



AALBORG  
UNIVERSITET

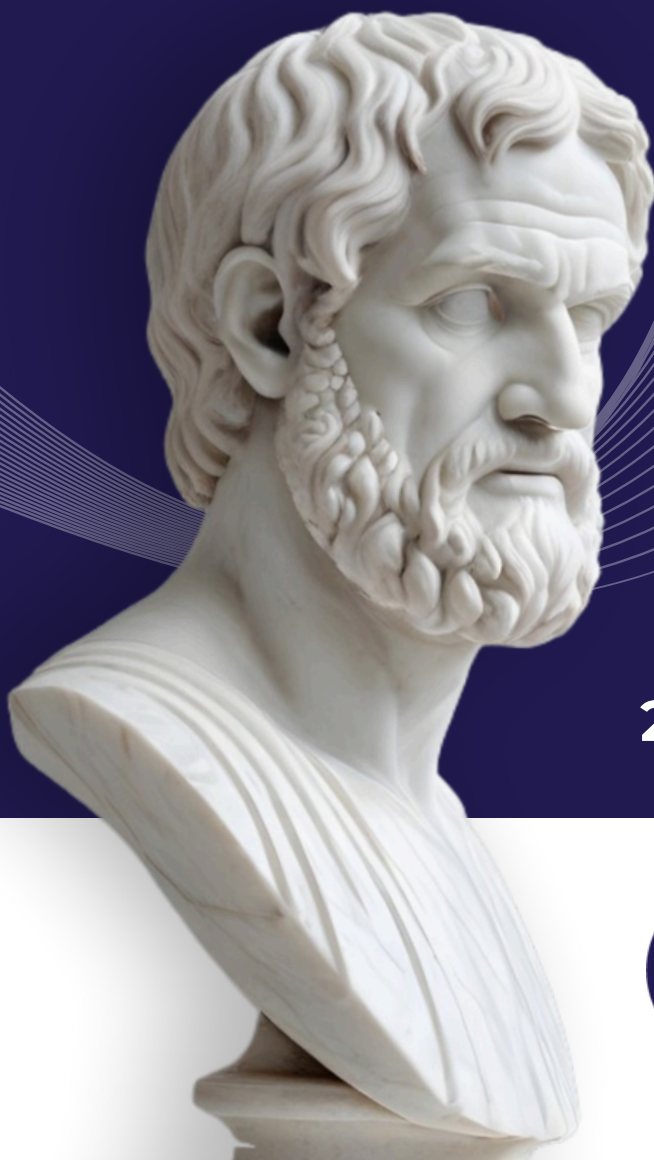
# Towards Ethical AI in Persuasive Technology

– an exploratory study

**Information Studies Master Thesis**

Written by  
**Edina Jakab & Frida Malene Cevro**

Supervised by  
**Frederik Stjernfelt**



2024





**AALBORG  
UNIVERSITET**

# **Information Studies**

## **Master Thesis**

**Towards Ethical AI in Persuasive Technology  
- An Exploratory Study**

<b>Written by</b>	Edina Jakab (20221347) Frida Malene Cevro (20220334)
<b>University</b>	Aalborg University Copenhagen
<b>Field of study</b>	MSc Information Studies
<b>Supervisor</b>	Frederik Stjernfelt
<b>ECTS</b>	30
<b>Date</b>	May 31, 2024
<b>Characters</b>	325 274
<b>Pages</b>	136

# Abstract

This exploratory study aims to examine the ethical implications of using artificial intelligence (AI) in persuasive technology. With the rapid advancements in AI and its persuasive capabilities, there is an increasing need to understand and address the ethical challenges posed by persuasive technologies. This is investigated through a systematic literature review, identifying key ethical principles and responsible stakeholders in relation to persuasive AI technology, and an investigation of the viewpoints of the different stakeholders. These insights are used to identify key focus areas for future efforts to ensure ethical AI in persuasive technology.

The project is a comparative study with an exploratory approach, comparing the viewpoints of different AI stakeholders. We collected qualitative data by conducting a total of nine semi-structured interviews with representatives from four different AI stakeholder groups: companies, developers, governments, and researchers. The interviews examined the viewpoints of the stakeholders on the five ethical principles in relation to persuasive AI technologies: beneficence and non-maleficence, fairness and justice, human autonomy and agency, and accountability and oversight. A thematic analysis was conducted to analyze emerging themes. We employed theories such as Human-computer Interaction, Persuasive System Design, ethics theory, and Self-Determination Theory to support our analysis with the goal of understanding the standpoint of these stakeholders, including challenges, attitudes, motivations, and practices.

The research identified end-users as a key responsible stakeholder group, in addition to the four previously recognized. Generative AI has emerged as a new type of persuasive technology requiring future examination from an ethical perspective. Based on our analysis, we propose a categorization of techniques and practices in persuasive AI based on their ethical implications. These suggested techniques and practices can be applied to improve ethical AI development in persuasive technology. Finally, we propose eight focus areas for future work: clear definitions and boundaries, stakeholder engagement and collaboration, continuous monitoring and transparency, addressing the “black box” challenge, regulation and ethical standards, education and AI literacy, justified use of AI, and balancing innovation and regulation.

**Keywords:** Ethical AI, Persuasive Technology, Stakeholders, AI Regulation, Human-Computer Interaction, Ethical Principles

# Acknowledgments

First and foremost, we would like to express our gratitude to our supervisor, Frederik Stjernfelt, for his support, guidance, and valuable insights throughout the course of this research. His expertise and encouragement have been crucial in shaping the direction and completion of this thesis.

We are also profoundly grateful to all the interview participants who made this study possible. Their willingness to share their expertise, experiences, and perspectives provided rich and nuanced data, forming the heart of this investigation, and we recognize the significant role they have played in this research.

Thank you all for your support and assistance in this research process.

# Table of Contents

<b>1. Introduction.....</b>	<b>3</b>
1.1. Motivation.....	4
1.2. Problem Discussion.....	4
1.3. Problem Formulation.....	6
<b>2. Related Work.....</b>	<b>9</b>
2.1. Literature Search.....	9
2.1.1. Literature search method.....	9
2.1.2. The search process.....	10
2.2. Literature Review.....	14
2.2.1. Persuasive AI technology.....	14
2.2.2. Human-centered AI.....	16
2.2.3. Ethical AI.....	17
2.2.4. AI stakeholders.....	24
2.2.5. Summary of findings.....	29
<b>3. Methodology.....</b>	<b>30</b>
3.1. Philosophy of Science.....	30
3.2. Research Design.....	32
3.3. Data Collection.....	33
3.3.1. Data collection methods.....	33
3.3.2. Research criteria.....	35
3.3.3. Participants.....	36
3.3.4. Procedure.....	38
3.3.5. Data collection process.....	39
3.4. Data Analysis.....	40
3.4.1. Thematic analysis.....	41
3.4.2. The analysis process.....	42
3.5. Validity, Reliability, and Generalizability.....	46
3.6. Ethical Considerations.....	46
<b>4. Theoretical Framework.....</b>	<b>48</b>
4.1. Human-Computer Interaction.....	48
4.2. Persuasive System Design.....	49
4.3. Self-Determination Theory.....	51
4.4. Ethical Theories.....	54
4.5. Summary.....	57
<b>5. Thematic Analysis.....</b>	<b>58</b>
5.1. Persuasive AI Technology.....	59

5.1.1. Defining persuasion.....	60
5.1.2. Techniques and technologies.....	62
5.1.3. Risks.....	65
5.2. Ethical Principles.....	67
5.2.1. Beneficence and non-maleficence.....	67
5.2.2. Fairness and justice.....	73
5.2.3. Human autonomy and agency.....	78
5.2.4. Transparency and explainability.....	83
5.2.5. Accountability and oversight.....	88
5.3. AI Stakeholders.....	90
5.3.1. Overview.....	90
5.3.2. Challenges.....	96
5.3.3. Attitudes and beliefs.....	99
5.3.4. Motivation.....	103
5.3.5. Responsibility.....	106
5.3.6. Practices.....	113
5.4. Summary of Results.....	114
<b>6. Discussion.....</b>	<b>115</b>
6.1. Standpoints of AI Stakeholders.....	116
6.2. Ethical Use of Persuasive AI Technology.....	119
6.3. Focus Areas for Future Efforts.....	121
<b>7. Conclusion.....</b>	<b>124</b>
<b>8. Reflection and Future Work.....</b>	<b>125</b>
<b>References.....</b>	<b>126</b>
<b>Appendixes.....</b>	<b>132</b>

# 1. Introduction

Communication and persuasion are central to human interaction and democracy. Humans naturally communicate to share attitudes and offer help, distinguishing them from many other species. This ability to communicate effectively is crucial in various domains, including politics and economics, where persuasion plays a vital role. AI systems, such as OpenAI's GPT-4, are being developed to communicate in human-like ways, making the study of AI persuasion salient. Persuasion is significant not only in marketing and politics but also in national security, where information operations can shape public opinion and actions. The power of ideas shows how they can lead to significant societal changes, even if the outcomes are unintended (Burtell & Woodside, 2023). Understanding ethical principles and the implications that make persuasion effective, such as trustworthiness, is essential in human and AI communication. AI systems, while not inherently intentional, can still influence beliefs, necessitating careful consideration of their persuasive capabilities.

AI breakthroughs are pervasive and embedded in daily tools and complex systems due to advancements in computing power, big data, and fast networks, enabling significant progress in various domains, such as personal assistance (Coeckelbergh, 2020). However, the impressive capabilities of AI have been raising concerns about the potential for machines to surpass human intelligence and control, evoking fears reminiscent of science fiction scenarios where AI undermines human autonomy, such as *The Terminator*.

To move beyond the hype about AI, it is helpful to consider historical narratives about humans and artificial beings, which are deeply rooted in both Western and non-Western cultures. These include ancient myths, like the Golem and Prometheus, and modern stories, such as Mary Shelley's "*Frankenstein*." These narratives, which often revolve around themes of creation, control, and competition, continue to shape contemporary fears and discussions about AI, encapsulated in the "Frankenstein complex," a term coined by Isaac Asimov to describe the fear of robots. This complex is echoed in popular culture and endorsed by figures like Elon Musk and Stephen Hawking, reflecting anxieties about the potential dangers of AI (Coeckelbergh, 2020). To address these anxieties, we can critically examine persuasive AI from the perspectives of ethical theories and the field of human-computer interaction (HCI), focusing on current practices in persuasive AI and addressing concrete ethical and societal challenges.

In this study, we engaged in conversations with AI stakeholders, such as companies, developers, governments, and researchers, to learn about their practices and viewpoints on specific ethical aspects of persuasive AI technology. The research took an exploratory and comparative approach to the different viewpoints, including attitudes, motivations, and responsibilities of AI stakeholders regarding ethical principles. The goal was to identify current challenges that are not being adequately addressed and ideate on possible focus areas for ensuring the ethical use of AI in persuasive technology. Based on the insights of the stakeholders, we proposed a framework for the ethical use of persuasive AI technology and

concrete focus areas that we found crucial to address going forward to solidify the safety of individuals and society.

## 1.1. Motivation

Persuasive technology is not a new concept; however, not so long ago, it started taking on new shapes and forms. We have been regular users of search engines and social media, and we have seen countless recommendations for content and products, even some that we only mentioned verbally while talking to someone else, without searching on the internet. These incidents raise questions about how much these platforms know about us and how transparent they are about the strategies they use to recommend content and keep us engaged to continue using their technology.

With the recent boom of emerging AI technologies, such as large language models (LLMs) and generative AI, and the discussions around them, we also noticed an increase in the number of artificial agents, such as ChatGPT-4o, Pi.ai, Perplexity.ai and many more. While testing these platforms, we noticed that the former often presented inaccurate information, also called “hallucinations,” and sometimes even recommended specific products from specific brands. This raised questions about the trustworthiness of these technologies and the potential risks of misinformation and disinformation. Furthermore, they have the potential to influence user preferences and choices through personalized recommendations. This also made us wonder: if these technologies can influence our purchasing decisions and mimic the behavior of social media influencers, in what other ways can they influence us?

Our motivation to investigate the ethical implications of persuasive AI technologies was to identify the potential risks they are associated with and which stakeholders are responsible for ensuring that individuals and society are protected from these. Furthermore, we were interested in learning what real challenges these stakeholders are currently facing and what they could focus on to reinforce their current efforts.

## 1.2. Problem Discussion

The term “technology,” derived from the Greek “techne” (art and skill) and “ology” (study of), encapsulates the knowledge of making things, leading some to propose renaming our species from *Homo sapiens* (the-wise-people) to *Homo technologicus* (they-who-use-tecnh) to better reflect our defining characteristic of tool use and creation (Warwick, 2016). A new branch of technology began to revolutionize the art of tool-making with the invention of the first digital computer in 1943. Shortly after, the first working AI programs were written in 1951. In 2019, Google introduced BERT (Base), the large language model (LLM) behind every English-based query administered via Google Search. However, the impact of LLMs only started getting noticed by the public when OpenAI released ChatGPT in November 2022. Today, anyone can have a human-like conversation with ChatGPT-4o, based on an LLM estimated to be running with hundreds of billions of parameters, compared to 110



million parameters of GPT-1 or BERT Base (Toloka Team, 2023). This fast-paced advancement highlights the potential of AI to become a significantly more powerful tool over the upcoming years.

In just 18 months, over a billion people have used AI chatbots, and some public figures, such as Mustafa Suleyman (CEO of Microsoft AI), have gone as far as suggesting that artificial agents will become an entirely new species. He has also emphasized that because AI is still currently being built and designed, now is the ideal time to define what it should or should not be used for. He highlighted that autonomy is a threshold over which societal risks increase. However, it is important not to fall into a pessimism aversion trap, overwhelmed by the fear of confronting potentially dark realities, which results in the tendency to look the other way (TED, 2024). Persuasion in the age of AI involves complex processes where AI entities generate, augment, or modify messages to shape, reinforce, or change human responses. Therefore, to be able to regulate it, it is essential that we understand the technology and its ethical implications.

The global tech debate has become increasingly critical of tech giants, emphasizing the need for regulation to balance the societal benefits of the technology with its potential harms, requiring democratic countries to lead in shaping a responsible technological future (Ministry of Foreign Affairs of Denmark, 2021). Initially employed for marketing and commerce, digital platforms now impact human actions in unforeseen manners, utilizing algorithms and convincing tactics that frequently compromise personal freedom and jeopardize democracy by regulating online conduct through digital structures, market pressures, and the absence of ethical and legal adherence (Botes, 2023). These risks highlight the need for strong privacy protections and management of technological impacts on democracy and society.

Most countries lack the resources and expertise to compete for AI leadership, making them dependent on wealthy corporations and states. This could worsen geopolitical power imbalances, although the benefits of AI will eventually spread globally, enhancing productivity and benefiting many nations, including those contributing to AI inputs. Washington and Beijing prioritize out-competing each other in AI development over potential societal risks, pouring resources into AI advancements and regulation to hinder each other's progress. This rivalry, rooted in mistrust, contrasts with the broader AI landscape where a few specialist companies control the technology, potentially undermining state power outside of China. Governments have successfully implemented regulations for older digital technologies, but AI's rapid pace outstrips the slow policymaking process, with the EU making some progress while the US lags behind (Bremmer & Suleyman, 2023), underscoring that AI has the potential to transform the global economy.

While China and the United States are competing to lead in critical technologies, the EU is focused on establishing favorable conditions to become a global tech leader while ensuring that technology benefits its citizens and maintains its rule-setting influence. The global distribution of the top ten largest tech companies in the world measured by market capitalization includes Microsoft (3 billion dollars), Apple (2.6 billion dollars), and Meta

Platforms (1,3 billion dollars), which are predominantly concentrated in the United States, with significant representation in Asia and Europe. Consequently, tech companies are expected to prioritize greater social responsibility, which involves enhancing consumer rights and privacy protection, combating misinformation, promoting transparency, and fostering international cooperation on responsible technological solutions while ensuring compliance with democratic values and human rights (Ministry of Foreign Affairs of Denmark, 2024). This highlights the geopolitical spread and market capitalization of leading tech companies, as well as their responsibility to prioritize ethical development.

For the next few years, AI's development is expected to be driven primarily by private businesses rather than policymakers. This will give technologists significant influence over a technology that could profoundly impact the power and relations of nation-states, presenting an unprecedented regulatory challenge for governments. Current regulatory efforts for AI are insufficient and lag behind its rapid development, making AI a moving target when it comes to regulation. With the EU's AI Act not fully effective until 2026, there is a need for innovative, global-level governance tailored to the unique challenges of AI and private sector control (Bremmer & Suleyman, 2023). Furthermore, European values emphasize upholding human rights and a multi-stakeholder approach to address the global digital challenges and ensuring that new technologies, such as AI, are used for the benefit of society (Ministry of Foreign Affairs of Denmark, 2024).

From the perspective of citizens, a survey showed that trust in AI systems varies significantly across countries, with high levels in emerging economies like China and low levels in France. People generally trust AI for accurate output and helpful services but are wary about safety, security, fairness, and privacy, especially when it comes to recommender systems. Overall, feelings toward AI are mixed, combining optimism and excitement with worry and fear, particularly in countries like France and Japan. The survey conducted across 17 countries (including Australia, Canada, China, Japan, Finland, France, Germany, the UK, and the USA) with more than 17,000 respondents indicated that only one-third of the respondents had high or complete confidence in governments and tech companies regarding the regulation and governance of AI development. Meanwhile, research institutions, universities, and defense forces were seen as more capable in this matter. Furthermore, global public support strongly favors organizations deploying AI systems to adhere to high standards in fairness, human oversight, transparency, accountability, risk mitigation, and AI literacy (Gillespie et al., 2023). Therefore, it is advantageous for organizations to comply with societal expectations regarding regulation and ethical development in AI, including the domain of persuasive technology.

### 1.3. Problem Formulation

Building on this understanding of the AI landscape today and the enormous potential AI has to influence and persuade users, this study aims to examine the ethical challenges posed by

AI in persuasive technology and possible ways to address these. Subsequently, we present the following problem statement (PS):

*PS:* What are the ethical implications of using AI to influence user behavior, and how can responsible stakeholders work towards ensuring the ethical use of AI in persuasive technology?

Furthermore, the research is structured around three key research questions (RQs) that explore the ethical use of AI in persuasive technology, focusing on stakeholder roles and responsibilities:

*RQ1:* What are the key ethical considerations associated with using AI to influence user behavior?

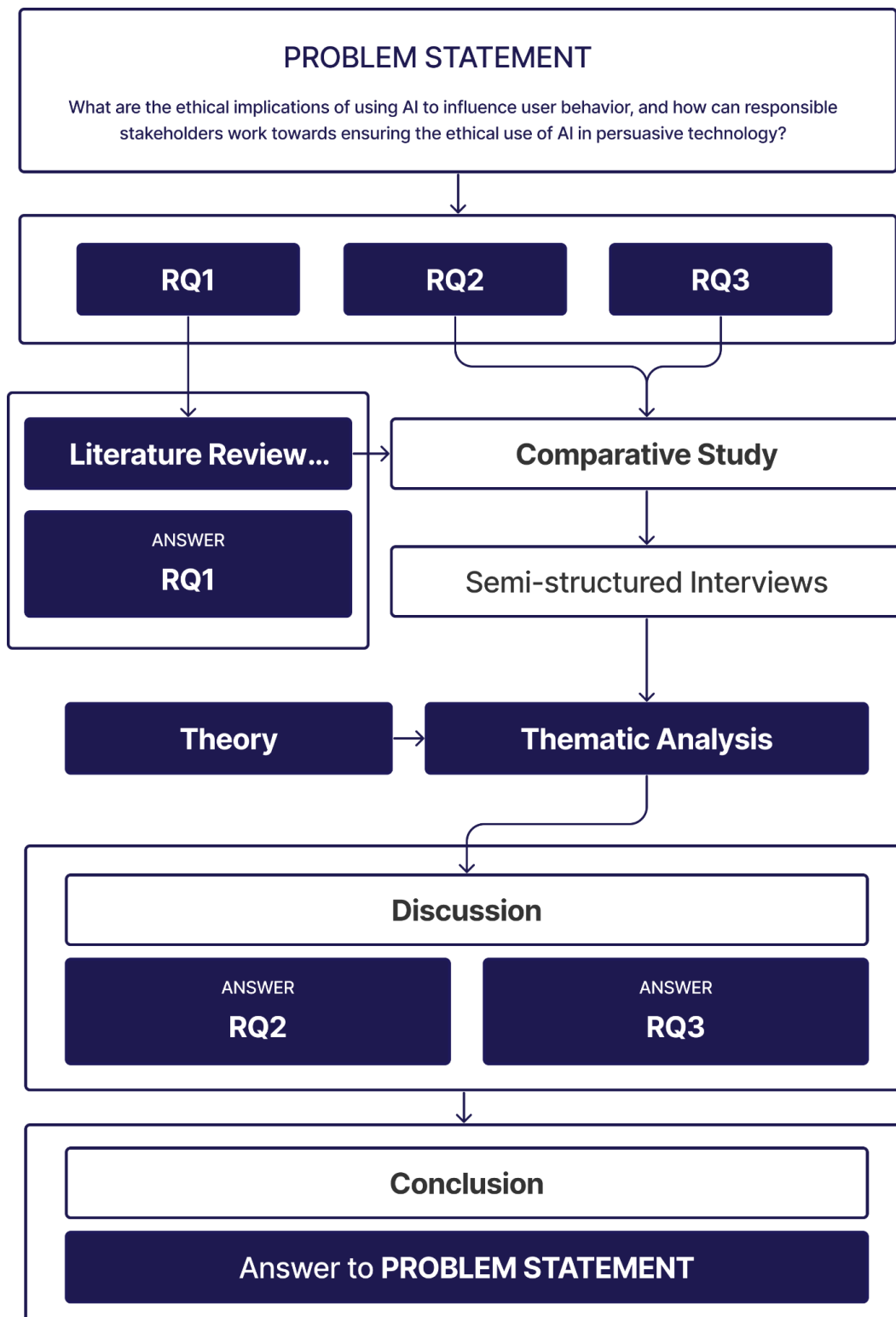
*RQ2:* What are the viewpoints of AI stakeholders about current practices for addressing the ethical implications of AI-driven persuasive technology?

*RQ3:* What should be the focus areas of stakeholders in future efforts to ensure ethical AI in persuasive technology?

The research questions will be addressed in different sections of this study. The process of investigating these questions and consequently answering the problem statement is visualized in the thesis structure, as shown in **Figure 1** below.

**Figure 1**

*Visualization of thesis structure*



## 2. Related Work

To address RQ1: *What are the key ethical considerations associated with using AI to influence user behavior*, we conducted a literature search and review of related work. The purpose of the literature review is to position our research within the existing body of literature in the field. This process provides context and helps in identifying gaps, thereby establishing the foundation of our study (Rowley & Slack, 2004, p. 32). In this section, we describe our process and the methods employed to conduct literature search and review before discussing the findings and the identified gaps in research.

### 2.1. Literature Search

A literature review is an objective and comprehensive summary and critical analysis of the available literature relevant to the topic being studied (Cronin et al., 2008, p. 38). This is a fundamental part of research as it establishes what is already known about the research topic. The goal of the literature review is to set the stage for new inquiries by building a comprehensive understanding of the background and justifying the need for further exploration (Bryman, 2016, p. 90). In our case, this means to examine the existing knowledge about AI-influenced user behavior and the ethical implications that come with it. A literature review starts with a literature search, therefore it is important to evaluate which literature search method is suitable for our study

#### 2.1.1. Literature search method

There are different approaches to conducting the literature review, notably distinguishing between narrative and systematic reviews (Bryman, 2016, p. 90). A systematic review, as the name implies, is a highly structured methodology. It involves explicit procedures to identify, select, and evaluate relevant literature, aiming to address specific research questions (Bryman, 2016, p. 91). While a narrative review primarily aims to provide background information for understanding the topic in question, the systematic review seeks to provide a complete list of studies on a specific topic (Cronin et al., 2008, p. 38-39). This approach has emerged as a response to a lack of thoroughness that has been observed in traditional narrative reviews and as a method for mitigating the risk of biases (Bryman, 2016, p. 98). Additionally, it ensures transparency through a clearly stated and replicable process. Therefore, the systematic literature review is argued to provide a more reliable foundation, providing a thorough understanding of the existing knowledge on the subject (Bryman, 2016, 104).

