/Befürchtungen haben Sie in Bezug auf den Einsatz von KI?

Haben Sie sich mit dem Problem "Fake News oder Deepfakes" auseinandergesetzt? Vertrauen Sie dem ORF, Fake News/Deepfakes zu erkennen und auszufiltern? Im Vergleich zu anderen Medien mehr/gleich/weniger? Warum?

Haben Sie Vertrauen in den ORF als Informationsverbreitendes Medienunternehmen?

Würde sich ihr Vertrauen verändern falls Journalist\_innen KI einsetzen würden?

### A.1.3 HR/Betriebsrat

Der ORF entwickelt ein KI-Tool für den redaktionellen Bereich, den AiDitor. Kennen Sie das Projekt? Was wissen Sie darüber?

Waren Sie in die Entwicklung des Tools miteinbezogen?

Welche Erwartungen haben Sie bzw. Chancen sehen Sie in Bezug auf dieses Tool?

Was ist wichtig für die Implementierung dieses Tools?

Welche Kompetenzen (bei Journalist\_innen) werden in Zukunft Ihrer Meinung nach wichtiger in der Zukunft? Welche weniger?

Wird sich das Berufsbild und Aufgabengebiet von Journalist\_innen ändern? Wenn ja wie?

Wird durch KI die Jobsicherheit der Mitarbeiter ein Thema?

### A.1.4 Vorstand/Räte/Management

Was erwarten Sie sich durch den Einsatz des Tools AiDitors?

Welche Chancen sehen Sie für den ORF?

Welche Herausforderungen sehen Sie in Bezug auf den öffentlich-rechtlichen Rundfunk und KI zukommen?

Welche Rolle soll in Zukunft der ORF einnehmen?

Welche Bedeutung wird in Zukunft die Zusammenarbeit mit anderen öffentlich-rechtlichen Rundfunkunternehmen in Europa haben?

Aktuell ist der AiDitor ein Pilotprojekt. Wie ist die weitere Ausrollung geplant? Welche Maßnahmen haben Sie geplant?

## A.2 Experten

### A.2.1 Allgemeine Fragen

Beschreiben Sie kurz Ihre Rolle beim ORF.

Wie würden Sie den AiDitor beschreiben?

Welche Probleme sollen durch das Tool gelöst werden?

### A.2.2 Menschliche Handlungsfähigkeit und Aufsicht

Inwieweit ist der AiDitor darauf ausgelegt mit den Benutzer\_innen zu interagieren, sie anzuleiten oder Entscheidungen zu treffen?

Wie wird den Benutzer\_innen bewusst gemacht, dass Entscheidungen, Ratschläge und Ergebnisse algorithmisch erzeugt wurden?

Haben Sie Verfahren eingeführt, um zu verhindern, dass sich Benutzer\_innen zu sehr auf den Output und die Funktionalitäten des AiDitor verlassen?

Wie bewerten sie die Möglichkeit, dass Entscheidungsprozesse der Benutzer\_innen auf eine unbeabsichtigte und unerwünschte Weise beeinflusst wird? Haben Sie Maßnahmen ergriffen, um das Manipulationsrisiko zu mindern?

Bestimmen Sie den Grad der Automatisierung:

1. Automatisiertes System
2. Human-in-the-Loop
3. Human-on-the-Loop
4. Human-in-command

Wie wurden die Benutzer\_innen bezüglich Beaufsichtigung/Überwachung geschult?

Gibt es eine Möglichkeit einzelne Funktionen oder Verfahren abubrechen/auszuschalten („Stopp-Taste“)?

Besteht die Möglichkeit das "Learning" rückgängig zu machen? Haben Sie eine Backup/Versionierungs-Strategie?

Welche Erkennungs- und Reaktionsmechanismen haben Sie eingerichtet, um festzustellen, ob etwas Ungeplantes passiert?

Haben Sie Überwachungs- und Kontrollmaßnahmen implementiert, um das Selbst-lernen oder den Grad der Autonomie des KI-Systems zu bewerten?

### A.2.3 Technische Robustheit und Sicherheit

Wie wurden die IT-Security Ziele wie, Vertraulichkeit, Integrität und Verfügbarkeit berücksichtigt?

Haben Sie verschiedene Arten von Schwachstellen berücksichtigt, z.B. Datenverschmutzung, physische Infrastruktur, Cyberangriffe und diese bewertet?