On the other hand, the systematic approach has some limitations. Bryman (2016) highlights that these reviews can struggle with research questions where boundaries are shifting, often overemphasize methodological details at the expense of deeper analytical insights, and face difficulties in objectively assessing the quality of qualitative research due to inconsistent evaluation standards (p. 105). Despite these challenges, we found that a systematic review

was suitable for our research question exploring ethical considerations associated with using AI to influence user behavior in design. Employing this method allowed us to thoroughly review and analyze all relevant literature on this complex topic. Using a structured approach allows the researchers to ensure that the exploration of the ethical issues in AI and design is comprehensive, reducing bias and providing a solid base for understanding and further research. This makes it a practical choice for our study, even with the acknowledged limitations of the approach.

## 2.1.2. The search process

The literature search was conducted based on methods for systematic literature review. Our process can be divided into four main steps, as outlined in this section: defining the scope, defining keywords and strings, selecting databases, and selecting the literature.

### Defining the scope

In order to ensure objectivity and relevance in a literature review, it is important to define the purpose and the scope of the review (Bryman, 2016, 99). This was achieved by defining a research question for examination and establishing specific inclusion and exclusion criteria for selecting literature relevant to this. As detailed in **Table 1**, we had specified inclusion criteria to ensure that the literature is scientific, up-to-date, and related to our field of research. These criteria are often determined during the search process and automatically applied to filter the results. On the other hand, the exclusion criteria are primarily guidelines for evaluating the content of the literature. This was both applied in the search phase in the form of certain keywords to exclude (see **Table 1**) and during the review of the results to ensure topical relevance.

**Table 1**

*Overview of inclusion and exclusion criteria for our search*

Inclusion criteria	Exclusion criteria
<ul style="list-style-type: none"> <li>• Peer reviewed</li> <li>• Max 5 years old</li> <li>• English</li> <li>• Keywords in abstract</li> <li>• Content type: (research) article, chapter of book</li> </ul>	<ul style="list-style-type: none"> <li>• Design of AI tools</li> <li>• Designing using AI tools</li> <li>• Chatbots/conversational AI or AI assistants</li> <li>• Unrelated fields like healthcare/medical science, disabilities</li> </ul>

### Defining keywords and search strings

We started the process of defining search strings by mind-mapping relevant keywords to our research topic. Since our study is exploratory in nature, we are interested in examining a

broad body of literature relating to AI ethics and how users are influenced while interacting with AI technology. After writing down a variety of keywords related to this topic, we sorted them into categories, resulting in five main topics: AI technology, user experience, user behavior, regulation, and ethics. Based on these categories, we elaborated on synonyms, abbreviations, and different terms for the topical keywords. This approach is referred to as 'building blocks' by Rowley and Slack (2004) and results in a thorough search covering a wide selection of literature within that field (p. 35-35). For words where part of it can vary, we used asterisks in combination with the stem of the word to indicate that this can be followed by different letters. For example, “ethic\*” can give results including both “ethics” and “ethical,” both of which are relevant to our search. These variations were put together with the Boolean operator “OR” to form an inclusive search string for each topic (Cronin et al., 2008, p. 40), as shown in **Table 2**. Furthermore, we also defined a sixth search string on the topic exclusion, which contains words related to our exclusion criteria. This will be added as “AND NOT” for each search conducted to filter out irrelevant literature to our scope. Lastly, we defined where in the literature the keywords should appear, marking them as either abstract or full text, depending on how important they are to the selection.

**Table 2**

*An overview of the selected topics and their search string for the literature search*

Topics	Inclusive search string	Where
AI technology	AI OR "Artificial intelligence" OR HAI OR HCAI OR "Human-centered AI"	Abstract
User experience	UX OR "user experience" OR HCI OR "human-computer interaction"	Abstract
User behavior	“user behavior” OR “user behaviour” OR manipul* OR “decision making” OR “persuasive design” OR automat* OR recommend*	Abstract
Regulation	regulat* OR politic* OR guidelines OR principles OR govern*	Full text
Ethics	ethic* OR responsible OR trustworthy OR transparen*	Full text
<i>Exclusion</i>	medic* OR healthcare OR chatbot OR conversational OR assistant	Abstract

After defining the various topics of interest, we consolidated them into broader themes for investigation by merging related topics. In answering RQ1, our objective was to explore how AI influences user behavior, identify the ethical considerations associated with AI, and examine the intersection between these two aspects. Consequently, this process led to the identification of three primary themes for our literature search: AI in UX, AI ethics and regulations, and AI ethics in UX. Each theme is presented in **Table 3**, accompanied by its related search string. These search strings are constructed from the topics listed in **Table 2**

using the Boolean operators AND and NOT.

**Table 3**

*An overview of how the topics from **Table 2** were used to make themes for the search*

No.	Theme	Search string based on topics
1.	AI in UX	(AI technology) AND (User experience) AND (User behavior) AND NOT (Exclusion)
2.	AI ethics and regulations	(AI technology) AND (Ethics) AND (Regulation) AND NOT (Exclusion)
3.	AI ethics in UX	(AI technology) AND (User experience) AND (User behavior) AND (Regulation OR Ethics) AND NOT (Exclusion)

## Database selection

To conduct the search, we used the AAU online library to browse for related databases. By looking at the recommended databases for the subjects of information science and IT and quickly evaluating the top results for our search strings, we decided to move forward with Scopus, ProQuest, and ACM Digital Library. The ACM Digital Library was a good fit as it is a computer science portal and would, therefore, provide a technical perspective on AI applications, setting a topical limitation to our search. On the other hand, Scopus and ProQuest are more general databases offering a wider selection of literature and subjects that could yield different perspectives on our topic. These were also more suitable for the topic of ethics and governance.


Based on the search strings defined in **Table 3**, we conducted the three searches in the three different databases. For this, we used the advanced search feature offered in all of the databases and narrowed the search by selecting the inclusion criteria where possible (see Appendix A). **Figure 2** showcases an example of one of the advanced searches.



**Figure 2**

*Example of an advanced search for theme 3 conducted in ProQuest*

Advanced Search [Command Line](#) [Recent searches](#) [Thesaurus](#) [Field codes](#) [Search tips](#)



AI OR "Artificial intelligence" OR HAI OR HCAI OR "Human-centered AI"	in	All abstract & summary text – SUMMARY* ▾
AND ▾ UX OR "user experience" OR HCI OR "human-computer interaction"	in	All abstract & summary text – SUMMARY* ▾
AND ▾ "user behavior" OR "user behaviour" OR manipulat* OR "decision making" OR "persuasive design" OR automat* OR recommend*	in	All abstract & summary text – SUMMARY* ▾
AND ▾ regulat* OR politic* OR guidelines OR principles OR govern* OR ethic* OR responsible OR trustworthy OR transparen*	in	Anywhere ▾
NOT ▾ medic* OR healthcare OR chatbot OR conversational OR assistant	in	All abstract & summary text – SUMMARY* ▾

## Literature selection

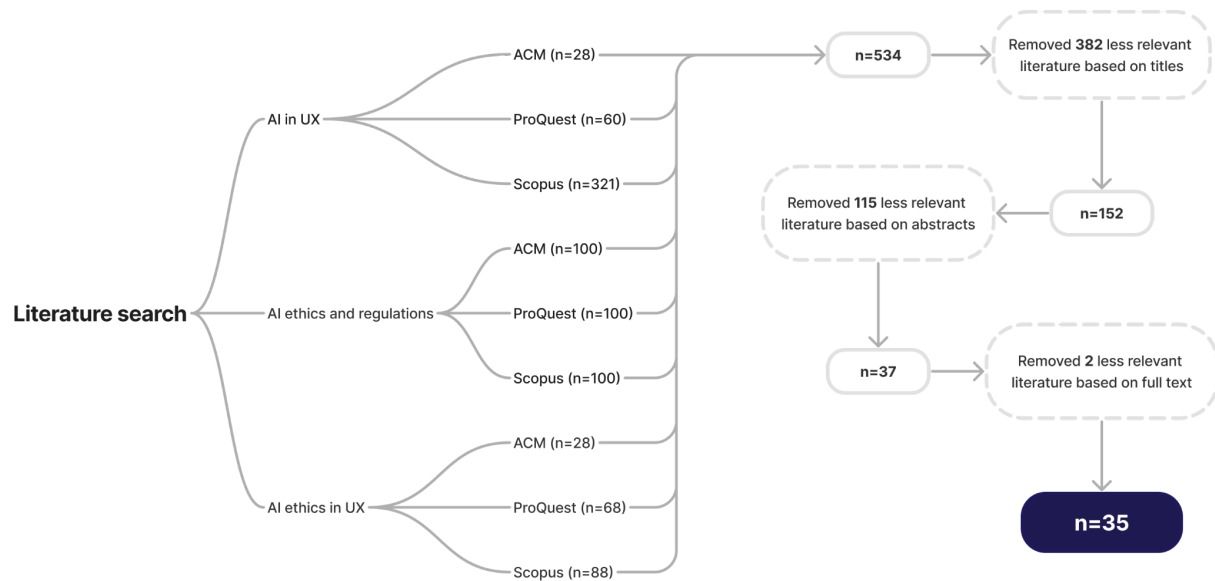
We started the literature selection for our research by searching across three different databases. If our initial search resulted in more than 1000 results, we refined our criteria to consider only the publications from the past three years instead of five. The goal was to narrow down the results and only focus on the newest research in this rapidly changing field. For searches that yielded more than 100 results, we sorted these by relevance and took only the top 100 for deeper evaluation.

After exporting the data, we combined all CSV files into one comprehensive file and removed any duplicate entries. This process left us with 534 unique titles. Next, we scrutinized these titles to filter out irrelevant ones, which reduced the list to 152 titles. Subsequently, we reviewed the abstracts of these articles to assess their direct relevance to our research questions while considering the number of citations associated with these. This step narrowed down the selection to 37 potentially suitable articles

Finally, we conducted a thorough review of the 37 articles by reading the full text for each. After this detailed review, articles that did not meet our criteria based on their content were excluded, resulting in a total of 35 articles. This rigorous selection process, as outlined in **Figure 3**, ensured that our literature review was focused, relevant, and comprehensive, providing a solid foundation for our research.

**Figure 3**

*Process of narrowing down literature selection*



## 2.2. Literature Review

In this section, we analyze and synthesize the resulting literature from our literature search and gather findings into clearly defined themes that emerged from the studies. Adopting a narrative synthesis approach, as outlined by Bryman (2016), we have arranged the literature into themes to present our findings clearly (p. 102). The main themes identified were persuasive technology, human-centered AI, ethical AI, and AI stakeholders. This synthesis helps in highlighting the commonalities across different studies and provides a structured way to explore complex issues like the ethics of AI and its impact on users.

### 2.2.1. Persuasive AI technology

As AI is a rapidly evolving, complex, and sometimes controversial technology, it increasingly intersects with the field of persuasion. Academics have examined the role of AI in persuasion from various perspectives, such as persuasive robotics, and robot persuasion. However, AI possesses unique characteristics that challenge existing theories of persuasion. As interactions with artificial agents become more prevalent and sophisticated, understanding AI's impact on communication and persuasion is crucial (Dehnert & Mongeau, 2022).

Dehnert and Mongeau (2022) describe persuasion as an intentional, message-based, and goal-directed symbolic process where a source aims to influence attitudes, behaviors, intentions, or perceptions. This is achieved through various forms of communication, broadening its scope to include diverse outcomes and both large-scale and one-on-one interactions (Dehnert & Mongeau, 2022, p. 387). Benner et al. (2022) point to concepts such

as gamification, gamblification, and digital nudging as methods for persuasion that have become widespread within many sectors. They highlight that since these concepts can be powerful tools for influencing behavior, they also raise ethical concerns. These concerns include lack of transparency and harming user autonomy, leading to decision-making processes that benefit third parties at the expense of the users' well-being (Benner et al., 2022, p. 549).

A paper by Gao et al. (2023) explores the significant advancements of persuasive AI in the advertising sector, pinpointing four critical areas: targeting, personalization, content creation, and ad optimization. These facets are enhanced by AI technologies such as machine learning and generative AI, which streamline advertising by providing data-driven insights, enabling real-time content adjustments, and improving audience targeting accuracy. While AI has the potential to revolutionize advertising, it also poses significant challenges and ethical concerns, such as algorithmic biases, privacy issues, and the need for transparency, necessitating collaborative efforts across multiple stakeholders to ensure its responsible use. They argue that addressing these challenges requires a *“clearer understanding of consumer perceptions of privacy and the expected use of personalized information”* (Gao et al., 2023, p.12). Additionally, a recent study by Matz et al. (2024) found that personalized ads generated by AI, particularly ChatGPT, significantly increased ad effectiveness and spending readiness, demonstrating AI's potential to enhance personalized persuasion.

Popular platforms like YouTube, Facebook, and Instagram operate on an economic model where user attention is harvested and sold to advertisers. This is achieved using technologies that manipulate user engagement and compromise autonomy by prioritizing commercial interests over user interests, often through opaque or deceptive practices. The platforms use persuasive techniques such as content recommendations and nudging to increase user engagement. AI-driven recommendation systems shape our perceptions and behaviors, and thereby challenge our autonomy by potentially misleading users who feel autonomous without realizing the bias. Furthermore, while developers may not aim to increase radicalization, they might ignore this effect if their algorithms, designed to optimize engagement, succeed by limiting content diversity. This raises ethical questions about how these goals affect user autonomy (Burr & Floridi, 2020).

When it comes to digital nudging, which is a technique that aims to influence decision-making through non-coercive methods, Sadeghian and Otarkhani (2023) highlight the challenges and issues of using such technology. They emphasize that the integration of AI and big data analytics into digital platforms has significantly advanced real-time, personalized nudging techniques. These subtly influence individual behaviors and decision-making across various domains like e-commerce, healthcare, and social media, promoting more effective and context-specific interventions. The nudging techniques raise concerns as they are commonly used for marketing objectives instead of supporting personal or social benefits, which can result in negative outcomes like financial abuse and undesirable behavior. This concern is also highlighted by Klenk (2020), stating that *“manipulative interaction, in particular, is detrimental to well-being because of its effects on user*

*autonomy*“ (Klenk, 2020, p. 83). He addresses a statement by Facebook's ex-president, according to which social media applications aim to maximize user engagement by exploiting human psychology, providing intermittent dopamine hits through social interactions like likes and comments. This type of persuasive design strategy creates a social-validation feedback loop, intentionally leveraging psychological vulnerabilities to increase user content and engagement (Klenk, 2020).

A recent paper by Matz et al. (2024) evaluates how large language models, like OpenAI's ChatGPT, as a new technology, can automate the creation of effective, psychologically tailored persuasive messages. This significantly enhances their efficiency and effectiveness across various domains, including consumer marketing, political appeals, and health messaging (Matz et al., 2024). The complexity of persuasion is amplified by AI, introducing "thick" and "thin" AI concepts that vary in the presence of AI and social cues. Thin AI involves minimal AI cues, often leading receivers to mistake AI-generated messages for those from human or corporate sources. Thick AI, on the other hand, features an obvious AI presence with ample social cues, facilitating direct, interactive human-AI relationships. These relationships can lead to anthropomorphism, where humans treat artificial agents as human-like friends, impacting trust and persuasive effectiveness. Future research should explore how these factors, along with individual differences and contextual elements, influence AI-based persuasion across various social, cultural, ethical, legal, political, and psychological domains (Dehnert & Mongeau, 2022).

Based on the reviewed literature, persuasive AI technology offers significant advantages by enhancing personalized communication and improving the effectiveness of advertisements, political appeals, and health messaging through advanced targeting and content creation. However, it also presents notable ethical challenges, such as compromising user privacy, undermining autonomy through manipulative tactics, introducing algorithmic biases that can skew information, and lacking transparency in how data and algorithms are utilized. These concerns necessitate careful consideration and regulation to ensure that the benefits of AI-driven persuasion do not come at the cost of ethical standards and user trust.

### 2.2.2. Human-centered AI

An emerging theme in the literature is the importance of the "human-in-the-loop" approach to AI, where the focus transitions from solely technical advancements to a more human-centric perspective. This approach, also known as Human-Centered Artificial Intelligence (HCAI), is recognized as a potentially promising direction for the advancement of AI, merging insights from Human-Computer Interaction (HCI) research and human factors (Chignell et al., 2023; Ozmen Garibay et al., 2023).

HCAI aims to place humans at the core of AI development to enhance safety, reliability, trustworthiness, and human capabilities (Shneiderman, 2020b; Bingley et al., 2023; Ozmen Garibay et al., 2023). This approach is diverse in its application, ranging from algorithm training to addressing broader societal impacts, with a common emphasis on the importance

of transparency and explainability (Bingley et al., 2023; Ozmen Garibay et al., 2023). By moving away from traditional AI development, which primarily assesses algorithm performance, HCAI now includes evaluating the impact of these technologies on human performance and satisfaction, promoting a more inclusive and comprehensive approach to technology development. Furthermore, it addresses the criticism outlined by Hagendorff (2020), who notes that AI ethics often overlook broader contexts such as care, welfare, and social responsibility (p. 103). This shift is supported by user-centered participatory design methods and engaging with a wide range of stakeholders to ensure that AI systems are aligned with human goals, activities, and values (Shneiderman, 2020a, p. 2).

Shneiderman (2020b) further elaborates on this approach, proposing a guideline that seeks both high levels of human control and high levels of automation (p. 495). He argues that transitioning to HCAI can revolutionize design thinking, leading to the development of applications that not only enhance automation but are also "*amplifying, augmenting, enhancing, and empowering people to innovatively apply systems and creatively refine them*" (Shneiderman, 2020b, p. 495).

Building on these concepts, Ozmen Garibay et al. (2023) outline six grand challenges in HCAI: ensuring AI enhances human well-being, is designed responsibly, respects privacy, adheres to human-centered design and evaluation frameworks, is governed and overseen appropriately, and interacts with humans while respecting their cognitive capacities. The study concludes that effective AI implementation requires not only careful algorithm and user interface design but also oversight spanning from individual developers to global organizations. This entails establishing governance structures that include HCAI training, codes of practice, and comprehensive standards and regulations (Ozmen Garibay et al., 2023). This is supported by Shneiderman (2020a), who underscored the importance of involving responsible stakeholders in every aspect of AI development and regulation, ensuring that AI systems are both human-centered and ethically aligned.

Furthermore, Sigfrids et al. (2023) stress the need for the governance of HCAI to evolve from traditional hierarchical models towards more inclusive and transparent frameworks. Such governance should facilitate stakeholder and citizen engagement and focus on the socio-technical and emancipatory potential of AI, ensuring that it serves the broader good and aligns with multidisciplinary needs and sustainability goals (Sigfrids et al., 2023).

### 2.2.3. Ethical AI

When exploring the ethical considerations of AI systems, understanding what constitutes ethical AI is crucial. However, defining ethical AI proves challenging as it involves synthesizing a broad spectrum of values and ideas that should guide the advancement of AI technologies. This challenge arises from the need to accommodate diverse and sometimes conflicting ethical standards and expectations across different cultures and stakeholders (Corrêa et al., 2023).

## Trustworthy AI

Despite these complexities, a recurrent term in the literature on AI ethics is "trustworthy," which is frequently used both as an attribute of ethical AI and almost synonymously with ethical AI itself. In the context of AI, "trustworthy" implies that the system is reliable, safe, and designed in accordance with ethical standards that prioritize human well-being. According to Banovic et al. (2023), this involves elements of competence, transparency, and fairness, which are crucial for technological innovation and societal acceptance. Trustworthy AI ensures task accuracy and efficiency, prioritizes safety and privacy, and provides clear justifications for its decisions. It also offers transparency about its development and the intentions of its creators, promoting honesty and aligning with the best interests of end-users. Furthermore, it commits to fairness and equity for all stakeholders, which are vital for broader acceptance and integration (Banovic et al., 2023).

Aligned with this view, Thiebes et al. (2021) argue that trustworthy AI is essential for ensuring that both individuals and societies can fully benefit from the advantages of AI technologies. They emphasize that the success and sustainability of AI in various sectors depend significantly on establishing and maintaining trust. This trust stems from AI's consistent demonstration of ethical behaviors such as respect for privacy, fairness in decision-making, and clarity in operations (Thiebes et al., 2021).

On the other hand, the concept of trustworthy AI is subject to debate, with some arguing that trust, traditionally considered a peer-to-peer human relationship, should not or cannot be extended to machines. This perspective is grounded in the philosophical definition of trust (Smuha, 2021). Ryan (2020) argues that while AI can meet the criteria for rational trust, it does not fulfill the emotional or moral requirements necessary for affective or normative trust, suggesting that the notion of trustworthy AI is misleading.

Despite these controversies, there is a significant focus on terms such as "trustworthy," "reliable," and "responsible" AI. This underscores the necessity for robust oversight and governance to ensure AI's compliance with legal, ethical, and technical standards. The emphasis is less on the terminology and more on the underlying requirements and the recognition that AI systems operate within a broader socio-technical framework. This perspective asserts that the trustworthiness of AI depends not only on the systems themselves but also on the actions and integrity of all associated actors, including their intended use and business models (Smuha, 2021).

When it comes to the landscape of AI trustworthiness in Europe, Todorova et al. (2023) argue that as AI is rapidly evolving, it is accompanied by several ethical dilemmas including bias and discrimination from flawed training data, privacy and security risks from extensive data collection, and autonomy in decision-making. These dilemmas raise questions about responsibility, and issues with transparency and accountability, making AI systems hard to trust. Therefore, challenges in fostering trustworthy AI persist despite the adoption of the EU AI Act (Todorova et al., 2023).

Moreover, Kreps et al. (2023) introduce the concept of a "trust paradox," where despite low levels of trust in AI technologies, there is a puzzling willingness among the public to support or use these technologies. This phenomenon is observed in contexts like social media, where users continue to engage heavily despite concerns over data privacy and misinformation. The paradox is explored through several hypotheses about why this is happening, including "fear of missing out" (FOMO), a cost-benefit analysis where the perceived benefits outweigh the risks, and optimism about future improvements in technology that might address current shortcomings. These dynamics suggest that while trust remains a critical issue, the decision to use AI may be influenced by factors beyond simple trust, implicating deeper psychological and social factors in the adoption of AI technologies (Kreps et al., 2023).

## **Ethical principles**

Since the mid-2010s, the rapid growth of AI and its potential negative impacts have led to extensive work on establishing ethical norms and guidelines to ensure the responsible development and use of AI globally and regionally. This is underscored by significant entities such as the EU, the Council of Europe, and the OECD, which have created expert groups to draft guidelines focused on protecting societal and individual rights (Koniakou, 2022).

Today, there are numerous ethical principles and frameworks related to the application of AI that serve as essential guidelines for the responsible development, deployment, and management of these technologies. There has been a boom in these principles in recent years, as highlighted by the recent work of Corrêa et al. (2023), reviewing 200 guidelines and recommendations for AI governance. These principles form the foundation for creating AI technologies that adhere to the standards of ethical and human-centered AI, as previously discussed.

Gutierrez (2023) classifies principles and guidelines as forms of soft law, which, while not legally binding, set substantive expectations and guide organizational behavior within the AI landscape. Due to its flexibility, soft law is widely used within the field of AI, allowing it to adapt to the constant changes in emerging technology without slowing down innovation. However, the drawback of such laws is their voluntary nature, giving organizations the choice to decide whether it suits their interests to implement them (Gutierrez, 2023, p. 792). This relates to organizations' motivation to adopt these principles, with Bélisle-Pipon et al. (2023) describing AI ethics as a strategic move by some to stay competitive rather than a genuine effort to address ethical concerns. Subsequently, they call for AI ethics to evolve into a truly inclusive, participatory, and transparent discipline, ensuring that it reflects a wide array of perspectives rather than the interests of a few.

In the reviewed literature, several principles and frameworks for ethical AI have been suggested. In a paper by Thiebes et al. (2021) on trustworthy artificial intelligence, they underscore five principles of ethical AI: beneficence, non-maleficence, autonomy, justice, and explicability. These characterize AI-based systems that are perceived as trustworthy.

Furthermore, Díaz-Rodríguez et al. (2023) present seven largely overlapping critical requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability. These support three foundational pillars of an integrated framework for trustworthy AI - legal, ethical, and technical robustness - as outlined in their paper.

There has also been extensive reporting of ethical principles within the field of HCAI, with contributions from various stakeholder groups. Main themes within this context include privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values (Shneiderman, 2020a, p. 2-3). Additionally, Shin (2023) presents the FATE framework as essential ethical guidelines for AI deployment, consisting of Fairness, Accountability, Transparency and Explainability (p. 18).

These varied principles and frameworks suggest a complex ethical landscape where we still struggle to align values and ideas of what ethical AI should be. However, in an attempt to synthesize the principles found in our literature review, we present five aggregated principles that represent the main concepts for ethical AI development in our context: beneficence and non-maleficence, fairness and justice, human autonomy and agency, transparency and explainability, and accountability and oversight. These align with the findings of both Corrêa et al. (2023) and Hagendorff (2020), who have conducted separate analyses of 200 and 22 ethical guidelines, respectively. Principles that have been frequently mentioned but not included in our synthesis include privacy, data protection, and safety. These principles are not emphasized in our study because they mainly concern the technical solution and can be operationalized mathematically, therefore playing a less visible role in influencing users. Furthermore, there has already been extensive research on these topics, making them more well-documented and understood than the five we will focus on (Hagendorff, 2020, p. 103).

### *Beneficence and non-maleficence*

According to Thiebes et al. (2021), beneficence is concerned with the development, deployment, and use of AI in a way that promotes well-being of humans and the environment while respecting basic human rights (p. 451). This principle emphasizes that AI systems should actively contribute to positive outcomes, such as enhancing health, education, and sustainability. Non-maleficence, on the other hand, focuses on avoiding harm, emphasizing the protection of people's privacy, security, and safety (Thiebes et al., 2021, p. 452).

Although presented as two distinct principles of ethical AI, we have combined them into an aggregated principle due to their shared focus on impact. Ensuring the principles of beneficence and non-maleficence requires organizations to consider both the positive and negative effects of their systems. This includes environmental issues, such as emissions from deploying AI, and societal issues, such as AI-driven chatbots providing genuine support rather than exploiting users for data (Thiebes et al., 2021, p. 452)



### *Fairness and justice*

The terms fairness and justice, often used interchangeably, emphasize non-discrimination and bias mitigation, ensuring that AI systems treat individuals fairly regardless of their sensitive attributes (Corrêa et al., 2023, p. 6). This is an important principle mentioned in most ethical guidelines (Hagendorff, 2020, p. 103; Corrêa et al., 2023). These guidelines largely exhibit the same perspective on what justice entails, though with slight variations that can be generalized into three main points: using AI to correct past inequities like discrimination, ensuring that the benefits of AI are shared and distributed fairly, and preventing AI from creating new harms and inequities. Central research themes concerning fairness and justice include identifying racial and other biases in current AI systems, quantifying fairness as well as lack of fairness in AI systems, and developing approaches to mitigate or avoid bias (Thiebes et al., 2021, p. 454).

### *Human autonomy and agency*

The principles of human autonomy and agency is concerned with preserving the autonomy of human decision-making during human-AI interactions (Corrêa et al., 2023, p. 6). According to Thiebes et al. (2021), the concept of autonomy in AI ethics includes promoting human autonomy, agency, and oversight while also considering the necessary restrictions on the autonomy of AI systems when required (p. 454).

Autonomy is also a crucial ethical consideration in persuasive technology. According to Benner et al. (2022), ethical persuasive design must protect users' autonomy, ensuring that systems influence users without coercing or manipulating them towards decisions that serve the interest of the designers over the well-being of users (p. 554). In practice, ensuring human autonomy and agency means organizations should implement proper oversight mechanisms, such as keeping a human-in-the-loop, to ensure AI systems support rather than replace human decision-making (Thiebes et al., 2021, p. 454).

In the book “Ethics of Digital Well-Being,” Burr and Floridi (2020) demonstrated how a model called METUX (Motivation, Engagement, and Thriving in User Experience) could be used to evaluate the experience of autonomy within technology environments, specifically the use of YouTube. The model is one of the most comprehensive frameworks for evaluating digital well-being and is based on self-determination theory (SDT) - psychological research conducted by Ryan and Deci in 2017. According to SDT, human well-being hinges on fulfilling three basic psychological needs: autonomy, competence, and relatedness. However, the focus here is specifically on autonomy due to its significant overlap with machine autonomy and its importance in ethical AI principles.

### *Transparency and explainability*

Transparency and explainability are essential principles in the ethical development and deployment of AI technologies. They ensure that the use and development of AI are clear and understandable to all stakeholders, including non-experts, and allow for necessary audits. Transparency can refer to the openness of an organization or the clarity of an algorithm, both of which are crucial for building trust and accountability in AI systems (Corrêa et al., 2023, p. 6).

Shin (2023) underscores the importance of transparency in personalized algorithms, meaning users should clearly understand the recommendations made by these systems. He relates algorithmic transparency to visibility, explainability, and interpretability, ensuring that the decision-making processes and intentions behind algorithms are both interpretable and accountable (Shin, 2023, p. 20).

Endsley (2023) highlights the ironies of AI, noting that as AI becomes more intelligent and adaptive, it paradoxically becomes harder for people to understand these systems (p. 1658). This lack of transparency not only reduces user trust but also complicates fault detection and mitigation for machine learning engineers (Panchal & Panchal, 2023). Several other studies point to this phenomenon as well, noting that AI systems often appear as "black boxes" with unclear internal workings (Thiebes et al., 2021, p. 455; Wang et al., 2021, p. 765; Kreps et al., 2023, p. 3; Wulf & Seizov, 2020, p. 617; Panchal & Panchal, 2023; Smuha, 2021; Bélisle-Pipon et al., 2023). The "black box" effect refers to the complexity and lack of interpretability in AI recommendation systems, which can lead to extreme user reactions, such as either distrust or over-reliance (Wang et al., 2021, p. 762).

Wang et al. (2021) found that interpretability significantly boosts user trust, perceived usefulness, and ease of use, and that transparency and procedural fairness are critical for ensuring user trust and system usefulness. This shows that users need to understand how and why recommendations are made to build trust in the system (Wang et al., 2021, p. 762). Supporting this view, Wulf and Seizov (2020) argue that AI development should shift focus from mere computing and predictive power to explainability and that AI explanations should be proportionate to the impact of its decisions (p. 621). Moreover, interdisciplinary cooperation is crucial for addressing the complexities of AI transparency and explainability (Wulf & Seizov, 2020, p. 622).

Explainable AI (XAI) is a research field aiming to make AI systems more understandable to end users, with techniques classified such as simplification, feature relevance, local explanations, and visual support (e.g., heat maps). Tools like LIME and SHAP help visualize how models make decisions, while approaches like Google Model Cards provide transparency about model training and testing data. Furthermore, designing inherently interpretable machine learning (ML) systems, such as rule-based or decision-tree models, enhances transparency. Subsequently, XAI is evolving to improve both the technology and the ways end users interact with AI systems (Laato et al., 2022). However, even though ML

systems improve on data handling, they often create opaque "black box" models that challenge explainability and transparency, highlighting a gap in providing meaningful, stakeholder-specific explanations that uphold the principles of responsible AI (Panchal & Panchal, 2023).

### *Accountability and oversight*

Lastly, the principle of accountability and oversight emphasizes that developers and deployers of AI technologies must comply with regulations and be responsible for their actions and the impacts of their technologies (Corrêa et al., 2023, p. 6). This ensures that in the case of an AI failure, someone can be held legally responsible (Thiebes et al., 2021, p. 455). The responsibility is enforced by oversight, with both internal and independent oversight being highlighted as important aspects of ensuring reliable, safe, and trustworthy AI (Shneiderman, 2020a, 2020b). According to Hagendorff (2020), a significant challenge in ensuring accountability is the combination of distributed responsibility and a lack of understanding of long-term or broader societal technological consequences (p. 113). This leads to a decreased sense of accountability and less awareness of the moral significance of their work among software engineers.

### **Ethics in practice**

An issue frequently addressed in the literature on AI ethics is its real-life application. Despite growing awareness and debate around AI ethics, it is apparent that a significant gap remains between theoretical guidelines and their practical implementation.

One critique is presented by Thiebes et al. (2021), who point to the generality of current ethical frameworks and note that they provide little practical guidance for implementation (p. 455). This vagueness often leaves professionals struggling to integrate these principles into their daily work effectively. Similarly, Ibáñez and Olmeda (2022) express their concern about the excessive focus on debating ethical principles rather than providing clear, actionable instructions for developers on how these principles can be practically applied in real-world settings (p. 1663). Specifically, Ozmen Garibay et al. (2023) call for training, codes of practice, standards, regulations, and/or legislation for achieving responsible AI in practice (p. 424). Another approach to this is presented by Hagendorff (2020), who advocates for transforming ethics into "microethics," which means creating more specific technology ethics related to technical disciplines such as machine, computer, data, and information ethics (p. 111).

The effectiveness of ethical guidelines is further compromised by the lack of enforcement mechanisms. Gutierrez (2023) discusses how many soft law programs lack incentives or clear methods for enforcement, which weakens the overall governance of AI ethics (p. 796). He calls for more research into how these guidelines can be enforced and how the application of AI affects stakeholder behaviors (p. 796). Hagendorff (2020) points to how this results in AI ethics being regarded as non-essential and an optional addition to technical concerns (p. 113).

Koniakou (2022) argues that human rights are identified as a suitable foundation for AI governance, with the UN human rights system providing a robust framework for making AI regulation human-centric and actionable. However, implementing this in the highly privatized AI industry is challenging due to the traditional state-centric model and jurisdictional limitations that affect enforcement against multinational corporations.

#### 2.2.4. AI stakeholders

##### **The global race of AI regulation**

As previously mentioned, soft law plays a critical role in guiding the development and implementation of AI technologies, allowing for a flexible and adaptive regulatory approach that can quickly respond to technological advancements. However, because of its voluntary nature, the AI landscape also requires the stability and enforceability that only hard law can provide (Gutierrez, 2023). Subsequently, as emphasized by an independent expert in the Council of Europe's Ad Hoc Committee on Artificial Intelligence (CAHAI) and the OECD Network of Experts on AI (ONE AI), the race to AI that we have seen over the last few years has also triggered a race to AI regulation (Smuha, 2021)

Several countries and regions have been actively participating in a global "race to AI," each striving to advance and harness the benefits of AI technology more rapidly and effectively than others. Notable examples include China's "Next Generation AI Development Plan" launched in 2017, aimed at achieving world leadership in AI; the European Union's AI strategy of 2018, which seeks to pioneer the development and use of AI for the benefit of all; and the American AI strategy initiated in early 2019, focused on enhancing the nation's AI leadership. International organizations, including the OECD, the Council of Europe, and UNESCO, have entered the regulatory arena, uniting a varied group of countries to collaborate on reaching a consensus regarding ethical issues presented by AI. The motivation behind this race is not solely about gaining a competitive edge in the global market but also involves considerations of national and economic security, suggesting an almost existential necessity for advancing AI. The competition in AI advancement has led to a concurrent "race to AI regulation," as governments face the challenge of finding an appropriate balance between protection and innovation amid uncertainties regarding the technology's impacts, the effects of regulatory interventions, and the consequences of inaction (Smuha, 2021).

To address these challenges, the EU published the European General Data Protection Regulation (GDPR) in 2016, encompassing principles that help protect human dignity and prevent discrimination by allowing individuals to understand and challenge AI decisions. As AI integrates into key areas such as employment, education, healthcare, and surveillance, the GDPR ensures it adheres to strict privacy and data protection standards, emphasizing transparency, legality, fairness, and data minimization. Considered a competitive disadvantage for Europe in the AI race, the GDPR has received criticism, arguing that it placed Europe behind China and the USA in AI development. However, it has also been

widely acknowledged that the GDPR succeeded as a regulatory export and contributed to a gradual convergence in global AI regulation (Smuha, 2021).

The significance of AI ethics committees and regulatory bodies is emphasized in a paper by Hoxhaj et al.(2023), according to which such stakeholders are essential in ensuring GDPR compliance, setting guidelines that govern AI's application to protect privacy and data integrity. Instruments like the European Commission's Ethics Guidelines for Trustworthy AI support these efforts, promoting international standardization and addressing global AI challenges.

Most recently, in March 2024, the European Parliament adopted the Artificial Intelligence Act (AI Act), introducing a risk-based regulatory framework for AI systems in the EU, classifying them into four categories: minimal or no risk, limited risk, high-risk, and unacceptable risk, each with tailored obligations and requirements. This approach, distinct from the Chinese government-centric and the US industry-owned data approaches, emphasizes a human-centric regulation focusing on AI usage rather than the technology itself. Key ongoing considerations include amendments for fundamental rights impact assessments, prohibited practices, the use of copyrighted content, and regulations for General Purpose AI Systems (GPAIS), which are versatile and can perform tasks beyond their initial training (Díaz-Rodríguez et al., 2023).

Over thirty countries, including major economies like the UK, France, Germany, and China, have implemented national AI strategies emphasizing ethical AI. At the same time, professional bodies and civil society organizations have also proposed various ethical principles to guide positive AI development. Additionally, several non-state actors, notably major technology companies such as Facebook, IBM, Microsoft, and others, have issued their ethical declarations, principles, and codes to express their commitment to the ethical and responsible development and deployment of AI and algorithms, both individually and through collective initiatives like the Partnership on AI. These companies often craft these narratives to reduce reputational risks, highlighting the necessity for stringent, enforceable regulations that work alongside ethical guidelines to guarantee the responsible incorporation of AI into society (Koniakou, 2022). A challenge of pursuing such hybrid governance, where companies and technical groups define ethical implementation, risks reducing complex ethical issues to simplistic, quantifiable terms. This "technical solutionism" may not adequately address societal needs, potentially distorting the human-centric goals of AI ethics frameworks. Therefore, transnational ethical initiatives are shaping a common AI ethics discourse, heavily influenced by private companies and tech communities focused on creating "trustworthy" AI tools (Palladino, 2022).

## **Stakeholder engagement**

On a national and regional level, Bélisle-Pipon et al. (2023) criticize the lack of transparency and inclusiveness in formulating AI ethics guidelines, which often exclude diverse and independent stakeholders. While ethical guidelines for AI come from diverse sectors,

including academia, industry, government, and civil society, there is a significant shortfall in stakeholder engagement; only 38% of guidelines involve stakeholders, and a mere 9% actively involve citizens. Documents that engage stakeholders tend to offer more comprehensive ethical guidance, highlighting a correlation between inclusiveness and the quality of guidance produced. The private sector shows the least tendency for stakeholder involvement. Overall, this notable lack of transparency and engagement with the general public in the creation of AI ethics guidelines underscores a need for greater inclusion of stakeholders and citizens in shaping ethical and societal norms for AI (Bélisle-Pipon et al., 2023).

A deliberative framework for responsible innovation in AI by Buhmann and Fieseler (2021) proposes three key stakeholders that should be involved, namely organizations that develop and apply AI, civil society actors, and investigative media in exploring the pathways for responsible AI innovation. Another piece of literature discusses the aspect of explainability in AI and identifies the following distinct stakeholders in AI systems: end users, including those who directly interact with and those affected by AI decisions; regulatory bodies; business overseers; and developers (Laato et al., 2022). In the context of Explainable AI (XAI), Panchal & Panchal (2023) name various stakeholders, including managers, domain experts, data scientists, ML engineers, developers, and regulatory bodies, who are responsible for ensuring system completeness, improving model performance, aiding in debugging and compliance checks, and fostering the development of new functionalities and methods. At the same time, Bélisle-Pipon et al. (2023) argue that the dominance of private sector and academic actors in AI ethics overshadows citizen participation, potentially reflecting a paternalistic attitude or systemic oppression.

## **Key responsible stakeholders**

The main responsible stakeholders identified and called out in a paper written by Ozmen Garibay et al. (2013) are researchers, developers, business leaders, and policymakers. They support this by arguing that stakeholders are everyone who actively contributes to the research, development, design, implementation, and regulation of AI technologies (Ozmen Garibay et al., 2023, p. 428). The product of the paper is six types of call-for-action recommendations for the four stakeholders in relation to human well-being, responsible design of AI, privacy, design, governance, and HCAI. Ibáñez and Olmeda (2022) highlight a challenge for the combined ecosystem of stakeholders and end-users in implementing proper AI ethics, noting that “*tensions in society may be produced for the stakeholders involved such as industry actors, developers, academics, government officials, and end-users*” (Ibáñez & Olmeda, 2022, p. 1664).

Researchers, developers, business leaders, and policymakers each have critical roles in advancing AI responsibly. Based on the action recommended by Ozmen Garibay et al. (2023), researchers should study AI's impacts, promote well-being, support cognitive capacities, ensure inclusivity, transparency, and ethical AI design, and improve AI governance and human-AI interaction. Developers must adapt AI to human needs, ensure

ethical practices, safeguard data, and provide mechanisms for human control and feedback. Business leaders need to promote well-being, prevent negative side effects, ensure transparency, and integrate human-centered design principles while enhancing AI governance. Policymakers should focus on ethical AI practices, protect user data, and ensure inclusivity and transparency in AI applications. All stakeholders must coordinate to standardize AI design, support ethical practices, and foster inclusiveness and accountability in AI development and implementation (Ozmen Garibay et al., 2023). However, challenges of such actions persist as *“coercion, self-interest, and legitimacy, are all active enforcement mechanisms at work in the AI governance ecosystem”* (Palladino, 2022, p.4).

In light of these findings, we have identified the following stakeholders in AI ethics: companies, developers, researchers, governments (and third-party auditors), and users (or consumers). However, for the purpose of the study, we will mainly focus on the first four as these play active and responsible roles in the development, deployment, and regulation of AI technology. We chose to use the word “companies” to encompass various terms used in the literature, such as “businesses,” “business leaders,” and “industry actors.”

### *Companies*

When it comes to the corporate sector, Buhmann & Fieseler (2021) suggest that due to self-interest, power imbalances, and information advantages, corporations are unlikely to tackle AI issues in a deliberative manner. The disclosure and sharing necessary for open participation may create agency problems for corporate entities by offering competitive advantages to rivals. However, corporations can benefit from engaging in these processes by acquiring critical information, facilitating learning, enhancing their reputation, and protecting against criticism of irresponsible behavior, thus managing reputation and risk effectively. Therefore, epistemological and reputational concerns are key motivators for businesses to participate in responsible innovation discussions (Buhmann & Fieseler, 2012). Other than reputational issues that may arise from the deployment of AI systems, private companies are concerned about the legal and financial risks from product failures, prompting a growth in initiatives to set ethical and governance principles (Palladino, 2023). Effective scaling of AI applications is closely tied to responsible AI governance. However, companies are facing challenges with model proliferation and team turnover that can result in loss of core functions and increased biases. A clear AI strategy, diverse teams, and ethical frameworks are essential for successful deployment and maximizing ROI (Eitel-Porter, 2021).

When it comes to legal risks, Gutierrez (2023) describes the avoidance of hard law repercussions as a strong incentive for companies to comply with AI regulation. Moreover, *“government-directed soft law can nudge, influence, or compel stakeholders to act in ways that benefit public objectives”* (Gutierrez, 2023, p.793). The reward-seeking behavior of companies, therefore, makes them open to implementing soft laws, such as guidelines and standards, as long as *“they are convinced that doing so will provide a concrete and worthwhile benefit”* (Gutierrez, 2023, p.795).

## *Researchers*

According to the reviewed literature, researchers have made minimal effort to ethically assess digital nudge design, particularly concerning fairness and autonomy (Benner et al., 2022, p. 554). Furthermore, researchers' interpretation of fairness also differs from "naive" users, who may interpret fairness in terms of social application. For instance, scientists expect XAI systems to ensure fairness in model performance (Panchal & Panchal, 2023).

On the other hand, Benner et al. (2022) also noted that researchers have highlighted for quite some time the apparent lack of ethical designs and policies that can be translated into law to regulate data-hungry companies. This deficiency can lead to unethical designs and outcomes that negatively impact user well-being (Benner et al., 2022, p. 550).

## *Governmental institutions*

Smuha (2021) mentions subsidies or tax benefits for AI development firms, alongside policies that simplify the immigration of AI-skilled workers to stimulate AI advancement, as government incentives for regulation that are enabling rather than protective. When it comes to the latter, governments aim to ensure ethical AI use and safeguard the interests of stakeholders in various sectors or scenarios through mandatory transparency requirements and ethical guidelines for AI developers and users (Smuha, 2021).

Another motivating factor for governments to impose ethical guidelines is the growing trend of AI ethics becoming a fashionable concern, with many organizations and governments being especially focused on staying competitive in the AI race and the market of ethical AI principles (Bélisle-Pipon et al., 2023). The emphasis on ethical and human rights-based AI development is also driven by strategic and pragmatic considerations, where governments fear public distrust could hinder technology adoption, similar to past examples like GMOs and nuclear power (Palladino, 2023).

When it comes to government-provided guidelines, *"in many cases, this consists only of guidelines fitting on one page that do not offer practical benchmarks, and do not appear to have been developed in consultation with experts and community members"* (Bélisle-Pipon et al., 2023). This statement suggests that often governments fail to consult with industry experts sufficiently when making policies, which creates significant challenges in the adaptation of these and necessitates increasing stakeholder engagement.

## *Developers*

According to Palladino (2023), AI developers and deployers tend to consider the effects of their technologies on individuals more thoroughly. He reasons that *"this is likely because it is more direct and evident, generates more social pressure, is more likely to have legal and economic consequences, and could undermine the development of the sector"* (p. 2). Furthermore, researchers and ethical guidelines emphasize the need for AI to be



understandable, explainable, and intelligible, reflecting a broad consensus on the importance of these features.

However, the results from a survey conducted with 70 developers who were tech professionals and Australian undergraduate students in HCI, software engineering, and ML suggest that there might be a disconnect between developer priorities and user experiences. Developers consider making AI understandable a low priority (6%), and it is rarely mentioned by users about their experiences (0% for positive, 4% for negative). Developers do prioritize ethics (20%), privacy (11%), and security (6%), aligning somewhat with ethical AI guidelines. In contrast, users infrequently mention privacy (5%) and ethics (4%), mostly in negative contexts, highlighting a disconnect between developer priorities and user experiences. The study demonstrates that while developers generally align with HCAI guidelines and are aware of AI risks, there remains a significant gap between HCAI's theoretical objectives and its practical implementation. Developers recognize the importance of enhancing AI functionality from users' perspectives but need a deeper understanding of AI's social impacts, which are crucial for improving user experiences and shaping AI ethics (Bingley et al., 2023).

### 2.2.5. Summary of findings

In the literature review, we have synthesized and presented the main themes that emerged during the assessment of the literature selected. The purpose of this synthesis was to gain a deeper understanding of the current research within the field of AI ethics and how users might be influenced while interacting with AI technology. This effort aimed to answer the first research question: *What are the key ethical considerations associated with using AI to influence user behavior?*

The literature has provided several perspectives on ethics in AI technology and revealed a multitude of different principles. However, the key ethical considerations most prominent in relation to AI influencing user behavior can be summarized in five aggregated principles: (1) beneficence and non-maleficence, (2) fairness and justice, (3) human autonomy and agency, (4) transparency and explainability, and (5) accountability and oversight. These principles are critical in guiding the development of ethical AI systems, providing an answer to RQ1.

However, the literature also highlighted a significant gap between these theoretical principles and their practical implementation. Hagendorff (2020) observed that ethical guidelines currently have minimal influence on the decision-making processes within AI and machine learning, suggesting a need for more impactful frameworks (p. 99). This concern is also particularly relevant in the domain of persuasive technology. A notable finding from Benner et al. (2022) highlights the lack of ethical design knowledge in persuasive system design, which can lead to unethical outcomes negatively impacting user well-being (p. 550). Addressing these issues within persuasive technology is crucial for advancing ethical AI development.

Furthermore, we found that there is a lack of consensus on the ethical principles and how they should be implemented among AI stakeholders. Based on the reviewed literature, we have identified four key stakeholder groups, which can be summarized as companies, developers, governments, and researchers. These stakeholders play vital roles in implementing ethical AI practices, yet there is a need for a clearer definition of responsibilities and collaborative efforts to ensure comprehensive and enforceable ethical standards. Additionally, the reviewed literature only marginally examined the standpoints and attitudes of the stakeholders in relation to ethical principles in persuasive AI technology.

Based on the insights gathered from our literature review, we can conclude that there is a large coverage of ethical principles related to AI technology and its impact on users. However there is a consensus in the field that further research is needed to effectively bridge the gap between ethical theory and practice. Given this finding and our understanding of the different stakeholders through the literature, we believe there is a need to further examine how the various stakeholders relate to the key ethical principles outlined. This insight will guide our further research, putting a focus on the four stakeholders in relation to the ethical principles and to the other stakeholders involved.

### 3. Methodology

In this section, we examine the research design and methods employed in our study to investigate the viewpoints of AI stakeholders about current practices for addressing the ethical implications of AI-driven persuasive technology (RQ2). Furthermore, we explore the rationale for choosing these methods over other alternatives. We do this by evaluating the pros and cons of various approaches to selecting the most appropriate methodology, ultimately choosing those that best fit the scope, resources, and duration of the study.

#### 3.1. Philosophy of Science

Human-Computer Interaction (HCI) is a dynamic and interdisciplinary field, incorporating diverse disciplines like industrial design, engineering, computing, psychology, and linguistics. This mix promotes integrating and adapting concepts and theories from various fields, continuously advancing HCI's state-of-the-art. However, this diversity also leads to different views on the nature and future direction of HCI research and how the field can progress (Duarte & Baranauskas, 2016). In his discussions on research methods, Bryman (2016) emphasizes the importance of understanding ontological and epistemological considerations in the social sciences, which we further discuss in the following paragraphs.