Ist der AiDitor für Cybersicherheit zertifiziert oder entspricht es bestimmten Sicherheitsstandards (z.B. TRUSTED AI by TÜV AUSTRIA)

Welcher Schaden entsteht im Falle von Risiken und Bedrohungen (Design- oder technischen Fehlern, Mängeln, Ausfällen, Angriffen) für den ORF?

Haben Sie Maßnahmen ergriffen, um die Integrität, Robustheit und Gesamtsicherheit des KI-Systems gegen potenzielle Angriffe über seinen Lebenszyklus hinweg zu gewährleisten?

Haben Sie überprüft, wie sich Ihr System in unerwarteten Situationen und Umgebungen verhält? Wurde der AiDitor einem Pen-Test unterzogen?

Falls vorhanden, beschreiben Sie Ihre Update-Strategie.

Haben Sie einen Prozess implementiert, um Risiken kontinuierlich zu messen und zu bewerten? Welche Metriken werden für die Risikobewertung verwendet?

Haben Sie das Risiko einer möglichen böswilligen, missbräuchlichen oder unangemessenen Nutzung des AiDitors eingeschätzt und bewertet?

Gibt es ein Backup System, dass im Falle von Problemen eingesetzt werden kann?

Haben Sie Maßnahmen ergriffen, um sicherzustellen, dass die verwendeten Daten für das KI-System aktuell, von hoher Qualität und repräsentativ sind?

Könnte eine geringe Genauigkeit des AiDitors kritische, kontroverse oder schädliche Folgen haben?

Wird die Genauigkeit des AiDitors gemessen, überwacht und dokumentiert?

Werden Benutzer\_innen über die Genauigkeit der Inhalte des AiDitors informiert?

Haben Sie Verfahren implementiert, um die Zuverlässigkeit des AiDitors zu überwachen und sicherzustellen?

Haben Sie Verfahren implementiert, um die Reproduzierbarkeit des AiDitors zu überwachen und sicherzustellen?

Haben Sie potenzielle Folgen in Betracht gezogen, wenn das KI-System neuartige oder ungewöhnliche Methoden erlernt?

### A.2.4 Privatsphäre und Data Governance

Wurden bei der Entwicklung personenbezogene Daten verwendet?

Wie würden Sie den Einfluss des KI-Systems in Bezug auf Datenschutz und Privatsphäre bewerten?

Wird die Privatsphäre der Benutzer\_innen geschützt?

Welche Möglichkeiten wurden implementiert, um Probleme im Zusammenhang mit Privatsphäre zu überwachen?

Wurden Maßnahmen gemäß der Datenschutz-Grundverordnung (DSGVO) berücksichtigt?

Haben Sie einen Datenschutzbeauftragten bestimmt?

Haben Sie sich mit Standards bezüglich Datenmanagement und Governance beschäftigt und gegebenenfalls implementiert?

### A.2.5 Transparenz

Wie wird die Qualität der Daten im gesamten Lebenszyklus (Entwicklung, Testphase, Go-live) des AiDitor sichergestellt?

Können Sie nachvollziehen, welche Daten/Entscheidungen vom KI-System verwendet wurden, um eine bestimmte Entscheidung/Empfehlung zu treffen?

Welche Maßnahmen haben Sie ergriffen, um die Qualität der Ergebnisse des KI-Systems kontinuierlich zu bewerten?

Werden Entscheidungen/Empfehlungen des AiDitors aufgezeichnet?

Inwieweit werden KI generierte Inhalte bei der Publizierung gekennzeichnet?

Haben Sie den Benutzern die technischen Einschränkungen und potenziellen Risiken des KI-Systems kommuniziert, wie beispielsweise die Genauigkeit und/oder Fehlerquoten?

Inwieweit wurden die Benutzer über die ordnungsgemäße Verwendung informiert/geschult?

Haben Sie Schulungs- und Informationsmaterial den Benutzern zur Verfügung gestellt?

### A.2.6 Diversität, Nichtdiskriminierung und Fairness

Inwieweit sind Sie bezüglich Bias im AiDitor sensibilisiert?

Welche Strategien wurden bei der Entwicklung verfolgt, um Bias des AiDitors zu vermeiden?

Haben Sie die Vielfalt und Repräsentativität aller Stakeholder in den Trainings- sowie Testdaten berücksichtigt?

Wie behandeln sie Vorkommnisse und/oder Probleme in Bezug auf Bias?

Gibt es bezüglich Fairness (gerechte Verwendung des AiDitors) Anforderungen beim ORF und inwieweit wurden diese bei der Entwicklung berücksichtigt?