In this section, we present ontological and epistemological considerations and how we apply them throughout the study. According to constructivism-interpretivism from an ontological perspective, reality is subjective and context-dependent, with multiple local and specific realities constructed differently (relativism). From an epistemological perspective, “*reality is*

*socially constructed and, therefore, the interaction between researcher and participant is central to capturing the participant's lived experience"* (Duarte & Baranauskas, 2016). Therefore, it is crucial to conduct the study in a way that enables participants to share their perspectives as authentically as possible.

Bryman (2016) defines constructionism, also known as constructivism, as an ontological position asserting that social phenomena and their meanings are actively constructed by social actors through interaction and are continuously revised. It emphasizes that the descriptions of the researchers of the social world are also constructions, presenting specific versions of reality rather than definitive ones. This view aligns with postmodernism, suggesting that knowledge is indeterminate. Constructionism challenges objectivism and, in its focus on the nature of knowledge, also opposes realism. The term primarily relates to how social phenomena and categories are seen as socially constructed, highlighting the subjective nature of knowledge in both the social and natural worlds (Bryman, 2016). Social constructionism, rooted in classical subjective idealism and constructivism, diverges from traditional scientific methods by emphasizing narrative analysis over experiments, as advocated by figures like Kenneth Gergen, who critiqued the empirical basis of social psychology and argued that the human behavior's unpredictability challenges the notion of fixed scientific laws, thus promoting the view that knowledge and truth in social psychology are socially constructed and subjective (Rosenthal & Rosnow, 2008). By exploring how stakeholders collectively construct meanings related to ethical AI, we gain insights into the shared understandings and divergent perspectives within the stakeholder community. The AI landscape continuously evolves with countless actors and influencing factors. Understanding the interplay between individual viewpoints and collective norms, therefore, contributes to a nuanced understanding of ethical considerations in AI.

Interpretivism reflects the distinctiveness of humans against the natural order. It requires the researcher to grasp the subjective meaning of social actions and to immerse themselves in the social context studied. This position contrasts sharply with positivism, which seeks objective truths through quantitative measures and applying scientific methods (Bryman 2016). Interviewing stakeholders allows researchers to understand how different individuals or groups perceive ethical issues in AI differently. We can delve into the stakeholders' worldviews, values, and beliefs regarding ethical AI through open-ended questions and probes. Furthermore, it is crucial to reflect on biases, assumptions, and subjectivities throughout the research process, especially when interpreting collected data. When interviewing stakeholders about ethical AI, it is crucial to consider that our own perspectives may influence the interpretation of the data. Finally, we will critically examine our role in shaping the interview process and the interpretations drawn from stakeholders' responses.

As we rely heavily on qualitative data, we aim to empathize with the subjects of the study and strive to understand their perspectives from a social viewpoint. We are particularly interested in the nuances of human behavior, motivations, and the meanings behind social phenomena. By adopting an interpretive approach, we can produce an in-depth understanding of how ethical considerations in AI are perceived and managed, acknowledging the complex,

subjective realities that influence stakeholders' views and decisions. Furthermore, applying interpretivism in interviews allows us to be flexible and adaptive in our approach, tailoring questions to follow interesting leads or explore unexpected areas that arise during conversations. This can lead to a more comprehensive understanding of the ethical landscapes in AI. When it comes to methodological stance, we engage in an empathetic interaction with the object of the investigation to collect qualitative data, which we later analyze and interpret based on the theories described in section 4. Theoretical Framework.

In summary, a constructivism-interpretivism approach to interviewing stakeholders about ethical AI offers a holistic framework for exploring the complex interplay between social dynamics, individual perspectives, and contextual factors in shaping ethical considerations in AI governance, development, and deployment.

## 3.2. Research Design

This research entails a comparative design primarily focused on comparing the perspectives of the four different AI stakeholders (governments, researchers, companies, and developers) to identify differences or similarities based on predefined criteria, namely the ethical principles identified in the literature review.

As described by Bryman (2016), *“the key to the comparative design is its ability to allow the distinguishing characteristics of two or more cases to act as a springboard for theoretical reflections about contrasting findings”* (Bryman, 2016, p. 68). The core of this approach is its emphasis on systematic comparison of various aspects of the subjects or cases under study to discern patterns, causes, or effects. This might involve comparing different organizational structures, leadership styles, policy impacts, cultural behaviors, or any other variables of interest across different settings or groups. The goal of this research is to enhance the understanding of the ethical AI phenomena by examining the approaches, challenges, and attitudes of different AI stakeholders when it comes to implementing ethical principles.

An alternative research design that could have been suitable for the topic is a cross-sectional design, which typically emphasizes the collection of data at a single point in time across a variety of subjects or phenomena. This approach captures a snapshot of a particular aspect, such as attitudes, behaviors, or conditions. The key characteristic of a cross-sectional design is its ability to provide a broad view of a population or multiple subgroups within a population at one instance. This design is beneficial for studying the prevalence of characteristics or for making comparisons between different groups within the sampled population at the time of the study (Bryman, 2016). Therefore, it could have been suitable for gaining an understanding of the attitudes, approaches, and challenges of the different stakeholder groups, and the association between these. Furthermore, a comparative design can often act as a hybrid, potentially being used as an extension of a cross-sectional design. However, the cross-sectional design is rather used for quantitative research *“in order to collect a body of quantitative or quantifiable data in connection with two or more variables*

(usually many more than two), which are then examined to detect patterns of association” (Bryman, 2016, p. 53). Considering that ethical principles are fundamentally qualitative, as they are based on values, norms, and moral considerations rather than numeric measurements, and the goal of the study is to gain a deeper understanding of the stakeholder perspectives, a qualitative approach was deemed more suitable for this research.

### 3.3. Data Collection

#### 3.3.1. Data collection methods

Comparative research is highly flexible and can accommodate both qualitative and quantitative approaches or a combination of both (mixed methods) (Bryman, 2016). As previously mentioned, we decided to collect qualitative data primarily because of the depth of qualitative data we aimed to examine.

When evaluating which data collection method would be most suitable for this study, we considered both questionnaires and interviews. Questionnaires or surveys are more commonly used in quantitative research to collect large amounts of data that can be statistically analyzed to compare frequencies, relationships, and trends across different groups or populations. On the other hand, conducting interviews allows researchers to gather detailed and nuanced information directly from participants, providing insights into their thoughts, perceptions, and experiences (Bryman, 2016).

When comparing self-completion questionnaires to interviews, the former tend to have fewer open questions that are easier to answer and have easy-to-follow designs, often to facilitate subsequent quantitative analysis. If shorter, they can reduce the risk of ‘respondent fatigue.’ Among other advantages are cheaper and quicker administration, more convenience for the respondents to complete according to their own pace and time, and the absence of interviewer effects. The latter suggests that “*characteristics such as ethnicity, gender, and the social background of interviewers may combine to bias the answers that respondents provide*” (Bryman, 2016, p. 222). However, few consistent patterns have emerged through research that specify what those biases might be. On the other hand, questionnaires primarily lack a facilitator and the possibility of asking clarifying questions from either side. They also tend to have lower response rates and a higher risk of missing data (Bryman, 2016). Since this study aims to dive deeper into the perspectives of different stakeholders, it would be challenging to ensure there is no missing data, as well as not being able to ask complementary questions. Therefore, the interview is a more suitable method for collecting data in this case.

Other data collection methods taken into consideration were ethnography (or participant observation) and case study (or “case”). The former is a research method where the researchers immerse themselves in a social setting for an extended period, observing behaviors, engaging in conversations, interviewing informants, collecting documents, and understanding the culture to ultimately write a detailed account of the setting (Bryman,

2016). Both ethnography and semi-structured interviewing are used in a way that the researchers “*have to orient their observations to that research focus, but at the same time maintain a fairly open mind so that the element of flexibility that is a strength of qualitative research is not eroded*” (Bryman, 2016, p. 441). They are inductive approaches to theorizing and enabling an exploratory approach to research that is more open-ended and less restricted when it comes to study findings.

Ethnographic research, however, would entail long periods of observation in an organization or a company, and even though micro-ethnography is an option, it would be challenging to observe participants in confidential settings such as governments, research centers, or even companies that do not wish to disclose details about their procedures. Therefore, interviews allow the researchers to ask participants directly about their perspectives and are more suitable for collecting qualitative data in the short period of time that is available for this study.

When it comes to interviews, it was relevant to consider both individual and focus group interviews as our goal was to gather data from specific stakeholder representatives and whether it should be structured, unstructured, or a hybrid of these semi-structured. The focus group technique is essentially a group interview where participants explore a specific theme or topic in depth. However, there can also be a distinction between group interviews and focus groups, the main difference being that the former is conducted so that it saves expenses and time for the researcher, while focus groups are conducted to discuss a certain issue as a member of a group, rather than simply as individuals (Bryman, 2016). Focus group interviews include observing interaction and discussion among participants, which can reveal the dynamics of social processes and generate a diversity of viewpoints and ideas. On the other hand, individual interviews provide deeper insight, more privacy, less influence from group dynamics, greater flexibility, and better control over the interview process than group interviews. Since the aim of the research is to discuss in-depth personal and professional views on ethical principles, it is crucial to prevent potential biases or participants being influenced by each other’s views; therefore, individual interviews were regarded as most suitable for this study.

Bryman (2016) argues that structured interviews are mainly used in quantitative research to maximize the reliability and validity of measuring key concepts. They are inflexible and discourage deviations from the main questions as they are usually regarded as a nuisance. An unstructured interview is a type of qualitative research method where the interview does not follow a strict format or sequence of questions. This approach is more like an open-ended, informal conversation than a formal interview. In comparison, in a semi-structured interview, “*the researcher has a list of questions or fairly specific topics to be covered, often referred to as an interview guide*” (Bryman, 2016, p. 468). As opposed to structured interviews, these are flexible methods with the primary goal of exploring the participant's thoughts and feelings in depth, allowing for discovering new insights that predetermined questions might not reveal. In this research, we focus on exploring the perspectives of AI stakeholders on specific ethical principles. Therefore, a structure is necessary to ensure that each aspect of ethical AI

is being investigated. A semi-structured interview allows us to have this predefined structure but also the possibility to gain further insights into aspects we find relevant to explore further. To make sure that all sub-topics of ethical AI are covered by the interviews and that they enable the researchers to identify and compare patterns in the different stakeholder groups, the questions associated with each topic were compiled in an interview guide (see Appendix B). The guide is further detailed in section 3.3.4. Procedure.

Additionally, Brinkmann and Kvale (2018) differentiate variations of interview types, including factual, conceptual, computer-assisted, narrative, discursive, and actively confronting interviews, each suited for different purposes and requiring different approaches. In this study, our interviews were a combination of conceptual and factual interviews. The purpose of a conceptual interview is to clarify and map out the conceptual structure and meanings of terms - for instance, the principle of “fairness and justice” - exploring their dimensions and interconnections within a conceptual network (Brinkmann, 2007, as cited in Brinkmann & Kvale, 2018). At the same, we aimed to obtain valid factual information in professional and historical contexts by asking about specific practices performed in an effort to ensure ethical AI principles.

### 3.3.2. Research criteria

The quality of research can be assessed and established based on multiple sets of criteria. Bryman (2016) discusses reliability, replicability, and validity as relevant criteria mainly to quantitative research but proposes an alternative set of criteria that should be applied in relation to qualitative research. While it is possible to adapt reliability and validity to qualitative research, the proposed criteria assess the research from the angle of trustworthiness and authenticity (Guba & Lincoln, 1994, as cited in Bryman, 2016).

#### **Trustworthiness**

According to interpretivism, trustworthiness emphasizes understanding the meanings and experiences of individuals from their own points of view and that researchers should immerse themselves in the social context they are studying (Bryman, 2016). Trustworthiness comprises four criteria in qualitative research: credibility, transferability, dependability, and confirmability. These correspond to specific realist criteria from quantitative research: internal validity, external validity, reliability, and objectivity.

Credibility in qualitative research refers to the accuracy and believability of the study and its findings, as judged by the participants, and the consistency of the researcher's observations and conclusions with the participants' realities (Bryman, 2016). There are two techniques to confirm the credibility of the study: respondent validation (or member validation) and triangulation. Pickard (2013) refers to the triangulation of investigators, theory, technique, or sources as a means of establishing credibility in qualitative methodology. In this study, we apply theories and techniques to support and interpret our findings, such as ethical theories, HCI theories, persuasive design principles, and self-determination theory.

Secondly, Bryman (2016) emphasizes the importance of providing rich, detailed descriptions, also called “thick descriptions,” of the research context, participants, and findings. When interviewing AI stakeholders about ethical AI, we provide comprehensive accounts of the interview process, including the setting, characteristics of the stakeholders, and the content of their responses. Thus, we ensure that the results of a study can be generalized or applied to other contexts or settings (Bryman, 2016).

The third criterion, dependability, requires the researcher to account for changing contexts that might influence the findings (Bryman, 2016). To ensure dependability, we documented extensively the problem formulation, participant selection, interview transcripts, and thematic data analysis so they could be inspected, as per Pickard (2013), by an external auditor, i.e., the study supervisor.

Finally, the criterion of confirmability refers to the degree to which the results are shaped by the respondents and research conditions rather than the biases, motivations, or interests of the researchers (Bryman, 2016). To meet this criterion, we establish transparency by attaching detailed descriptions of interview procedures, transcriptions, and analyses. Furthermore, we rely on theories to interpret the viewpoints and motivations of the various stakeholders when it comes to implementing ethical AI.

## **Authenticity**

The criteria from the angle of authenticity address a broader range of issues regarding the wider political impact of research (Bryman, 2016) and evaluate how truthfully the research portrays the diverse realities and experiences of the participants involved. These criteria ensure that the findings represent all viewpoints fairly, contributing to a deeper understanding of the context and the people studied. Therefore, ensuring authenticity, equitable representation of diverse perspectives, and fostering reflection and possible transformative actions in persuasive AI will be crucial criteria for guiding the discussion of the study findings. Concrete steps taken in this regard will be, for example, an extensive and in-depth thematic analysis that considers all data collected equally, regardless of the socio-demographics of the participant.

### **3.3.3. Participants**

In this study, we aimed to gain an understanding of the viewpoints of four different AI stakeholder groups. The stakeholder groups have been selected based on our findings in the literature review, resulting in companies, developers, governments, and researchers. A stakeholder can be defined as anyone affected by the performance of a system and is often categorized into primary, secondary, tertiary, and facilitating stakeholders (Dix et al., 2004, p. 458). In our project, companies and governments can be viewed as both secondary and tertiary stakeholders. This means they are either not directly affected by the system but are involved with its input or output, or they are directly affected by the system’s success or



failure. Developers and researchers are facilitating stakeholders, meaning they contribute to the design, development, and maintenance of the system (Dix et al., 2004, p. 459). We have chosen not to include primary stakeholders, namely the end-users, in our study. This is because our focus is on investigating the stakeholders involved in the development, deployment, and regulation of AI-driven technology and how they perceive ethical implications and responsibility.

The number of interview subjects needed in qualitative studies should be sufficient to uncover necessary insights, balancing between too few subjects, which limits generalizability, and too many subjects, which impedes in-depth analysis (Brinkmann & Kvale, 2018). In accordance, interviewing one participant per stakeholder group might be insufficient in terms of generalizability. Lazar et al., (2017) also highlighted that the amount of participants is a trade-off between the amount of information you collect and the cost of collecting it, further underscoring the need for balance. Therefore, we aimed to interview at least two but not more than four, considering the limited period for the study and the cost of doing two-on-one interviews.

While evaluating which research instrument was most suitable for the study, it is necessary to decide what kind of sampling pairs with it to determine what type of population is suited for the investigation and how the instrument should be administered. Bryman (2016) distinguishes three types of non-probability sampling strategies: convenience sampling, snowball sampling, and quota sampling. As opposed to probability sampling, which is often used in quantitative research where participants are selected randomly and the goal is to estimate population parameters and ensure unbiased representation, in non-probability sampling, the selection is often subjective and based on the judgment of the researchers. The advantage of this type of sampling is that it is often more time-efficient and less expensive than the former. It is particularly useful in qualitative research, where the focus is more on depth and detail than generalizing findings. As the primary goal of this study is to explore new or undiscussed perspectives of stakeholders rather than generalizing stakeholder groups as a whole, non-probability sampling allows us to gain a deep understanding of the phenomena of ethical AI without generalizing findings to the entire population.

The participants for this study have been selected through a mixed sampling strategy based on the four stakeholder groups. The first four participants were chosen by convenience, meaning professionals available to us by accessibility, for instance, through the university or previous work. Then, a snowball sampling was applied, where the previously mentioned participants aided in establishing contacts with an additional three. To increase the generalizability and gain different perspectives from a population with various backgrounds, the remaining two participants were selected based on quota sampling, meaning that the researchers did desktop research on search engines and social media, i.e., LinkedIn, to find potential subjects on a set of criteria, such as age, location, gender, job position, organization type (or stakeholder type), experience level. The goal of quota sampling was to make sure the interview panel was diversified, which would help mitigate biases. Furthermore, we employ purposive sampling to ensure that *“each new research participant contributes characteristics*

*differing from preceding participants. This allows for multiple perspectives on the phenomena under study*” (Pickard, 2017, p.14). However, this sampling strategy proved more challenging in terms of attracting participants. Consequently, despite reaching out to numerous potential candidates, only two agreed to participate in an interview, resulting in only one representative from the developer stakeholder group in our sample. Nevertheless, some of the other participants were able to provide insights from the developer perspective based on previous experience.

### 3.3.4. Procedure

The openness of qualitative interviews, lacking standard procedures, relies heavily on the competence of the researchers, preparation, and ability to make informed, on-the-spot decisions (Brinkmann & Kvale, 2018). Therefore, it is essential for the researcher to understand the methodological options and their implications. Defining clear procedures is also key in ensuring the replicability of a study (Bryman, 2016). In this section, we describe the steps taken to formulate the interview questions and provide instructions on how to collect data through the semi-structured interviews and conduct respondent validation.

When planning an interview investigation, it is essential to clarify the purpose of the study, gain pre-knowledge of the subject matter, and familiarize oneself with various interviewing techniques, as method choices depend on the goals of the study (Brinkmann & Kvale, 2018). The scope of the study is to understand the challenges different stakeholders address in implementing ethical AI, what measures they are currently taking to ensure AI ethics, and what they believe should be done to enhance ethical practices in AI development.

Before asking the interview questions, Brinkmann and Kvale (2018) emphasize the importance of briefing as the first minutes of the interview are decisive. In this stage, the interviewers have the opportunity to learn briefly about the interviewee and introduce them to the goal of the study. Furthermore, at this point it is crucial to ask for permission to audio record the conversation so it can be transcribed for the data analysis and ensure the participants that anonymity would be provided for them as individuals and the organizations they represented. Alternatively, we could have video recorded the interviews, which would allow us to capture and interpret body language. However, this option has been discarded due to the limited scope and period of the study.

In the first part of the interview, we asked questions about the backgrounds of the participants and explored the aspects of AI technologies in their work, specifically whether they had hands-on experience with persuasive technology. Collecting this information helped us in the analysis and discussion stages, ensuring that the participants had diverse backgrounds and evaluating potential biases.

In the second part of the interview, we focused on three aspects of ethical AI principles: (1) what are the current practices applied towards trustworthy AI, (2) what are the main challenges of implementing these practices, and (3) what would be the avenues in addressing

the challenges to ensure ethical principles in persuasive AI technology. This part was divided into the five aggregated principles that emerged during our literature review: (1) beneficence and non-maleficence, (2) fairness and justice, (3) human autonomy and agency, (4) transparency and explainability, and (5) accountability and oversight.

Furthermore, we were interested in the views of the participants on the four stakeholders we identified and interviewed. Therefore, we asked them to elaborate on their professional opinion on the roles and attitudes of companies, developers, governments, and researchers. The aim of this question was to gain a better understanding of how the participants view the different actors in the AI landscape, including their own stakeholder group, when it comes to current practices and perceived responsibilities.

Finally, we prepared follow-up and clarifying questions to determine if there was sufficient time left from the allocated 50 minutes to gain deeper insights into certain aspects, such as personal motivation and the importance of discussing ethical principles when developing AI technology. As a debriefing, we concluded by asking about any other potential interview subjects that the participants found relevant for our research (snowball sampling) and thanked them for their time. These steps have been outlined and detailed in an interview guide (see Appendix B) that was later used while conducting the interviews to make sure that all previously specified topics were covered.

### 3.3.5. Data collection process

Nine interviews were scheduled and conducted online, with the exception of one in-person interview. We used email and LinkedIn for communication and the virtual meeting platforms Google Meet and Zoom at the convenience of the participants. First, the participants were introduced to the scope of the study and requested consent to record the interviews for transcription purposes. We took turns facilitating the interviews, which has been also helpful in mitigating possible biases imposed by the researchers. The interview guide (Appendix B) prepared beforehand has been used as a template for asking the questions, however, the order and formulation have been altered to fit in the context of the discussion and create a natural flow for the conversation. This approach has been applied to ensure that the interview was engaging and did not feel like an interrogation for the participant. While one researcher was interviewing, the other one took up the role of a note-taker and observer. The researcher who did not facilitate the interview had the responsibility to ensure that all areas of interest had been covered and had the opportunity to ask follow-up questions deemed relevant to gain a deeper understanding of certain viewpoints.

The interviews have been conducted with representatives from the four stakeholder groups - two company representatives, one AI developer, two government officials, and four researchers. Demographically, each participant is located in Denmark or is part of a Danish organization. To mitigate possible biases, the participant panel was diversified in relation to age, gender, education, and seniority level, as represented in **Table 4** below:

**Table 4***Demographics of interview participants*

Code	Stakeholder Group	Age	Gender	Education	Position
C1	Company	26	Female	MSc Business Analytics BA Data Science	AI Governance, Risk & Compliance Analyst
C2	Company	48	Male	MSc IT Development	Digital Strategy Director
D1	Developer	25	Male	MSc, Information Studies BA Computer Science	Entrepreneur/mobile and AI Developer
G1	Government	37	Male	MSc, Philosophy and Psychology	Senior Tech Advisor for AI
G2	Government	25	Female	LLM, Law	Head of Section (Tech)
R1	Researcher	47	Male	PhD Robotics and AI MSc Computer Science	Professor, Head of Section, Biorobotics
R2	Researcher	38	Male	PhD Information Studies	Associate Professor
R3	Researcher	44	Male	PhD Recommender Systems and Computer Science MSc Computational Linguistics and AI	Associate Professor and CSO
R4	Researcher	50	Male	MSc Philosophy PhD Information Studies	Professor of data and AI ethics

Finally, we used Microsoft Word as a tool to automatically transcribe the voice recordings, which enabled us to speed up the process and automatically add timestamps. These timestamps were helpful in referencing and keeping track of specific quotes later on. Once the interviews were transcribed and reviewed, the voice recordings were deleted to ensure the privacy of the participants. These transcriptions formed the basis for the subsequent data analysis, as outlined in the following section, 3.4. Data Analysis.

### 3.4. Data Analysis

When analyzing the data, we apply a combination of inductive and abductive approaches that involve observing instances to draw general conclusions. The inductive approach assumes we can examine a consistent entity across various instances to establish broad understanding (Brinkmann & Kvale, 2018). To explore different viewpoints on ethics in persuasive AI technologies, we have selected five specific principles based on which we formulated the interview questions, with the goal of exploring and observing the attitudes of participants in relation to these. However, qualitative interviews can be rather dynamic and unpredictable in certain aspects. Therefore, an iterative approach, abduction, is suitable for understanding the complex attitudes of the human mind (Kvale & Brinkmann, 2019). By combining the two approaches, we can apply prior knowledge from the literature review about ethical principles and the AI ecosystem to set a direction for the desired outcomes and, at the same time, have an exploratory, iterative approach for understanding and interpreting these.

In the next steps, we discuss which type of qualitative data analysis suits the study. Bryman (2016) highlights thematic analysis as one of the most common approaches to qualitative data analysis, while Kvale and Brinkmann (2019) distinguish three modes of interview analysis: focused on meaning, focused on language, and bricolage, which is a mix of different types of analyses. Meaning coding involves attaching one or more keywords to segments of text to facilitate later identification and analysis of statements. This can be either concept-driven, using pre-developed codes based on existing literature, or data-driven, developing codes through iterative readings of the material (Brinkmann & Kvale, 2018). While both thematic analysis and content analysis, which is a type of analysis focused on meaning, involve systematic coding and analysis of qualitative data, they differ in their purposes and approaches. The thematic analysis aims to uncover underlying meanings and patterns within the data (Clarke & Braun, 2017), while content analysis focuses on quantifying and categorizing the content to identify trends or patterns (Kvale & Brinkmann, 2019). Revisiting the research question we are investigating, more specifically, what are the viewpoints of AI stakeholders about current practices for addressing the ethical implications of AI-driven persuasive technology, we are interested in understanding perspectives and the motivations behind on a deeper level, rather than quantifying or identifying trends among stakeholder groups. Therefore, thematic analysis is the most suitable method when analyzing data from the interviews to gain a deep understanding of stakeholder standpoints and the meaning behind these.

### 3.4.1. Thematic analysis

Thematic analysis (TA) is a method for identifying, analyzing, and interpreting patterns of meaning within qualitative data, offering a flexible tool adaptable to various theoretical frameworks and research paradigms, prioritizing an organic approach to coding and theme development within a qualitative paradigm. It offers systematic procedures for deriving codes and themes from qualitative data, where codes represent the smallest units of analysis capturing relevant data features, forming the basis for larger patterns of meaning called themes. The main advantage of TA is its flexibility in terms of theoretical approaches and research designs, allowing for the exploration of patterns within and across data related to participants' experiences, views, and behaviors (Clarke & Braun, 2017). Since our approach is exploratory in the sense that we want to observe the respondents and their attitudes, this type of flexibility allows us to dive deep into different aspects of the theories we employ to support our interpretations.

TA is a versatile method due to its ability to be conducted in various ways. It spans three key continua in qualitative research: inductive versus deductive data coding and analysis, experiential versus critical orientation to data, and essentialist versus constructionist theoretical perspectives (Braun & Clarke, 2006, as cited in Braun & Clarke, 2012).

An inductive approach to data coding and analysis is bottom-up, with an experiential orientation, deriving codes and themes directly from the data content. This method closely

aligns with the semantic content of the data. In contrast, a deductive approach is top-down, where the researcher applies predefined concepts, theories, or ideas to the data. Consequently, these external frameworks influence the codes and themes more than the data itself. In practice, TA often combines both approaches, as it is challenging to be purely inductive or deductive (Braun & Clarke, 2012). After thorough consideration, we decided to combine the two approaches as we have an exploratory strategy to understand the different perspectives of AI stakeholders. At the same time, we employ theories to identify patterns in the qualitative data collected, examining how realities are constructed and the underlying ideas and assumptions in the data, which aligns with our constructionist approach.

### 3.4.2. The analysis process

According to the approach of Clarke and Braun (2017), TA consists of six phases: (1) familiarization with the data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes and (6) producing the report. Additionally, thematic maps aid as visual or text-based tools in identifying and mapping out main themes and their connections (Clarke & Braun, 2017). These stages are not strictly sequential but involve a continuous interplay between conceptualization and data review, providing a general guide to the main elements and their interconnections in TA (Bryman, 2016). Subsequently, the six stages presented by Clarke and Braun (2017) form the basis for our iterative data analysis process of synthesizing the standpoints of AI stakeholders. Furthermore, Richards and Hemphill (2018) emphasize the need to balance rigor, transparency, and trustworthiness in data analysis while managing the challenges associated with analyzing qualitative data in research teams, such as coordination between researchers. In their study, they outline a six-step process for collaborative qualitative analysis, including preliminary organization and planning, open and axial coding, development of a preliminary codebook, pilot testing the codebook, final coding process, and review of the codebook and finalize themes (Richards & Hemphill, 2018). Although not strictly followed, these principles guided our approach and collaborative efforts.

#### **Familiarizing with the data**

The initial qualitative analysis phase involves immersing oneself in the data by reading, rereading transcriptions, and listening to audio recordings. The aim is to become deeply familiar with the data, making informal observational notes to aid in later coding and analysis. These notes serve as memory aids and are typically informal, intended for the researcher's use or sharing within the research team (Bryman, 2016; Clarke & Braun, 2017). In this phase, we also did preliminary organization and planning in line with the collaborative qualitative data analysis process outlined by Richards & Hemphill (2018). We familiarized ourselves with the data when proofreading the transcripts by listening to the original audio recordings. We did this while observing potential patterns or repetitions in each transcription, which allowed us to be more time-efficient.

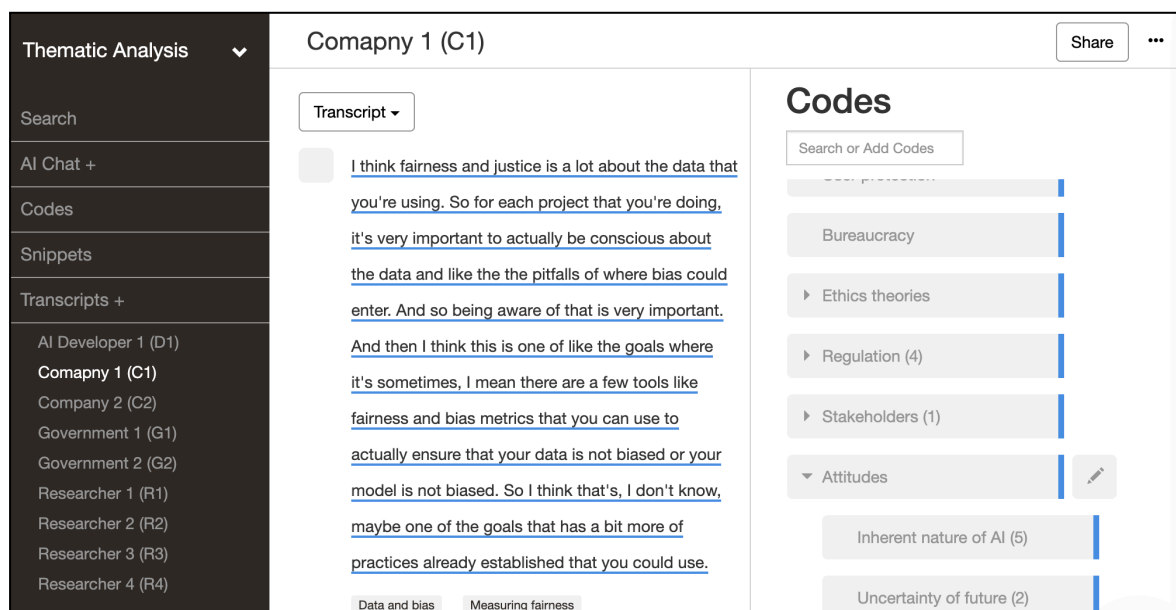
## Generating initial codes

The second phase of TA focuses on coding the data, where codes serve as labels for features relevant to the research question. Coding and categorizing are fundamental methods in social science text analysis. This involves the use of keywords to identify text segments and categorization, allowing systematic conceptualization for quantification, both of which can be concept-driven or data-driven, and necessitate detailed “code memos” for tracking and reflection (Brinkmann & Kvale, 2018). Coding can be semantic, reflecting the content of the data, or latent, offering deeper interpretations. This process includes reading each data item thoroughly, identifying relevant snippets, and creating codes. This phase ends when all data are fully coded and collated (Clarke & Braun, 2017).

In this phase, we used a qualitative data analysis software called Delve to aid the coding process and facilitate efficient research collaboration. It helped address the challenge of collaborative qualitative data analysis being difficult to coordinate as well as time consuming to conduct (Richards & Hemphill, 2018, p. 226). The Delve tool allowed us to highlight snippets of the transcripts and assign suitable codes easily (see **Figure 4**). The codes were then added to a database for our project within the software, allowing us to easily reuse the same code for other parts of the text and keep track of how many snippets have been assigned to each code. This code database was continuously updated for all owners of the project, making it easy for us as researchers to work simultaneously using the same codes. The software also enabled us to arrange the codes into a hierarchical structure, allowing for greater granularity of the data while revealing overarching themes.

**Figure 4**

*Interface of the Delve tool used for coding*



*Note.* The figure represents a screenshot of the Delve tool platform interface, which allows users to collaborate in adding codes to the same transcripts and arranging the codes in a hierarchy. Retrieved from *delvetool.com*

We initiated the coding process by analyzing one interview transcript together, thereby agreeing on an initial outline of codes and aligning on how we assign meaning to the various statements. The main concepts driving the analysis were the five ethical principles derived from related work, accompanied by selected theories, as described in section 4. Theoretical Framework. The remaining transcripts were divided between the researchers and coded individually. However, this was done while sitting together, making sure to stop and discuss whenever we were uncertain of the interpretation of what codes were most suitable. Usually, a snippet or a section of text was assigned several codes as they relate to multiple aspects of our investigation.

## **Searching for themes**

After all the data was fully coded and collated, we moved on to the third analysis phase, where the focus shifted from codes to themes, representing significant data patterns related to the research question. This process involves actively constructing themes by reviewing and clustering similar codes into coherent patterns. Here, miscellaneous codes are temporarily grouped until they fit into themes or are discarded (Clarke & Braun, 2017). When searching for themes, various strategies can be employed, such as looking for repetitions, indigenous typologies, metaphors, transitions, similarities and differences, linguistic connectors, missing data, and theory-related material, which ensures that identified themes are relevant to the research questions and justifiable in their significance and connections (Ryan & Bernard, 2003, as cited in Bryman, 2016, p 586). After coding all of the transcripts, we were left with 669 coded snippets of text distributed on 243 different codes. Since the Delve tool allowed us to group codes into hierarchical structures, some of these codes served as main categories with several sub-categories within them. Subsequently, themes in our data were revealed through the creation of these main categories. Furthermore, since we primarily employed selected theories as a way of triangulation, these formed a solid basis for identifying themes regarding stakeholder attitudes and motivations, in addition to other patterns, such as repetitions, similarities, and differences.

## **Reviewing themes**

The next phase involved a recursive quality-checking process by reviewing and refining developed themes in relation to the coded data and the entire data set to ensure they meaningfully capture the most important and relevant elements in relation to the research question (Clarke & Braun, 2017). Considering the extensive qualitative data collected through interviews, an iterative approach to defining themes ensured that we emphasized the most prevalent opinions and attitudes of the participants and adhered to the study scope and research questions. To achieve this, we went through all of the codes and their related snippets collaboratively, rearranging the codes and the code hierarchy into a clear structure that reflects the main themes and patterns in our data. This thorough process was facilitated by a feature of the Delve tool that makes viewing the data by code possible, presenting only snippets assigned to these. This resulted in the omission and merging of several codes, leaving us with a total of 196 codes.

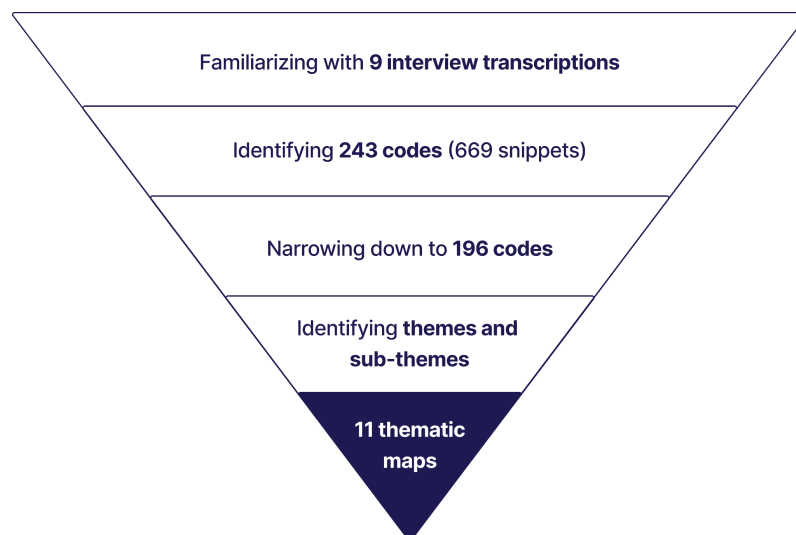


## Defining and naming themes

The fifth phase aims to create a coherent story that integrates detailed analysis with broader conceptual insights. When defining themes, each should have a unique and specific focus, can be clearly summarized, and directly address your research question. Themes should be distinct yet related and may include subthemes for overarching patterns. Furthermore, naming themes is crucial for clarity and should be informative and memorable, often using direct quotes from the data (Clarke & Braun, 2017). Some writers propose that constructing networks of themes and subthemes can effectively portray these interconnections (Attride-Stirling, 2001; Grogan et al., 2013, as cited in Bryman, 2016, p. 588). For this purpose, we employed thematic maps, which are defined as a “visual [...] or sometimes text-based [...] tool to map out the facets of your developing analysis and to identify main themes, subthemes, and interconnections between themes and subthemes” (Clarke & Braun, 2017, p. 60). The thematic maps were created while thoroughly examining the codes collaboratively, agreeing on the main themes, and filling in related subthemes. The TA process outlined in **Figure 5**, resulted in the creation of 11 thematic maps encapsulating the themes of the five ethical principles, the four identified stakeholder groups, as well as users and persuasive technology. These align with the main topics addressed in the interview guide.

**Figure 5**

*Thematic Analysis process: from transcriptions to thematic maps*



## Producing the report

Finally, a report was produced in which writing and analysis are intertwined throughout the process. The aim of the report is to tell a compelling, clear, and complex story that answers the research question, with themes presented logically to build a coherent narrative (Clarke &

Braun, 2017). Our analysis aims to investigate RQ2 by presenting the viewpoints of AI stakeholders in relation to current practices for addressing the ethical implications of AI-driven persuasive technology, and the result of this final phase is presented in section 5. Thematic Analysis.

### 3.5. Validity, Reliability, and Generalizability

Brinkmann & Kvale (2018) reinterpreted the traditional terms of validity and reliability for qualitative research. Reliability is concerned with the consistency and trustworthiness of findings, including whether they are reproducible and consistent across different settings. Validity, traditionally about truth and correctness, in qualitative research refers to whether a method investigates what it intends to. This broader conception allows for valid scientific knowledge from qualitative research by focusing on the quality of knowledge produced and the defensibility of knowledge claims. Instead of fixed criteria, three general approaches for validating interview knowledge are suggested: (1) emphasizing the craftsmanship of the researcher, (2) the communication of findings, and (3) their practical implications (Brinkmann & Kvale, 2018). These three approaches are crucial when reflecting on the quality of the interviews and the findings of the study.

When the findings of an interview study are deemed reliable and valid, the question of their generalizability arises. When it comes to generalization methods, interview-produced knowledge can be approached through statistical and analytical generalization. Statistical generalization requires a random selection of subjects and quantification of findings. However, this method is limited to random samples and cannot apply to self-selected samples (Brinkmann & Kvale, 2018) which has been the main sampling approach for the participants of this study. Analytical generalization, on the other hand, involves assessing the extent to which findings from one study can guide expectations in other situations and requires high-quality, extensive descriptions of the interview process and the results (Brinkmann & Kvale, 2018, p). Ultimately, the responsibility for determining the applicability of findings lies with the information receiver. Therefore, high-quality reporting of interview studies is crucial for facilitating analytical generalization, highlighting the importance of detailed and thorough documentation of the interview process and results (Brinkmann & Kvale, 2018). For this reason, an extensive description of the interview procedure has been provided, and the generalization of the findings has been detailed and supported by in-depth snippets of qualitative data collected (section 6. Discussion).

### 3.6. Ethical Considerations

The popularity of qualitative research interviewing has grown due to internal scientific recognition of its suitability for studying human experiences and external cultural shifts from industrial to consumer societies. This transition emphasizes softer, more empathetic forms of power, making qualitative methods appealing. Despite its rise, qualitative research is often portrayed as inherently ethical, a notion termed "qualitative ethicism." However, this view

can obscure the power dynamics and cultural contexts inherent in research. The asymmetrical nature of interviews, where interviewers control the process and interpretation, highlights the hidden power relations (Brinkmann & Kvale, 2005). Therefore, when conducting social research, it is crucial for qualitative researchers to acknowledge the ethical complexities that accompany the study.

Ethical principles often focus on harm to participants, lack of informed consent, invasion of privacy, and deception (Diener & Crandall, 1978, as cited in Bryman, 2016, p. 125-126). Ethical issues arise throughout the research process, from thematizing the study's purpose to designing with informed consent, considering the interview's impact, ensuring confidentiality in transcription, analyzing with integrity, verifying knowledge responsibly, and reporting while protecting privacy. Each stage requires careful ethical deliberation to balance scientific inquiry with human dignity and welfare (Brinkmann & Kvale, 2018). As researchers, it is critical that we ensure that the integrity and reputation of our own and of the participants are preserved in all stages of the study, especially in cases where AI stakeholders are in a delicate or vulnerable position.

Ethical guidelines stress the importance of anticipating and minimizing harm, maintaining confidentiality, and considering researcher safety. Challenges arise in qualitative research where anonymity is difficult to guarantee. Additionally, confidentiality can create dilemmas, especially when researchers witness unethical behavior (Bryman, 2016). Professional ethical codes and philosophical theories such as Kantian duty, utilitarian consequences, and Aristotle's virtue ethics guide ethical decisions in interview inquiries but do not provide definitive answers. Key ethical considerations include obtaining informed consent by clearly explaining the purpose of the study, risks, and the rights of the participants. Furthermore, confidentiality is critical, as it ensures that private information is not disclosed without consent and that ethical demands are balanced with scientific transparency (Brinkmann & Kvale, 2018). Drafting an ethical protocol will, therefore, help us make informed reflexive choices during the study design and prepare us to address sensitive issues that may arise throughout the investigation.

Ethical considerations in interview studies involve both micro- and macro-level perspectives. Micro-ethics focuses on personal implications for interview subjects, while macro-ethics addresses the broader social impact of the knowledge produced. Positive interview experiences for participants can lead to problematic social consequences, such as in studies used to enhance consumer behavior (Brinkmann & Kvale, 2018).

From a micro-level aspect, to protect the integrity of the researchers and study participants, we first introduced the participants to the aim of the study. We explained its beneficial consequences, which consist of a deeper understanding of stakeholder attitudes in implementing ethical principles in AI-driven persuasive technologies and possibly new insights on how to overcome the challenges that come with it. Before starting with the interview questions, we informed the participants that anonymity would be provided for their

names and the organizations they represent to prevent any possible negative consequences. Furthermore, we obtained informed consent to record the process for transcription.

Finally, from a macro-level perspective, this study partly contemplates the potential long-term impacts of challenges in implementing ethical principles in persuasive AI. Therefore, it may influence public opinion about AI ethics in persuasive technology, and it is crucial to ensure that the report is transparent and unbiased to maintain public trust in both AI technology and the research process.

## 4. Theoretical Framework

This section outlines the theoretical framework guiding our research, which is crucial to increasing the credibility and, thus, the trustworthiness of our findings. By doing this, we adhere to the significance of triangulation in qualitative research, as highlighted by Pickard (2013) and Bryman (2016). The theories selected have a key role in investigating *what are the viewpoints of AI stakeholders about current practices for addressing the ethical implications of AI-driven persuasive technology (RQ2)* and *what should be the focus areas of stakeholders in future efforts to ensure ethical AI in persuasive technology (RQ3)*.

In the following segments, we present the main theories that support our findings and discuss how these can be applied to interpret aspects of RQ2, such as the viewpoints, attitudes, motivations, and practices of stakeholders. Furthermore, we employ system design and evaluation theories to understand better specific technologies and techniques presented by the study participants.

### 4.1. Human-Computer Interaction

Human-Computer Interaction (HCI) is a multidisciplinary field that holds a central place in computer science and systems design. HCI involves designing, implementing, and evaluating interactive systems that are optimized for the user's tasks and workflows (Dix et al., 2003, p. 4). Because we are investigating the influences of AI-driven systems on user behavior, this is a central subject for understanding how these technologies affect user interactions and decision-making processes. Furthermore, in examining ethical considerations of persuasive AI technology, HCI provides insights into designing systems that are both user-friendly and ethically aligned with human needs and values.

Breaking down the term “human-computer interaction,” the “human” aspect refers to the users of a system. This includes a wide variety of user types, ranging from individual persons using a single computer to groups of users interacting with complex systems. Likewise, the “computer” aspect encompasses a range of technologies, including mobile devices, embedded systems, and large-scale computer systems. Lastly, “interaction” refers to both direct and

indirect communication between a user and a computer in order to accomplish a goal (Dix et al., 2003, p. 4).

Dix et al. (2003) outline three criteria for evaluating the success of products in HCI, stating that they should be “useful,” “usable,” and “used.” “Useful” means that the product effectively achieves its intended purpose, and “usable” indicates that it should be easy and intuitive to operate. “Used” is a more recent factor, focusing on making products that are engaging and attractive, motivating people to use them. Together, these criteria ensure that technology not only fulfills practical requirements but also enhances the overall user experience (Dix et al., 2003, p. 4).

The field of HCI has undergone several changes since its beginning in the early 1980s. It is continuously evolving to account for technological advancement, societal needs, government funding priorities, and user frustrations (Lazar et al., 2017). In recent years, one significant factor driving change is the increasing integration of AI in the technology we interact with on a daily basis, thereby changing both the computer aspect and the interaction aspect of HCI (Li & Hilliges, 2021, p. 9). Likewise, HCI plays an important role in AI development, and according to Li and Hilliges (2021), HCI is crucial for AI research to ensure the full benefits of emerging AI technology (p. 5). Additionally, they state that HCI tools are required to achieve fairness, privacy, and efficiency in real-world AI applications (Li & Hilliges, 2021, p. 6).

The increasing integration of AI points to the significant ethical dimension of HCI, with the principles of fairness, accountability, transparency, and explainability (FATE) frequently appearing in interaction design literature. Therefore, within the context of HCI, researchers should focus on developing more accountable intelligent systems that should not only explain their algorithms but also be usable and useful to people (Sharp et al., 2019, p. 379).

## 4.2. Persuasive System Design

Fogg (2003) defines persuasive technology as “*interactive computing systems designed to change people’s attitudes and behaviors*” (p. 1). We interact with this type of technology all the time, both online and in the real world, and with technological advancements, it has become increasingly integrated into everyday life, making it less visible to users (Fogg, 2003, p. 3).

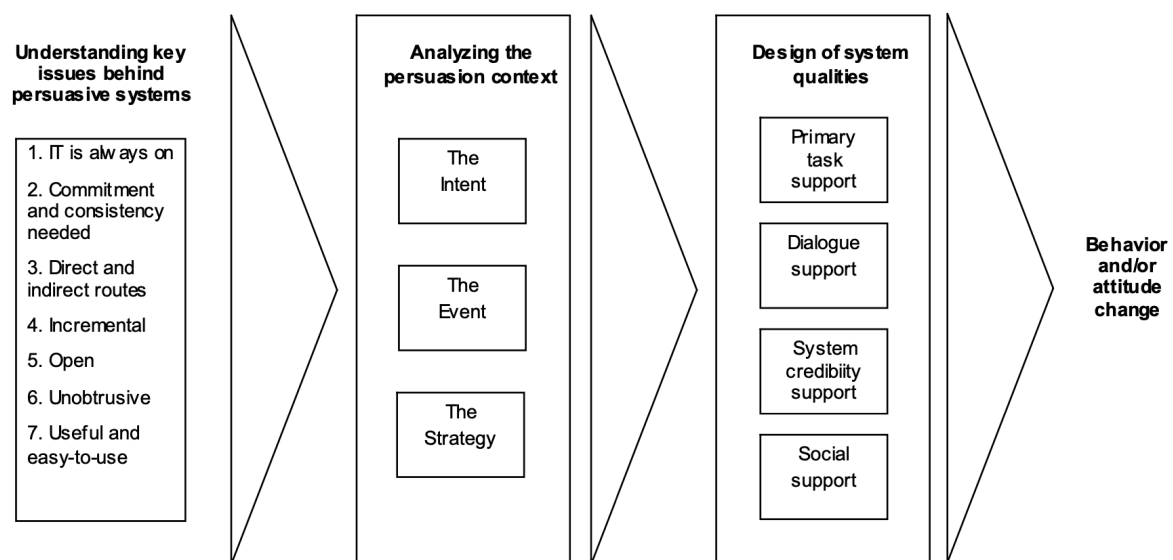
It is important to note that persuasion does not cover the concepts of coercion or deception. Coercion implies that the change occurs by force, while persuasion entails a voluntary change. Similarly, deception means to change someone's behavior or attitude by presenting false information (Fogg, 2003, p. 15). This distinction is specified in Oinas-Kukkonen and Harjumaa’s (2008) definition of persuasive systems as “*computerized software or information*

systems designed to reinforce, change or shape attitudes or behaviors or both without using coercion or deception” (Oinas-Kukkonen & Harjumaa, 2009, p. 486).