### A.2.7 Gesellschaftliches und ökologisches Wohlergehen

Könnten durch den AiDitor generierte Inhalte negative Auswirkungen auf die Gesellschaft oder die Demokratie haben?

Haben Sie die gesellschaftlichen Auswirkungen der Nutzung des AiDitor berücksichtigt und bewertet?

Haben Sie sich bezüglich der Umweltauswirkungen des AiDitors Gedanken gemacht?

Bestehen negative Auswirkungen des KI-Systems auf die Umwelt? Wenn ja, welche potenziellen Auswirkungen identifizieren Sie?

Haben Sie Mechanismen etabliert, um die Umweltauswirkungen der Entwicklung, Bereitstellung und/oder Nutzung des KI-Systems zu bewerten (z.B. Energieverbrauch)?

### A.2.8 Verantwortlichkeit

Wurde der AiDitor von unabhängigen Dritten geprüft?

Haben Sie (bis auf die TU Studie) eine externe Beratung oder ein Auditverfahren durch Dritte vorgesehen, um ethische Bedenken und Maßnahmen zu überwachen hinsichtlich einer Rechenschaftspflicht?

Haben Sie Schulungen zu Risiken organisiert, und wenn ja, informieren diese auch über das potenziell geltende rechtliche Rahmenwerk?

Haben Sie einen Prozess für die Nutzer etabliert um potenzielle Schwachstellen, Risiken oder Biases des AiDitor zu melden?

### A.2.9 Abschlussfragen

Habe ich etwas vergessen zu Fragen?

Möchten sie abschließend noch etwas sagen?

## A.3 Users

### A.3.1 Allgemeine Fragen

Beschreiben Sie kurz Ihre Rolle beim ORF.

Wie würden Sie Ihre Einstellung gegenüber der Verwendung von KI-Tools im öffentlich-rechtlichen Rundfunk beschreiben?

Haben Sie Ängste und Bedenken haben Sie in Bezug auf KI im öffentlich-rechtlichen Rundfunk?



### A.3.2 Menschliche Handlungsfähigkeit und Aufsicht

Wie ist die Benutzer Interaktion mit dem AiDitor gestaltet? Gibt es Funktionen, die Sie bei der Entscheidungsfindung unterstützen?

Inwieweit ist ihnen bewusst, dass Entscheidungen algorithmischer Natur sind?

Wie viel Vertrauen haben Sie in die Richtigkeit der Ergebnisse, die der AiDitor liefert?

Welche Funktionen des AiDitors nutzen Sie?

Überprüfen Sie die Inhalte des AiDitors bevor Sie diese verwenden? Wenn ja: Worauf und wie?

### A.3.3 Technische Robustheit und Sicherheit

Was machen Sie beim Auftreten von Fehlern oder unerwartetem Verhalten des AiDitors?

### A.3.4 Privatsphäre und Data Governance

Wissen Sie, ob personenbezogene Daten verwendet werden?

Wenn ja, wurde dafür eine Einwilligung zur Verwendung dieser von Ihnen geholt?

### A.3.5 Transparenz

Wurden Sie bezüglich der ordnungsgemäßen Verwendung des AiDitors informiert?

Wurden Sie über die technischen Einschränkungen und potenziellen Risiken informiert?

Welche Unterlagen (Schulung; Information ...) haben sie bekommen?

Können Sie die Quellen der generierten Inhalte des AiDitors nachvollziehen?

### A.3.6 Diversität, Nichtdiskriminierung und Fairness

Wurden Sie bezüglich Bias sensibilisiert?

Welche Möglichkeiten verwenden Sie, um Bias in den generierten Inhalten zu erkennen?

Beschreiben Sie Ihre Interessen in Bezug auf Diversität, Nichtdiskriminierung und Fairness.

### A.3.7 Gesellschaftliches und ökologisches Wohlergehen

Welche Auswirkungen auf die Gesellschaft können sie sich in Bezug auf die generierten Inhalte vorstellen?

Haben Sie sich bezüglich Umweltauswirkungen Gedanken gemacht?

Unabhängigen Schätzungen zufolge kostete alleine das Training des GPT-3 rund 1.300 Megawattstunden, was einem Ausstoß von 550 Tonnen (CO<sub>2</sub>) Kohlendioxid entspricht.

### A.3.8 Verantwortlichkeit

Besteht die Möglichkeit potenzielle Schwachstellen und/oder Risiken zu melden?

Wem und wie melden Sie potenzielle Schwachstellen oder Risiken?

### A.3.9 Abschlussfragen

Habe ich etwas vergessen zu Fragen?

Möchten sie abschließend noch etwas sagen?