Oinas-Kukkonen and Harjumaa (2009) expand on the concept of persuasive technology by introducing the Persuasive Systems Design (PSD) model, which provides a framework for designing and evaluating persuasive systems. The PSD model highlights the importance of understanding the key issues and analyzing the context, including the intent, event, and strategy of the persuasive system, before designing its features (Oinas-Kukkonen & Harjumaa, 2009, p. 486). These are the three phases of PSD, as illustrated in **Figure 6**.

**Figure 6**

*The phases of Persuasive Systems Design*



*Note.* Retrieved from *Persuasive Systems Design: Key Issues, Process Model, and System Features*. by H. Oinas-Kukkonen, & M. Harjumaa, 2009, *Communications of the Association for Information Systems*, 24, p. 487.

The PSD model categorizes design principles into four main categories: primary task support, dialogue support, system credibility support, and social support, each providing guidelines to enhance the persuasiveness of the system within different areas (Oinas-Kukkonen & Harjumaa, 2009, p. 492). For example, primary task support includes principles such as reduction and personalization to help users achieve their goals more effectively. Tailoring, tunneling, reduction, and social comparison have been the most studied methods for persuasion (Törning & Oinas-Kukkonen, 2009, p. 1).

Due to its inherent aim of influencing people, persuasive technology raises significant ethical concerns (Fogg, 2003). Whether persuasive technology is unethical depends on how it is used, and according to Fogg (2003), ethical concerns can arise from several factors. One of

the main ethical concerns is the potential for abuse, as persuasive technologies can be designed to exploit users' vulnerabilities. For instance, systems that use emotional cues to influence behavior without the ability to reciprocate emotionally can create an imbalance that puts users at a disadvantage. Additionally, persuasive technologies can be persistently proactive, continuously influencing users without fatigue, which can lead to over-persuasion (Fogg, 2003).

Assessing the ethical implications of persuasive technology involves inquiry into three key areas: intentions, methods, and outcomes. Persuasion requires intentionality, meaning planned persuasive effects of the technology. Therefore, the intention should be ethical, promoting positive goals like health or education rather than exploiting users for reasons like increasing sales. Furthermore, methods must be transparent and empower users without aspects of coercion or deception. Lastly, both intended and unintended outcomes must not harm or exploit users, and creators must take responsibility for mitigating any negative effects (Fogg, 2003).

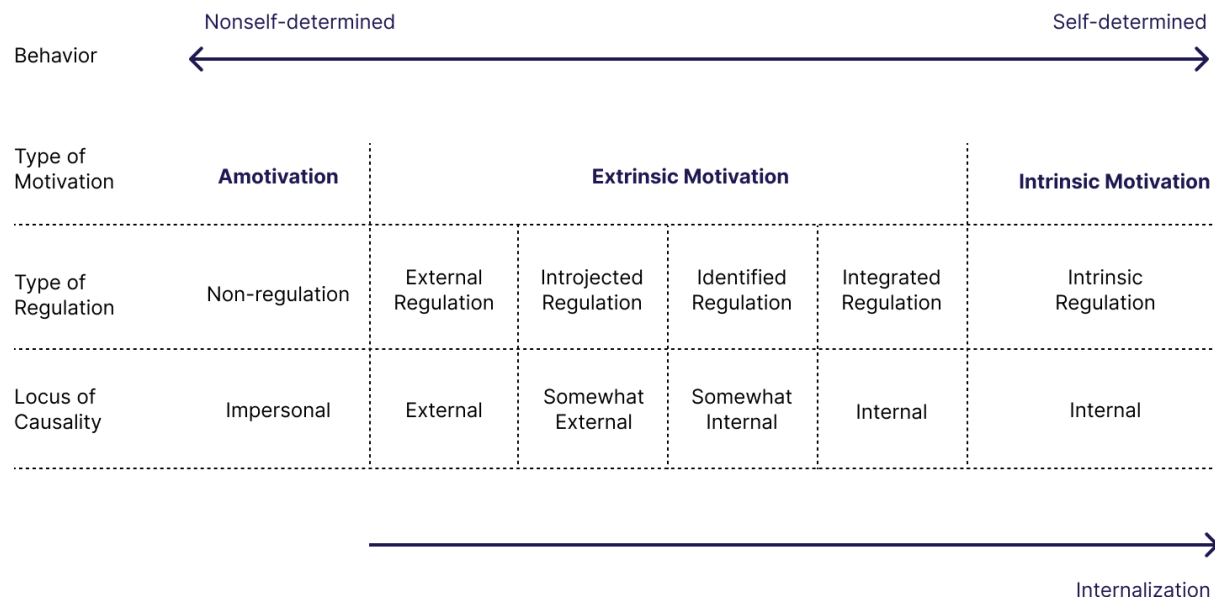
PSD is highly relevant to interpreting the results of our analysis regarding ethical AI principles in persuasive technology as it provides a framework for understanding how design choices can influence user behavior and decision-making. By applying this theory, one can analyze how ethical considerations are integrated into the development of AI technologies. This perspective helps to identify potential ethical issues and biases in AI systems, facilitating discussions about accountability, transparency, and the impact of persuasive technologies on user autonomy and well-being.

### 4.3. Self-Determination Theory

According to Self-Determination Theory (SDT), human well-being depends on satisfying basic psychological needs for autonomy, competence, and relatedness. The theory also emphasizes how environments that fail to meet these needs can lead to distress and ill-being, highlighting the importance of these elements in creating supportive settings (Deci & Ryan, 2000). As mentioned in the literature review, Burr and Floridi (2020) described a set of characteristics essential for conceptualizing human autonomy within the field of AI technology. This definition argues that for human autonomy to be supported in technological environments, several conditions must be met. First, individuals must experience a sense of willingness, volition, and endorsement in their interactions with technology. Secondly, there should be an absence of any pressure, compulsion, or the feeling of being controlled. Lastly, it is essential that there is freedom from deception or any deliberate misinformation (Burr & Floridi, 2020). These characteristics and conditions are crucial when it comes to observing how the ethical principle of human autonomy is influenced by AI systems that threaten basic psychological needs, endangering the psychological well-being of the users.

**Figure 7**

*The self-determination continuum and the types of motivation*



*Note.* The self-determination continuum shows the motivational, self-regulatory, and perceived locus of causality based on the degree of self-determination and the process of internalization. Adapted from *The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior*, by E. L. Deci & R. M. Ryan, 2000, *Psychological Inquiry*, 11(4), p. 237.

Another angle to the theory is the aspect of motivation, the experience of being “moved to do something” (Ryan & Deci, 2000, p.1). As **Figure 7** shows, SDT makes a fundamental distinction in types of motivation based on the underlying reasons or goals that inspire action. Intrinsic motivation involves engaging in an activity because it is inherently interesting or enjoyable. In contrast, extrinsic motivation involves performing an activity to achieve a separate outcome (Ryan & Deci, 2000). SDT can be a crucial tool for researchers when it comes to the interpretation of the qualitative data collected. In the context of this study, we employ SDT to examine the attitudes of various AI stakeholders and their motivations regarding the implementation of ethical AI principles.

Intrinsic motivation is described as engaging in an activity for its inherent satisfaction, not for any separable outcome. Humans, from birth, exhibit a similar innate motivation that is essential for cognitive, social, and physical development, as it drives them to learn and explore based on inherent interests. Intrinsic motivation is crucial throughout life, influencing performance, persistence, and well-being. This understanding has practical implications for task design, aiming to enhance motivation by making tasks more intrinsically interesting. The research into intrinsic motivation, largely informed by Self-Determination Theory (SDT) and Cognitive Evaluation Theory (CET), indicates that human intrinsic motivational tendencies are expressed under certain conditions, suggesting that social and environmental factors can



either support or undermine intrinsic motivation by affecting feelings of competence and autonomy (Ryan & Deci, 2000). Intrinsic motivation, therefore, plays a key role in influencing the way users engage with AI systems but can also contribute to the deliberate actions of stakeholders, especially AI developers, when applying practices concerning ethics in the development of AI-driven systems.

Intrinsic motivation is vital, however, it is often limited by social roles and demands as individuals grow older. On the other hand, extrinsic motivation, which is driven by external rewards, contrasts with intrinsic motivation, which is driven by the enjoyment of the activity itself. SDT posits that extrinsic motivation can vary in autonomy and emphasizes the internalization of values, where behaviors move from external regulation to self-determined actions. This internalization improves engagement, performance, and well-being. To foster internalization, environments need to support competence, relatedness, and autonomy (Ryan & Deci, 2000). Initial experiments by Deci demonstrated that monetary rewards diminish intrinsic motivation, reducing post-reward behavior below baseline levels. This shift occurs because extrinsic rewards change the perceived locus of causality from internal to external, making individuals feel controlled rather than self-determined (Deci & Ryan, 2000). Research also supports that autonomy-supportive environments enhance internalization and integration of values, leading to more self-determined behaviors (Ryan & Deci, 2000). The extrinsic motivation aspect of the stakeholder attitudes can be evaluated by assessing the degree to which stakeholders internalize ethical AI principles and transition from external regulation to self-determined, autonomous actions within autonomy-supportive environments.

Extrinsic motivation can be broken down into four types: external regulation, introjection, identification, and integration. External regulation involves performing actions to obtain rewards or avoid punishments, leading to compliance or reactance. Introjection is driven by ego involvement, where individuals act to gain approval from themselves or others. Identification occurs when individuals consciously value an activity and endorse its goals, leading to self-endorsement. Finally, integration represents the highest level of extrinsic motivation, where the activity is fully assimilated with one's values and goals, resulting in a harmonious and congruent synthesis of goals.

Finally, autonomous and controlled activities are both motivated behaviors involving regulatory processes, whereas amotivation represents a lack of intention and motivation, arising when individuals feel neither efficacious nor in control regarding a desired outcome, contrasting with both intrinsic and extrinsic motivation.

SDT, therefore, distinguishes between behaviors driven by autonomy and those driven by pressure, emphasizing that intrinsic motivation, arising from interest and satisfying psychological needs, represents self-determined behavior, while extrinsic motivation can become more self-determined through internalization and integration (Ryan & Deci, 2000). Unlike drive theories that focus on quiescence and need satisfaction, SDT views individuals

as inherently active and growth-oriented, engaging in activities out of interest or importance rather than solely to satisfy needs.

## 4.4. Ethical Theories

Ethical issues in AI are complex and debated, requiring an understanding of ethics that encompasses moral intuition, explicit morality, ethical theory, and metaethics (Stahl, 2021). Employing theoretical approaches sets the stage for examining current ethical challenges in AI through an empirically based approach.

The field of digital ethics, shaped by the evolution and societal integration of digital technologies, addresses both longstanding ethical concerns like privacy and data security in new digital contexts and entirely new issues such as automated decision-making and AI risks, evolving from terms like "computer ethics" to more encompassing terms like "digital ethics" better to capture the relevance and application of modern technologies (Müller, 2022).

Unlike traditional ethics, which builds on longstanding philosophical debates, digital ethics focuses on real-world implications and the ethical dilemmas posed by modern technologies. This shift aims to influence design processes and address novel ethical uncertainties, such as the responsibilities of artificial agents, thus broadening the scope of philosophical inquiry. By integrating insights from digital ethics back into traditional philosophical discussions, the field not only maintains its technical rigor but also enriches classical debates, thereby contributing to Immanuel Kant's fundamental philosophical questions about knowledge, duty, hope, and human nature (Müller, 2022).

A paper by Hagendorff (2020) that discusses the ethics of AI ethics suggests that while normative guidelines should include technical instructions, they are insufficient on their own. To enhance the application and fulfillment of AI ethics, one must consider broader ethical theories, particularly virtue ethics and deontology. Current approaches to philosophical ethics, like consequentialism and deontology, focus on the rules and outcomes of actions, whereas virtue ethics, rooted in ancient Greek philosophy, focuses on developing a virtuous character to guide individuals in living the "good life," emphasizing human flourishing, social interconnectedness, and practical reasoning, and has been modernized to apply to today's technology-driven society (Stahl, 2021).

### **Virtue ethics approach**

Virtue ethics, inspired by Aristotle and later developed by philosophers such as Philippa Foot and Rosalind Hursthouse, emphasizes character dispositions, moral intuitions, and virtues (Hagendorff, 2020). The term "virtue" in contemporary philosophical ethics, derived from the Latin "virtus" and the Greek "arête," meaning excellence, refers to stable traits that enable individuals to excel in fulfilling their distinctive functions, such as honesty, courage, and patience, which are cultivated through habitual practice and study. Unlike popular or historical connotations associated with Victorian morals or mere advantages, virtue in this

context signifies moral excellence essential for a good life. This idea, deeply rooted in ancient ethical theories from Aristotle and other traditions, suggests that virtues appropriately align an individual's feelings, desires, and actions, contributing to personal and communal flourishing (Vallor, 2016).

In the 21st century, reconnecting modern technological living with these timeless virtues is crucial for achieving true happiness and moral well-being. The concept of "technomoral virtues" is an essential approach to ethical decision-making in the realm of technology. These virtues include honesty, justice, courage, empathy, care, civility, and generosity (Shannon Vallor, 2016, as cited in Hagendorff, 2020). Virtue ethics focuses on the individual level, aiming to cultivate a moral character within technology developers rather than merely adhering to a set of rules (Hagendorff, 2020). These ethics are applicable when analyzing whether stakeholders value traits such as honesty, integrity, and responsibility in AI practitioners and whether they prioritize the cultivation of virtuous qualities.

### **Deontological approach**

Deontology, rooted in the works of philosophers like Immanuel Kant, focuses on strict adherence to rules, duties, and imperatives. This approach is characterized by fixed, universal principles that technology developers should follow. The traditional type of AI ethics, as outlined by Mittelstadt (2019), aligns with deontological concepts, emphasizing a set of prescribed ethical guidelines and maxims. For instance, AI ethics guidelines often propose principles such as fairness, transparency, and accountability that developers must adhere to (Hagendorff, 2020).

Deontology, particularly in the Kantian tradition, emphasizes the importance of inalienable individual rights and intrinsic human value, focusing on duties to each person rather than collective outcomes. It critiques the society-centered approach of utilitarianism that prioritizes measurable outcomes and social impact, advocating instead for a user-centered perspective that upholds fundamental rights such as the right to freedom from inhumane treatment, privacy, and equal treatment before the law. Furthermore, Kant's moral theory centers on the categorical imperative, a universal principle that evaluates actions based on their ability to be universally applicable without contradiction, such as false promises. It also recognizes that all persons are ends-in-themselves: *"Act in such a way that you treat humanity, whether in your own person or anyone else's, never merely as a means, but also always as an end"* (Kant, 1785, as cited in Mougan & Brand, 2023, p.4). Deontology argues for the moral necessity of respecting these rights irrespective of societal pressures, suggesting that this approach can improve fairness metrics, especially where utilitarian methods have failed (Mougan & Brand, 2023).

A deontological approach to interpretation allows us to interpret whether stakeholders believe persuasive AI development adheres to universal ethical principles, such as fairness, transparency, and respect for privacy. Moreover, we can analyze whether stakeholders emphasize the importance of following ethical guidelines or principles in AI development.

## **Utilitarian approach**

Utilitarianism, a prominent version of consequentialism first outlined by Jeremy Bentham and later refined by John Stuart Mill, posits that the morality of an action is determined by its outcomes, specifically its impact on overall happiness or welfare. It operates on three core principles: the moral value of an action is defined by its consequences, this value is assessed in terms of the welfare it produces, and the ultimate aim is to maximize welfare across all individuals equally. While the adaptability of utilitarianism allows it to reinterpret other ethical theories and remains influential in fields like AI, it faces significant challenges in measuring and comparing utility and criticisms from philosophers regarding its simplicity and potential ethical oversights (Lorente, 2022). Therefore, combining it with other ethical approaches allows us to address these challenges.

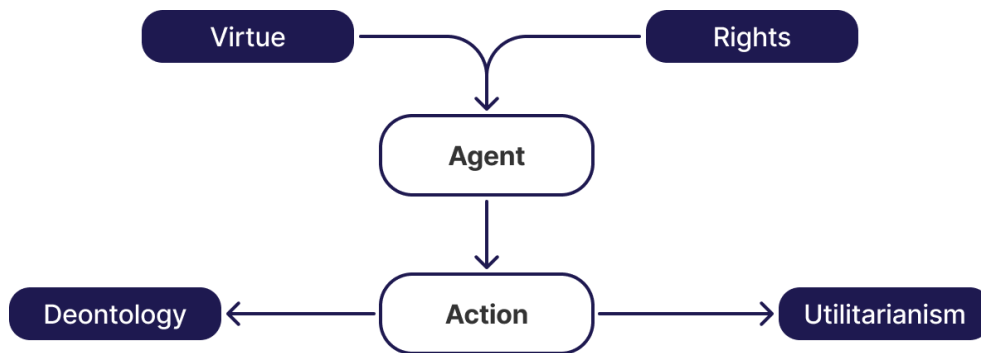
We employ the utilitarian approach when interpreting whether stakeholders are primarily focused on the positive or negative consequences of AI implementation. Furthermore, we look into aspects such as whether stakeholders are concerned with the impacts of persuasive AI on society, such as privacy issues and social inequality, and whether they support persuasive AI projects that aim to benefit the majority, even if some individuals may be adversely affected.

## **Rights-based approach**

The natural rights theory, highlighted by Locke's assertion of inherent rights to "life, liberty, and estate," influenced eighteenth-century revolutionary American and French political documents, especially Jefferson's Declaration of Independence (Wenar, 2023). Human rights, fundamental to all individuals by virtue of their birth, encompass moral claims, legal frameworks, and cultural practices that protect core human interests and dignity. Human rights are increasingly being integrated into AI design to ensure ethical accountability and address societal harms, emphasizing the need for alignment with established human rights principles to guide the development and deployment of AI technologies (Prabhakaran et al., 2022). In contrast to deontology and utilitarianism, rights impose duties on others to respect and protect these claims and are not necessarily tied to the outcomes of actions (Wenar, 2023). This approach serves as an additional angle when it comes to observing whether stakeholders are concerned with ensuring that persuasive AI systems respect and uphold human rights.

**Figure 8**

*Ethics theories considered*



*Note.* Deontology and utilitarianism assess the ethics of actions based on intentions or outcomes, whereas virtue ethics evaluates the character of the agent performing the actions while inherent rights are imposed. Adapted from *Moral exemplars for the virtuous machine: the clinician's role in ethical artificial intelligence for healthcare*, by S. Hindoch & C. Badea, 2022, *AI and Ethics (Online)*, 2 (1), p. 169.

Hindocha and Badea (2022) describe a suitable framework for ethical AI and healthcare, as it aligns with the bottom-up learning approach of machine learning and mirrors the ethical training of medical students, though they acknowledge that deontological and consequentialist frameworks also have roles in maintaining the importance of the clinician. Virtuous machines must learn from human moral exemplars, as machines cannot independently develop the necessary ethical decision-making skills, particularly in the ever-evolving context of healthcare, where clinicians will always be needed to guide and update machine learning models with new data and ethical insights (Hindocha & Badea, 2022). Considering the importance of human agency in using AI technologies, we adapted this framework to the context of our study, where we interpret the role of the agents (responsible stakeholders and end-users) and the ethics that drive their attitudes and behavior in shaping ethical AI.

## 4.5. Summary

To summarize, the selected theories are used to understand the design, implementation, and evaluation of interactive systems, how they align with user needs and ethical principles in HCI, and to understand and influence user behavior through Persuasive Technology. Deontology emphasizes the duty to create interfaces that respect user rights and privacy, ensuring ethical standards are met regardless of outcomes. Utilitarianism focuses on maximizing overall user well-being, promoting designs that enhance user satisfaction and minimize harm. Virtue ethics encourages designers to adhere to virtues like honesty, empathy, and responsibility, fostering trustworthy and user-centric interfaces. Meanwhile, rights-based

ethics underscores the protection of individual user rights, such as the right to be informed and to opt out, ensuring that technology serves to empower and not exploit users.

Finally, we analyze the motivations and ethical attitudes of AI stakeholders through Self-Determination Theory (RQ2) and evaluate how AI development can adhere to ethical standards (RQ3) through various ethical frameworks, including virtue ethics, deontology, utilitarianism, and rights-based approaches.

## 5. Thematic Analysis

In this section, we will conduct a comparative thematic analysis of interviews with stakeholders involved in AI technology. The analysis is structured around emerging themes from the interviews, as outlined in the methodology section, which closely align with the themes in the interview guide. These themes include the five ethical principles: (1) beneficence and non-maleficence, (2) fairness and justice, (3) human autonomy and agency, (4) transparency and explainability, and (5) accountability and oversight. It also considers perspectives on different AI stakeholders, including companies, developers, governments, and researchers, as well as the user. This will be presented as a comparison of the stakeholders in relation to their challenges, attitudes, motivations, responsibilities, and practices. Additionally, we will start by presenting the emerging views on persuasive technology in relation to AI. This analysis approach allows us to systematically explore and compare how different individuals and groups address the various themes.

When citing participants in this section, we maintained word repetitions and lines of thought to preserve their manner of speech and more accurately reflect their sentiments.

### Background

Before delving into the analysis and presentation of the different themes, we provide a brief introduction to the participants from the various stakeholder groups. As previously mentioned, our population includes two company representatives, one developer, two government representatives, and four researchers.

The first representative of the company stakeholder group, Company 1 (C1), has a background in data science but is currently working as an AI governance, risk, and compliance analyst at a leading Danish AI company that delivers advanced AI solutions to businesses. The second representative, Company 2 (C2), works as a digital strategy director at an accelerator for digital solutions. The developer stakeholder group is represented by Developer 1 (D1), an entrepreneur and mobile app developer of an AI nutrition app recently launched in the Apple App Store.

The government stakeholder group consists of participant Government 1 (G1), who works as a senior tech advisor for AI, emerging technology, and online platforms for a mission

representing Denmark in the United States as part of the Ministry of Foreign Affairs of Denmark. Participant Government 2 (G2) also works for the Ministry of Foreign Affairs but in Denmark as the head of the tech section within the mission. In their line of work, they engage directly and indirectly with big tech companies such as OpenAI, Microsoft, Meta, and Apple. Therefore, they were able to provide experiential insights about the attitudes of these.

The last stakeholder group, researchers, is the most well-represented in our population, with four representatives working at three different universities in Denmark. Researcher 1 (R1) has a background in computer science, robotics, and AI, has previously worked with multi-robot systems, and is currently working as a professor in a drone center. Researcher 2 (R2) is an associate professor in information science and is doing research in collaboration with Oxford University and Mannheim University on the persuasiveness of AI-generated news. Researcher 3 (R3) is an associate professor in computer science and information science and serves as Chief Scientific Officer of a major Danish research center for fundamental and interdisciplinary AI research. Lastly, Researcher 4 (R4) is a professor of data and AI ethics, a member of a center for AI ethics, and a former member of a council on ethics.

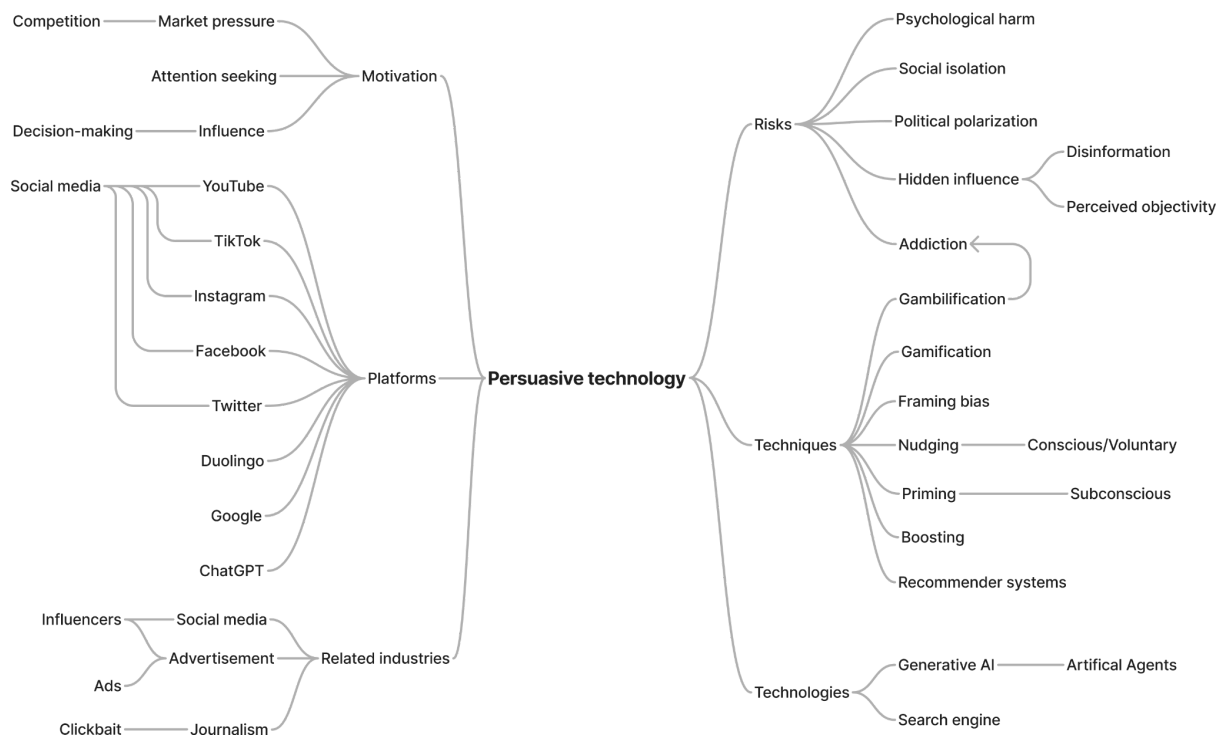
## 5.1. Persuasive AI Technology

We initiated the interviews by introducing the participants to the topic of our study and giving a short introduction to persuasive technology and what that entails in the context of AI. Following this, we asked whether the participant had experience with persuasive AI technology, revealing various degrees of experience. The majority of participants had worked specifically with AI technologies associated with some persuasive characteristics, such as recommender systems (R2, R3, C2, D1, G1) and generative AI (R4, G1). Others have mainly worked with other types of AI technology, such as robotics (R1) and within the ethical AI compliance fields (C1, G2). However, they all had personal experience as users of persuasive AI technology and were able to relate their professional knowledge and experience to provide perspectives on ethical principles in persuasive AI.

We visualized the main themes and subthemes identified in this initial stage of the interviews in a thematic map shown in **Figure 9** below. In the following sections, we describe the three main themes and the prevalence of their subthemes: (1) defining persuasion, (2) techniques and technologies, and (2) the risk associated with these.

**Figure 9**

*Thematic map representing main themes and subthemes associated with persuasive technology in AI*



### 5.1.1. Defining persuasion

When asked about their experience with persuasive technology, R4 emphasized that I know that *“it's really, really important here to distinguish between different ways of manipulating people and different ways of persuading people”* (R4, Appendix K) and highlighted the omnipresence of persuasion even outside of technology by stating that:

*“Persuasion happens in marketing and PR context, and it happens between people, it happens between teachers and students. It happens everywhere all the time. What distinguishes persuasion is that it's typically, you know, a cognitive and rational process is about. Persuading people to hold a certain belief. And then there's also, it's an an open way of influencing people”* (R4, Appendix K)

R4 defined persuasion as *“[...] the way we do it, go along with it all day long, it is autonomy preserving and autonomy protecting”* (R4, Appendix K) while describing manipulation as being *“about influencing autonomy in ethically problematic ways. It sort of bypasses autonomy. If you give people information, you give them arguments, you strengthen autonomy”* (R4, Appendix K). Furthermore, R4 explained that:



*“[...] manipulation is not an open act. It's not something that happens openly in language. It's not about giving a reason, providing an argument, it is hidden, and manipulation covers everything from outright lying, hiding facts, subliminal influence, and even some ways of nudging people.” (R4, Appendix K)*

On the other hand, R1 highlighted and elaborated how the market competition plays a role in pressuring stakeholders to optimize their solutions, and it is the reality of how the market works, suggesting that it is not inherently malevolent:

*“I think we are subjected to persuasive AI whether all of that or whether that's just the [...] optimization of [...] websites and the interaction paradigms and so forth that we are exposed to and that's basically optimized by [...] the free market” (R1, Appendix H)*

*“The algorithms [...] sort of work jointly, no whatever any mean intent or anything like that. There's no evil supermind behind it. It's just the fact that all these algorithms are optimizing for getting our attention, and it's also optimizing the way we work to create content that gets the attention” (R1, Appendix H)*

The aspect of market pressure as a drive for developing such technologies is also supported by G1, according to whom *“[...] the competition on like the economic competition and gaining market access and selling most and becoming the main source of some specific service, can overshadow their wants to make it ethical”*(G1, Appendix F).

When it comes to the most prevalent industries where persuasive technology is being used, many of the stakeholders mentioned social media platforms and their intent, for instance:

*“The Facebook, the Instagram, [...] the TikTok, the Google, the Reddit, let's throw Twitter in as well, mimics what they think you want or what you really do not want. So they try to. They only take what is already there, and then they multiply it”*(G1, Appendix F)

Additionally, R1 mentioned that journalism, in addition to social media, is also in competition for consumers attention:

*“[...] the fact is that you have many, many algorithms running in many different places, social media, newspapers that may be testing figure out like what headlines work and so forth that are optimizing what materials are exposed those but also change in the way that journalists they produce news and so forth like I mean, if you go to any website you can get free news like it's clickbait, clickbait, clickbait, clickbait.” (R1, Appendix H)*

According to these statements, industries that use persuasive technology are competing for the attention of consumers by optimizing algorithms and persuasive strategies.

### 5.1.2. Techniques and technologies

When talking about persuasive technology, a main theme of persuasive strategies emerged. As summarized in **Table 5** below, the most prevalent subthemes associated with this have been recommender systems (C2, D1, R1, R3, G1), nudging (D1, R2, R4), boosting (R2), gamification/gamblification (C1, R1), emotional framing/framing bias (R2, R4), and priming (R4).

**Table 5**

*Summary of subthemes identified within persuasive AI strategies*

<b>Recommender system</b>	<p><i>“we reach that point of time where we will have an AI <b>recommending</b> something” (C2, Appendix D)</i></p> <p><i>“[...] we ask them if they want to lose weight, gain weight or manage the current weight. And based on that we give them a number of calories they should take to hit this goal” (D1, Appendix E)</i></p> <p><i>“A <b>recommender system</b> can be defined as persuasive technology [...] We're also going to look at job recommendations, so <b>recommending</b> jobs to candidates.” (R3, Appendix J)</i></p>
<b>Nudging</b>	<p><i>“[...] our goal is to basically <b>gamify</b> it and help the users. And <b>nudge</b> them to use the app” (D1, Appendix E)</i></p> <p><i>“Some of the <b>nudging techniques</b> may be considered manipulative because they sort of fuck with people's minds.” (R4, Appendix K)</i></p> <p><i>“with <b>nudging</b>, you're more in this kind of context that human autonomy is lost to a certain degree and and then, of course, it is debatable, right, whether even if it is for a good cause” (R2, Appendix I)</i></p>
<b>Boosting</b>	<p><i>“<b>boosting</b> that is on how to change shopping behavior and turn shopping behavior into more greener shopping” (R2, Appendix I)</i></p>
<b>Gamification Gamblification</b>	<p><i>“[...] like casinos of that whole like, you know, way of <b>gamifying</b> something or using colors and all those things that we know that people like like to see and we get kind of addicted to.” (C1, Appendix C)</i></p> <p><i>“[...] algorithms somehow could <b>gamify</b> our psyche to keep us scrolling” (R1, Appendix H)</i></p>

Emotional framing Framing bias	<p><i>“[...] another question is what <b>emotional framing</b> actually works best for people to lead to actual climate action [...] but then also looking at can AI and in in this case GPT 4 replicate the <b>emotional framing</b>” (R2, Appendix I)</i></p> <p><i>“[...] in one case they focused on the mortality and mortality rate and the other they focus on the survival rate. And it makes a huge difference and that and that influence on people's choices is hard to describe as a rational inference. Because it's <b>exactly the same information</b>. So somehow something sort of short circuits it in the minds of people [...]” (R4, Appendix K)</i></p>
Priming	<p><i>“[...] this was in the early days of persuasive design [...] <b>priming</b> is sort of using colors and symbols, and so they don't have any sort of, in some cases [...] they are not part of a language in any ordinary sense of being part of the language.” (R4, Appendix K)</i></p>

C2, D1, and R3 mentioned recommendations as part of the software they have been working on. Examples of recommender systems have also been mentioned several times in the context of explaining concepts like beneficence, autonomy, accountability, and transparency when discussing ethical principles. For instance, R1 used a recommendation case to discuss accountability:

*"You have a human being that gets a recommendation from the system. Now it might be that the human being is [...] not sure that is right, you know? And so for the custody case [...] my back is covered if I just go with the recommendation.” (R1, Appendix H)*

According to Persuasive System Design (PSD) theory, recommender systems are user-friendly interfaces that effortlessly provide valuable suggestions and are always available. They continuously analyze user behavior to provide personalized recommendations (IT is always on). When these algorithms are transparent, they allow users to understand and trust the recommendations. Furthermore, they have the potential to influence user preferences and choices through personalized recommendations.

When it comes to nudging, D1 described how they use gamification to nudge the user to use their app. While he expressed that their *“goal is to basically gamify it and help the users and nudge them to use the app so they can track the calories.”*(D1, Appendix E), D1 also emphasized that *“nudging [...] can be also detrimental if the system is nudging the users despite the best interest of the users”*(D1, Appendix E). This aspect has also been highlighted by R4, according to whom *“some of the nudging techniques may be considered manipulative”* (R4, Appendix K).

R2 has been working with both nudging and boosting techniques and has described them as being similar but distinguishable. With nudging, *“human autonomy is lost to a certain*

*degree*” (R2, Appendix I), and *“the system might do the changes in the background”* (R2, Appendix I). With boosting, on the other hand, one *“would basically give the users ideas or strategies on how to handle that problem themselves”* (R2, Appendix I) and *“present users with some additional information, and the users why they're reading that now and basically to reflect about their actions”* (R2, Appendix I). Therefore, R2 describes boosting as a more transparent and less autonomy-undermining persuasive technique.

Based on PSD, nudging can manifest in the form of prompts and reminders to guide user behavior without requiring active decision-making. It contributes to small, incremental changes to the environment to gradually influence behavior and they can be unobtrusive through subtle prompts that do not interfere with user autonomy. Its intent is to gently steer user decisions in the desired direction. This means that it can be used to subtly alter user behavior. Boosting, on the other hand, focuses on enhancing user skills and decision-making abilities and establishing credibility through transparent and reliable information.

Less prominent subthemes, such as gamification, have been named by two participants, C1 and R1, who compared scrolling on social media with casinos (gambification) as a way to increase user engagement. R2 also mentioned the role of emotional framing in journalism, where organizations try to persuade people to be more sustainable, and the possibility of AI replicating this technique in the future. R4 supported this aspect, explaining how important the way information is presented is and that sometimes people can be emotional rather than rational. Additionally, R4 indicated that sometimes persuasion is not linguistic but rather visual, using colors or symbols that have the potential to persuade user behavior.

If evaluated based on PSD, gamification and gambification entail reward systems and progress tracking to keep users engaged and consistent in their behavior. Their intent is to engage users through game-like elements to motivate certain behaviors. Emotional framing (and framing bias) directly influences user emotions and perceptions through a specific presentation of information. They are used with the intent of shifting attitudes and perceptions through emotional appeals. Meanwhile, priming indirectly sets the stage for desired behavior through subtle cues with the intent of preparing and influencing future behavior.

Additionally, G1 added a new angle to persuasive technology, predicting a transition from search engines like Google to generative AI integrated into artificial agents (AAs) such as Siri.

*“Like five years ago, we would talk about persuasive technology when Google would promote my link over your link, not because my link is better, but because it's better search engine optimized or because I paid you to put it at the top and like does Google, who's not even a search engine, but it's more like an indexation machine, became persuasive.”* (G1, Appendix F)

*“If you can, we can imagine technology as human-like, as persuasive, as argumentative as ChatGPT integrated in every new Apple product, it's gonna substitute search function.” (G1, Appendix F)*

According to G1, these technologies may seem objective to most of the population; however, it is unclear how the output data is generated and whether there are hidden strategies behind them:

*“[...] we can't ask our ChatGPT powered Siri next week... Why are you telling me to buy Snickers? Because you think that my voice sounds annoyed and frustrated. Siri, what's going to tell us because it would help. OK, but why does Siri know that that would help? Because of the power of marketing after the ‘you're not you when you're hungry’ Snickers commercials. [...] Or is it going to be that [...] OpenAI and Apple partnered with Nestlé as well? And we have, like, a trifecta of world vendors?” (G1, Appendix F)*

When looking at generative AI through the lens of PSD, it continuously generates personalized content, suggestions, and interactions based on real-time user data and behavior (IT is always on). The intent is to produce content that aligns with user goals by enhancing the persuasive impact, therefore it can influence user behavior and attitudes by creating highly personalized and engaging content.

### 5.1.3. Risks

As most stakeholders stated, not knowing the hidden strategies behind the recommendation systems comes with certain risks. According to them, persuasive AI technology can negatively impact not only individuals but also society, revealing two main subthemes associated with risks summarized in **Table 6**. Each stakeholder type (C1, D1, R1, G1) acknowledged psychological consequences, such as addiction, loneliness, lack of self-esteem, and developmental issues in children. On the other hand, participants (D1, G1, G2) mentioned geopolitical issues that might arise from the hidden influences of persuasive technology, such as political polarization.

**Table 6**

*Summary of negative impacts of persuasive AI technology on individuals and society*

<b>Impact on individual</b>	<p><i>“if an AI is going to suggest, for instance, a diet plan [...] That could be seen as being well; you're causing <b>inside harms</b> because you know people they want to [...] themselves as being lean or not overweight or something like that” (R1, Appendix H)</i></p> <p><i>“YouTube shorts or TikTok or whatever [...] it's <b>highly addictive</b>” (R1, Appendix H)</i></p> <p><i>“all those things that we know that people like like to see and we get kind of</i></p>
-----------------------------	--

	<p><i>addicted to</i>” (C1, Appendix C)</p> <p><i>“some kids feel like they are <b>addicted</b>, some kids felt like OpenAI’s My AI chatbot had <b>persuaded</b> them to something” (G1, Appendix F)</i></p> <p><i>“[...] we just let our kids watch YouTube Kids and so forth, which is [...] just <b>garbage, empty calories for the brain</b> [...] companies now that are using AI to generate content for kids to put on on YouTube Kids. And it’s [...] the most <b>brain-dead stuff</b> [...] that doesn’t make sense. [...]” (R1, Appendix H)</i></p> <p><i>“[...] this thing can also in some ways support <b>loneliness</b> more” (D1, Appendix E)</i></p>
Impact on society	<p><i>“It could <b>polarize</b> certain types of people or certain groups of people to be more aggressive and <b>give certain opinions</b> that they wouldn’t do themselves in the public.”(D1, Appendix E)</i></p> <p><i>“[...] increasingly geopolitical role that Silicon Valley’s private tech industry tech sector will have on both <b>national politics</b> and <b>geopolitics</b> as a whole and <b>economies</b> as well.” (G1, Appendix F)</i></p> <p><i>“[...] the private organizations and the political politicians and the legislative organizations [...] fear that it’s going to <b>interfere in in political elections.</b>” (G2, Appendix G)</i></p>

On the other hand, G1 highlighted challenges when it comes to impact on individuals, expressing that it is difficult to determine psychological effects when it comes to technology:

*“It’s a problem we always hit in modern psychology that **it’s hard to isolate psychological effects**. If you do it on a large enough scale, like thousands, 10s of thousands of people, you would start to see some credible scientific effect.” (G1, Appendix F)*

*“[...] we can quantify how many people felt something, but **we can’t quantify if they felt something right**. Like, are you addicted? To your phone. I feel addicted. OK, how are we going to? How are we going to measure it?” (G1, Appendix F)*

Furthermore, G1 described hypothetical scenarios that present the complexity of the risks that artificial agents impose and their potential impact on individuals and society if they are being manipulated by unknown actors:

*“What if Al Qaida puts out a chatbot that is super nice, that is designed to understand the intricacies of being mixed race in a Scandinavian country? That understands the the quiet, backhanded racism that permeates a Scandinavian country. And it’s just so supportive. It is so nice. It is so great, and at some point, it starts arguing: Well,*

*maybe you should just go out and do shit. Like the quiet, tacit manipulation that can happen from something that feels like a friend is wild.” (G1, Appendix F)*

In summary, the participants expressed concerns about the hidden strategies of persuasive technologies and their impacts, such as addiction, loneliness, self-esteem issues, and political polarization. The potential for generative AI to transition from search engines to persuasive artificial agents raised further ethical questions about transparency and manipulation.

## 5.2. Ethical Principles

The main part of the interview was concerned with the five ethical principles for AI and the views and experiences of participants. In this section, we analyze each of the five principles in detail. We present the main themes that emerged during the interviews with the different participants, highlighting their practices, challenges, and insights in relation to implementing these ethical principles in AI development and deployment. Through this analysis, we aim to provide a comprehensive understanding of how these principles are perceived and applied in practice by the different participants, as well as the complexities and considerations involved in upholding ethical standards in AI.

As with the previous section about persuasive AI technology, the analysis of each principle is summarized in a thematic map visualizing the main themes and subthemes discussed. These also showcase some additional connections and specific examples not explicitly discussed in the analysis.

### 5.2.1. Beneficence and non-maleficence

When initially asked about beneficence and non-maleficence, something that was largely discussed among the participants was the need to challenge defining various aspects of beneficence and non-maleficence. When asked about his view on the ethical principle and what practices he is familiar with for promoting the well-being of users and ensuring AI systems do not cause harm, R1 answered:

*“Yes, it seems very nice. You know, like everybody could agree. Yeah, it should not do any harm. [...] But then when you dig down and when you try to define the words a bit more, you'll find out that sometimes there's trade-offs between these.” (R1, Appendix H)*

C1 also mentioned a certain consensus around beneficence and non-maleficence and how it is a fundamental rights question, with most people, especially in the EU, having a similar view on what that entails. However, several of the researchers expressed more uncertainty and questioned how exactly to define and measure these aspects. R4 responded: “[...] *the more philosophical question is what is beneficence?*” (R4, Appendix K). He further pointed out that in some cases, beneficence or the direct benefits of using AI are fairly obvious, such as in the healthcare context. However, for other technologies, such as recommender systems for

daily activities, it is more difficult to pinpoint the real benefits of using them. R3 shared this view in regard to the impact of recommender systems, talking about the potential effect on users. An example he gave in this context is when a user gets recommended something like a job, they don't always know what they are missing out on: *"[...] are you being harmed by it? Maybe not so much. But you're not benefiting from it either, right. So I feel that some aspects of this are harder to measure or harder to quantify than others."* (R3, Appendix J).

R4 elaborated on the issue of identifying benefits: *"In philosophy, we would use the notion axiology. What is really of value? So you have to come up with a credible axiology. And that's I think that's definitely a challenge"* (R4, Appendix K). In line with this idea of assigning value, D1 underscored the necessity of offering value to customers for a system to be successful: *"I think, in general, all apps are trying to somehow help the users because if they are installing it they need to have a certain purpose or it needs to have a certain value."* (D1, Appendix E)

When it comes to non-maleficence, there is a consensus that AI systems should not cause harm to their users. However, this aspect also causes difficulty when specifying what harm constitutes in different contexts, which R4 defines as "axiological problems." He elaborated on different degrees of harm, noting that some negative effects are so small that they are almost negligible. He used an example of using recommender systems for finding holiday destinations: if the recommendations are bad, the user might spend a bit more time finding the right option and may experience some frustration. On the other hand, he also pointed to more serious harms: *"[...] there is one general and non-negligible harm, and that is the carbon footprint, the climate effects of developing these models."* (R4, Appendix K). Like in the case of healthcare, he points to climate effect as an obvious and serious harm to certain AI technology.

At the same time, R2 stated that while AI technology should not cause too much harm to people, he also believes that there are some necessary risks associated with AI development:

*"I would say a little bit of harm is acceptable, right, to me, and I'll explain you why. Because I think that the problem is with digital technology, you can never get it right at the very first go, right?"* (R2, Appendix I)

He further supported this statement by explaining how design is an iterative process. Therefore, it is necessary to acknowledge that everything might not work as it should initially, but this is a necessary step to reiterate the product.

This idea of necessary risks relates to what R1 expressed about the existence of trade-offs between beneficence and non-maleficence. In his explanation of accepting certain harms, R2 also stated that a machine learning algorithm can never be 100% accurate. Consequently, in order to have a system that works well, you also need to accept some issues. This entails that in order to get the benefits a machine learning model can provide, like increased efficiency,



some level of imperfection or harm is unavoidable and seemingly accepted. He further touched upon this subject of trade-offs while talking about one of his own projects:

*“[...] we are to some degree, especially if you use nudging strategies, limiting the person's own free will to set the set sense because you're trying to like push them to that more healthy option. But in the end, it's good for them.” (R2, Appendix I)*

In this statement, he justifies certain harms, such as limiting the autonomy of the user, by the fact that the end goal, improving the health of the user, is beneficial, aligning with a utilitarian perspective. It highlights the necessity of balancing the positive outcomes with the potential negative impacts and underscores the issue of trade-offs in AI development.

Based on the insights from the interviews, the impact of AI technology can be categorized into two main areas: its impact on individuals and its impact on society. As for the impact on individuals, the aspect of physical and psychological health is highlighted, as well as increased efficiency. For example, as a developer of a nutrition app, D1 aims to promote physical health by making it easier to track the calories and nutritional values of food. He also talks about a different type of AI technology, a type of conversational artificial agent (AA), and how that can impact the psychological health of users:

*“If it can be useful for certain types of people that are like on the verge of suicide, and it could help them, basically release their emotions and talk more about themselves. So it could give them a second chance. But this thing can also, in some ways, support loneliness more” (D1, Appendix E)*

Here, he pointed to both positive and negative potential impacts of the same system. On the positive side, he described how it could help people in critical emotional states. At the same time, while he uses the words “support loneliness,” which might sound like a positive outcome, the rest of the conversation suggests that he is pointing to the negative effect of promoting loneliness by AAs replacing normal human contact. This is a good example of how the same AI system can be beneficial to some users while it could potentially harm others.

The category of impact on society extends beyond individual users to broader societal effects. One such impact is the environmental effects of AI, as classified by R4 as a non-negligible harm. He elaborated:

*“But you know, the 100 million prompts on ChatGPT, 100 million daily prompts on ChatGPT, you could see they use one GWh of energy every day, which corresponds to the daily energy consumption by 33,000 American households. [...] They have a significant footprint, carbon footprint.” (R4, Appendix K)*

Furthermore, R2 emphasized the importance of considering the broader implications of AI: “[...] it's important that to have that kind of perspective on machine learning and AI for social good” (R2, Appendix I).

Moving on to the practical aspects of implementing beneficence and non-maleficence, C1 highlighted an approach that relies on clear definitions and boundaries:

*“I think for a lot of AI systems, it starts with actually defining the intended use of the system. Because once you're you have defined what you're actually going to use it for, you also need to define what you definitely cannot use it for.”* (C1, Appendix C)

She further explained that by clearly defining the intended use and limitations of an AI system, it becomes easier to identify and mitigate potential risks associated with its implementation. This approach helps in understanding and managing the risks involved in deploying the AI system.

To ensure alignment of the technical aspects, C1 mentioned the definition of principles as a specific method:

*“And it's obviously also a lot of technical aspects and actually implementing it, making sure that you have some kind of principles that you're living up to that you have to define beforehand.”* (C1, Appendix C)

She elaborated on these principles, stating that they need to be tailored to each specific system by customizing existing guidelines. However, she expressed that this approach is challenging because AI systems can have multiple uses, as D1 exemplified previously, requiring multiple definitions and tailored principles in order to meet ethical standards.

Furthermore, ensuring that AI systems adhere to the principles of beneficence and non-maleficence requires ongoing effort even after the development stage, as highlighted by C1:

*“[...] while any system is in deployment, then you also need to, you know, continuously monitor it to actually make sure that you're not stepping over the line that you have defined before you actually started implementing this system.”* (C1, Appendix C).

Here, she referred to the boundaries defined before the implementation of the system, emphasizing the necessity of continuous oversight to maintain ethical standards. C1 further emphasized the importance of monitoring as a practice to reduce the challenges of ensuring this ethical principle: “[...] the challenges are mostly making sure that we are keeping to what we have promised so that we could monitor the models and what we're doing” (C1, Appendix C).

Other challenges that were pointed out in relation to ensuring beneficence and non-maleficence include uneducated decisions and a lack of consensus in the field. C1 stated that it is important to make sure that the people who are developing the models have made the right decisions: *“I think there's something dangerous about people not knowing what is actually like the theory behind everything, just implementing things”* (C1, Appendix C). Furthermore, G2 pointed to a very fragmented AI approach. This fragmentation applies to most ethical principles, with many different organizations and institutions providing guidelines, making it difficult to know whom to listen to. *“[...] it's so fragmented, there's no general rule book to look towards”* (G2, Appendix G).

C1 also pointed to different ethical views and values as a significant challenge based on her experience in the field:

*“[...] what we are experiencing or hearing sometimes from some of these groups is that when Meta, for example, is they have a very set way of wanting things to be so that it benefits them and not necessarily the society or anything like that.”* (C1, Appendix C)

Here, she explained how large corporations, often referred to as big tech, prioritize their own interests, which may not align with broader societal or ethical considerations. The statement points to a lack of deontological and utilitarian ethics among these companies, where the interests of a few are prioritized over the larger social benefits. It also contrasts what C1 previously stated about beneficence and non-maleficence being a fundamental rights question, which emphasizes the importance of a rights-based ethics approach to the principle.

When talking about the responsibility of different stakeholders, several participants mentioned the aspect of user responsibility. D1 explained how although he, as a developer, assumes the responsibility of providing an accurate AI product, the users also have a certain responsibility:

*“[...] but at the end of the day, we try to tell our users that the results are not accurate and it's in their judgment to decide if they want to use it or not.”* (D1, Appendix E)

In line with this, G1 expressed that companies or parties supplying the technology should not be held accountable for how their products are used or misused by users. He drew parallels with the case of Pirate Bay, noting that although the platform is often used for downloading pirated content, it was originally designed to share academic data and promote open-source knowledge, which is not illegal. This argument underscores the idea that AI, like any tool, should be assessed based on its intended use rather than its potential for misuse. It points to a discord between deontology and utilitarianism, where despite the beneficent intention of a tool, it ends up being used for malevolent reasons. According to this angle, the creator of the tool is absolved from responsibility due to their benevolent intention. He illustrated this perspective by explaining that if an AI system is used for racist actions, *“that's a problem with the racists. It's not a problem with the tools they use for racism”* (G1, Appendix F).

Several different views have been presented while discussing the topic of beneficence and non-maleficence. From the interviews, it is clear that there is a need to define various aspects of these principles, such as axiology and use. Participants expressed a significant degree of uncertainty regarding how to achieve and measure this. This is demonstrated by the answer given by C2 when asked how to determine if an AI system is beneficial for humans and does not cause harm: “[...] *we can't know that. We can try our best to make sure that it doesn't, but we can't be 100% sure*” (C2, Appendix C). Other challenges related to this principle are uneducated decisions among developers, lack of consensus in the field, and different ethical values among stakeholders. However, the interviews also uncovered some practices for ensuring beneficence and non-maleficence in AI systems, including striving for accuracy in the AI models and the use of guidelines, as mentioned by D1. He highlighted the HAX toolkit as “*a human-centered design guideline for a application on how they should create a safe and user friendly application which harness that which can harness the AI features. And try to not harm people*” (D1, Appendix E).

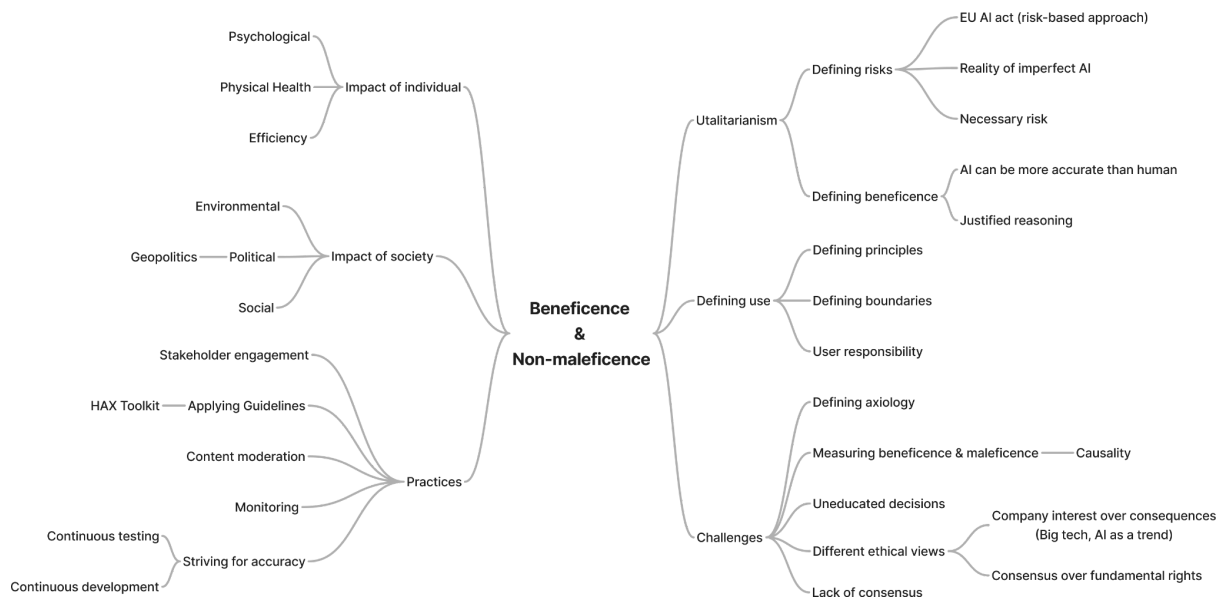
Lastly, a view shared by several researchers is that actors in the AI industry should be more critical about when to apply AI. They emphasize that before developing or employing AI systems, it is essential to question whether AI is actually useful and ensure there is a clear, significant benefit. As R2 said, “[...] *but not by all means do we have to implement AI everywhere. And so that's why I think it is necessary to have that critical dialogue of whether it actually has something good for us, right?*” (R2, Appendix I). This perspective aligns with the principles of human-computer interaction (HCI) and utilitarianism, which stress the importance of user-centered design and maximizing overall happiness and well-being.

In line with this, several participants pointed out that AI is often used simply for the sake of using AI, rather than because it offers a real advantage. This trend can lead to the implementation of AI systems that do not provide meaningful improvements or benefits. R4 expresses this as an important aspect of future, responsible AI development:

*“I think in the future, the developers of AI systems and users of AI system, they will have to face and answer this single question. Is this a case and situation where we should use AI or is it a case or situation where we shouldn't use AI?”* (R4, Appendix K).

**Figure 10**

*Thematic map representing main themes and subthemes associated with beneficence and non-maleficence*



### 5.2.2. Fairness and justice

Fairness and justice are crucial aspects to ensuring ethical and responsible AI, as highlighted by G2:

*“One of the things that I hear people say when they say that AI brings a lot of promise, but also a lot of peril is exactly the injustice and the biases and the discrimination and all these things that AI can bring with it.” (G2, Appendix G)*

However, much like beneficence and non-maleficence, the principle of fairness and justice raises significant challenges in terms of definition. This uncertainty was emphasized by R2, who stated, *“that’s difficult to say right because it’s also very difficult to I think to define fairness in the first place, because what is fairness, right?”* (R2, Appendix I). Furthermore, the subjective nature of fairness was underscored by R1:

*“[...] so when you say you have a bias right then and you always assume that there is a point where you have no bias, that that is possible. But that’s never going to be the case, because bias is subjective, like in the sense that what some people will perceive as a bias might originate from how they they want the world to be rather than what the world is like.” (R1, Appendix H)*

This statement reflects the inherent challenge of creating unbiased AI systems. R1 pointed to bias as an inevitable part of human nature and perception, and because AI systems are created by humans and trained on human-generated data, they reflect these biases. This notion is

supported by R3, who, in talking about his work in job recommendations, stated, “So mistakes can happen, right? Mistakes will also happen by recruiters. I mean recruiters have bias.” (R3, Appendix J). G1 further highlighted this as a problem, asking “[...] how do we teach a chat bot, how do we teach an AI to understand systemic racism? Which has an output of objective data, but based on biases.” (G1, Appendix F)

Highlighting the complexity of fairness in AI, R1 provided a recent example demonstrating the unintended consequences that can occur when attempting to mitigate certain biases, such as lack of inclusivity or racism:

*“I think a recent example that caught the eye of the press was that the Google Gemini model started to generate images of diverse Nazis, right? So they were asked about, like the prompt for generating soldiers from, German soldiers, from Second World War. And they came out black and Asian and all those other things that Nazis are not, right.”* (R1, Appendix H)

This suggests that while trying to mitigate on bias, other issues might arise, such as historically inaccurate and culturally insensitive outputs. In referring to the same example, G1 questioned how we can define who is a vulnerable or targeted group and who is not. He also emphasized the creation of new issues by stating “Now we just removed one level of racial bias. But it can be derogatory towards white people or Asians or Native Americans” (G1, Appendix F).

Delving more deeply into the concept of fairness within AI systems, R2 elaborated:

*“[...] that's maybe a bit of the problem in the machine learning or in the yeah or computer science perspective on things, right, that you somehow forget about that fairness is not only a mathematical concept, but fairness is also a sociological concept that is based on human values.”* (R2, Appendix I)

This statements establishes fairness as a sociological concept, and R2 further talked about how this requires interdisciplinary collaboration to ensure these human values are being adhered to.

Defining fairness is further complicated by its context-dependent nature. R3 illustrated this by explaining that fairness can vary significantly across different domains and stakeholders: “It is really domain specific, right? When something is fair or when something is unfair and it really depends on the stakeholder” (R3, Appendix J). He emphasized that defining fairness requires engaging with all relevant stakeholders to understand their perspectives and needs. R2 also shared this view, stating “and basically depending on what kind of stakeholder you are, you might have different interests, right?” (R2, Appendix I). For example, in his work on a job portal, R3 involved job seekers, companies, recruiters, and job centers to develop a shared understanding of fairness and to establish specific fairness metrics. Subsequently,

stakeholder engagement is highlighted as an important practice for ensuring fairness and justice in AI technology.

R3 further illustrated the context-dependent nature of fairness with an example underscoring the importance of proportional representation in specific contexts, aligning fairness with realistic and context-specific benchmarks:

*“So of course, if there are 20% men in the nursing profession and 80% women and you can't expect suddenly it to be 50/50, right? But it also shouldn't be 90/10. It should be that 20/80 proportion, right?”* (R3, Appendix J)

A critical aspect of fairness and justice is discrimination. In relation to this, R4 clarified that bias does not equal discrimination:

*“It's important to distinguish here between bias, and that's really, really important, to distinguish between bias and the ethical problem of discrimination. Because you're gonna have a lot of biased systems but it's not clear that I'm being discriminated.”* (R4, Appendix K).

He elaborated, explaining that not all biases can be classified as discrimination and that individuals do not have a right to be treated equally in all aspects. Differential treatment happens all the time, and sometimes, an algorithm is biased because it personalizes content based on user data. Therefore, he underlines that *“[...] in some context, maybe we prefer for the systems to be biased. We would prefer them to treat us, you know, in different ways.”* (R4, Appendix K). However, he provided another example where a man and a woman receiving different sentences for the exact same violation of the law. *“Then we would all say, ah, here we have a clear cut case of discrimination, because if you commit the same crime and all the circumstances are the same, you cannot pass a different sentence.”* (R4, Appendix K) This is discrimination because *“Usually you define discrimination as morally unjustified differential treatment.”* (R4, Appendix K).

This issue is also tied to the level of risk associated with the system. R3 noted, *“[...] this is a use of AI and this is marked as a high-risk area because you are dealing with people's lives”* (R3, Appendix J). As highlighted by several of the participants, the EU AI Act reflects this risk-based approach, emphasizing the importance of assessing and mitigating risks in AI systems, particularly in contexts with significant impacts on individuals' lives and rights.

Addressing bias in AI requires recognizing its different forms and origins. R4 noted, *“there are different kinds of bias you can bias in your data and you could also have bias in your algorithms. You can have bias in your tagging also”* (R4, Appendix K). Therefore, C1 stressed that it is crucial to understand and be aware of the various sources of bias to effectively manage them in AI systems. Similarly, R1 expressed that rather than trying to correct all biases, we should acknowledge their existence and approach the issue with transparency:

*“I’m more against trying to do that and instead just saying this is what the model was trained on. If you want to inject some preferences, you can do it like this and like that, but then let people use the model instead of having a committee trying to decide is it a fair and unbiased model because you can never do that in an objective way.” (R1, Appendix H)*

Based on the insights from the different interviews, there are several established practices for ensuring fairness and justice. One fundamental approach is stakeholder engagement, which we already established as critical for understanding the diverse perspectives and needs that define fairness within specific contexts. Another practice is meticulous preparation and monitoring of AI systems, as highlighted by C2: *“We do our best in the training data and the setup of the AI, and then we continuously follow and evaluate all outputs that come from it to be as close as 100% sure as we can be”* (C2, Appendix D). C2 further noted that fairness and justice heavily depend on the data used: *“So for each project that you’re doing, it’s very important to actually be conscious about the data and like the pitfalls of where bias could enter. And so being aware of that is very important”* (C2, Appendix D).

Furthermore, several of the participants noted that there are already established tools and metrics for measuring bias. C1 stated:

*“I mean there are a few tools like fairness and bias metrics that you can use to actually ensure that your data is not biased or your model is not biased. So I think that’s, I don’t know, maybe one of the goals that has a bit more of practices already established that you could use.”* (C1, Appendix C)

Here, she specified that this is a field that seems to be more advanced in practice compared to many of the other principles. Also, C2 and R3 mention these pre-made metrics and how they can provide a kind of score for bias. In working on ensuring fairness and justice, R3 states that *“We haven’t made our own because if there are like 50-60 metrics out there already, then we we didn’t think it made sense to reinvent the wheel, right?”* (R3, Appendix J). He further underscores how there already is sufficient practice for measuring fairness and justice by stating, *“[...] with so many metrics, I have to assume that a lot of situations have been mapped over the last five to 10 years”* (R3, Appendix J).

Although various practices are in place for ensuring fairness and justice, significant challenges are still associated with the principle. One difficulty lies in dealing with "black box" models, as C1 highlighted:

*“Especially for AI models, it could be difficult if you have these black box models. So if you’re not always sure what’s going on in the model, or why it’s making the decisions that it’s making, then it’s also very hard to mitigate the bias that it has. So that’s that’s definitely. One of the main challenges, I think.”* (C1, Appendix C)



Another challenge stems from GDPR regulations that protect user privacy by requiring a justifiable reason for storing certain data. While this regulation is crucial for privacy and contributes to less bias in models by not providing sensitive information, this regulation paradoxically complicates efforts to test and monitor bias in AI models. R3 articulated this dilemma:

*“Does someone directly benefit from me knowing that they're a female or male or a non binary or something? Probably not because we don't want to use it in the algorithm to recommend because that would make it biased, but we do need that data to figure out whether we are biased or not right?”* (R3, Appendix J)

R3 highlights a form of backtracking as a practice for ensuring model fairness during use. However, if they are not allowed to store sensitive attributes such as gender and age, it complicates the processes of ensuring no bias in that regard and might result in less accurate assessments of the models. Furthermore, when making adjustments to mitigate bias and ensure fairness and justice, R3 also mentioned the importance of monitoring its effect on other aspects of the model. He stated:

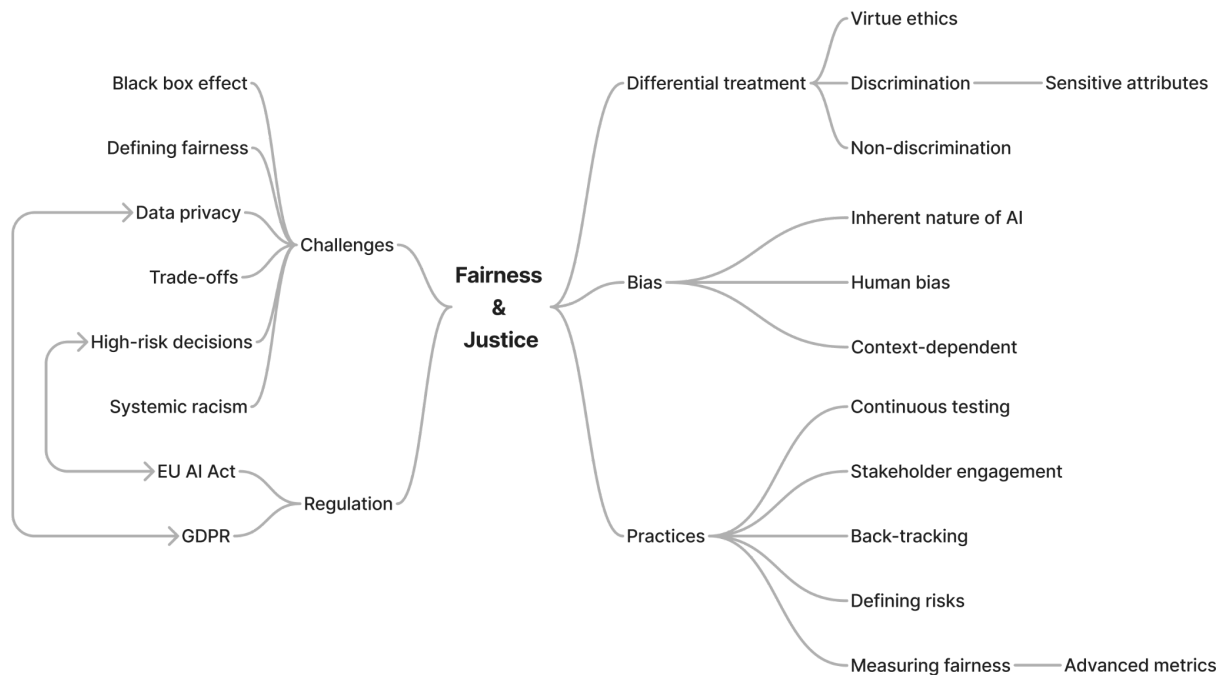
*“Like if we implement this new algorithm, does that make it better? Does accuracy go down with fairness, really shoot up, or can we find a balance right? Because it's typically a bound balance, you can't just say, well, it has to be only fair, right? [...] there's a trade off there”* (R3, Appendix J)

This underscores the importance of balancing fairness with other critical aspects, such as accuracy and efficiency in AI models. Ensuring fairness and justice is not only about removing bias but also about understanding the broader implications of these adjustments on the overall functionality and performance of AI systems.

In summary, fairness and justice are fundamental principles for ethical AI, yet defining these concepts is challenging due to their subjective nature. We have explored how bias is inherent in AI systems, reflecting human biases from training data, making it impossible to achieve total fairness. Furthermore, context-dependent fairness requires stakeholder engagement to establish relevant metrics, and there must be a clear distinction between discrimination and bias. Practices for ensuring fairness and justice include recognizing biases, using existing tools to measure them, and balancing fairness with other factors like accuracy. Challenges include managing "black box" models and navigating GDPR regulations, highlighting the complexity of maintaining fairness in AI.

**Figure 11**

*Thematic map representing main themes and subthemes associated with fairness and justice*



### 5.2.3. Human autonomy and agency

Human autonomy and agency in the context of AI is a complex and multifaceted concept. G1 emphasized the need to understand what human autonomy entails, stating, *“Well, that is a super interesting question that sort of prerequisites a discussion about what is human autonomy?”* (G1, Appendix F). One thing that might make this challenging is that the principle of human autonomy often intersects with other principles and issues. As R4 pointed out, many privacy and data control issues are essentially violations of autonomy rather than direct harms: *“So, but usually all these privacy and control of data issues are not considered harms. They are considered violations of autonomy”* (R4, Appendix K).

R1 illustrated how we are constantly exposed to influencing factors, especially through social media and their algorithms made to keep our attention:

*“It's just the fact that all these algorithms are optimizing for getting our attention, and it's also optimizing the way we work to create content that gets the attention. So in a way, we are subjected to these algorithms, not evil ones, just some that work because they're constantly in competition with each other and connected well, right. So yes, it does take human autonomy away very much.”* (R1, Appendix H)

As human autonomy can be difficult to define, a challenge that emerges is how to measure it. G1 highlighted the difficulty of scientifically assessing the influence of persuasive AI on human autonomy, suggesting that it would require controlled studies comparing the experiences of users with and without AI interaction:

*“If we argue that humans are autonomous and we do have free will. Then the only scientific way to understand how influential or how persuasive persuasive AI would be, would be to have a controlled group and ask the exact same people having the exact same experiences and the exact same situations complex questions and some of them use GenAI and some of them do not” (G1, Appendix F)*

G2 also supported this view, stating that *“It's very hard to know sometimes when you're actually getting influenced by something” (C1, Appendix C)*. This underscores identifying issues and measuring human autonomy as significant challenges in ensuring this principle.

Although measuring and ensuring autonomy can prove difficult, this is a very important as the influence of AI on human autonomy can have significant societal implications. G2 observed that many private organizations and legislative bodies were concerned about AI's potential, and reasoned, *“Probably because of the fear that it's going to interfere in in political elections” (G2, Appendix G)*. This concern points to the importance of ensuring autonomy to maintain democratic processes and individual freedoms.

The principle of human autonomy and agency is closely connected with persuasive technology and many of its aspects. One such aspect is the distinction between persuasion and manipulation, as previously discussed. R4 argued that while persuasion could enhance autonomy by providing information and arguments, manipulation bypassed rational processes and undermined autonomy. Subsequently, this differentiation and the ability to distinguish between what is persuasion and what is manipulation in AI technology are crucial for developing ethical AI systems that respect and preserve human autonomy and agency.

Furthermore, the practices and views on human autonomy and agency are deeply interconnected with aspects of transparency and user rights. Types of user rights that were frequently discussed by the participants in relation to human autonomy and agency were the right to be informed and the right to opt out of or contest AI decisions. In relation to this, R3 emphasized that users should have the ability to contest AI-generated recommendations and understand the data behind the decisions:

*“People should have a way of contesting these results in a sense, right? They should ask OK, well, why? Why am I being recommended this? There should be some some way of doing this because that's a helpful feature. People should also be allowed to figure out well, what data is being used to generate these recommendations in general [...]” (R3, Appendix J)*

This capability is underscored as essential for maintaining user autonomy, as it allows users to question and potentially reject AI-driven decisions.

Furthermore, R4 emphasized the importance of information transparency, referring to Article 14 of GDPR: *“And so in Article 14, you have the right to meaningful information about the*

*logics involved in automated decision making. And that's the right to transparency or the right to explainability.”* (R4, Appendix K). C1 also stressed the importance of informing users about the workings of AI systems to allow them to make informed choices: *“We also need to make sure that we're informing the people using an AI system of what is happening so that they also get the choice of whether they want to use it or not”* (C1, Appendix C). Therefore, informing users about AI operations ensures that they can exercise their autonomy by opting out of systems they are uncomfortable with.

A problem that was highlighted regarding the right to be informed is the lack of awareness among current users. R3 highlighted this in relation to data sharing on platforms like Facebook, noting:

*“People have shown that people don't necessarily know. First of all, they don't know what they're sharing. They don't know how much they're sharing over time. For instance, if you compare, I think it's 2005 and 2014. I have this diagram that shows you how much information you were sharing by default, that's visible by default, about your friends and your photos.[...] And that that has increased over the years. So first of all, people don't know this”* (R3, Appendix J).

This lack of awareness can lead to unintentional data sharing, undermining user autonomy and control over personal information.

The right to opt out of or contest AI decisions, as already mentioned by R3, is another critical aspect of human autonomy. C1 highlighted the necessity of allowing users to opt out of data sharing and AI interactions: *“Well, I think it's important that people are actually aware of what they're interacting with and that they give the opportunity to say no and opt out of whatever that AI is doing”* (C1, Appendix C). Also, this right is covered in the GDPR: *“And then we have an Article 22. If you are subjected to automated decision making, you have a right to to state your opinion on the automated decision making and to contest decisions, and so this is the contestability right.”* (R4, Appendix K).

However, C1 pointed out that often, users do not have a straightforward way to do this, for example, when it comes to cookie policies on websites:

*“There are a lot of times when that pops up on your phone and you don't actually get the very concrete choice of saying yes or no. Very often that yes, you can use my data or oh, you only have to go into this little box and check and a lot of other things in order to actually opt out of sharing your data”* (C1, Appendix C).

This is an example of bad practice for human autonomy and agency, where the complexity and inconvenience of opting out can lead to users unknowingly consenting to data sharing. Such practices undermine the user's ability to make informed decisions and maintain control over their personal information. Elaborating on this, C1 also stressed the need for users to have the option to reject data use outright:

*“It's very important that users actually get the possibility, and it's not necessarily when interacting with the AI system, but in all the cases where you're doing something on your phone and they're saying we're going to use your data, we would hopefully always get the option to say no, you're not allowed to use my data” (C1, Appendix C).*

This ensures that users can protect their personal information and maintain control over how their data is used in a clear and accessible way. Similarly, G2 underscored the necessity of having safe environments where users can choose whether to engage with AI. She highlighted the importance of this by stating, *“It's almost walking around blind if we don't do the clear separation”* (G2, Appendix G), emphasizing that in Denmark and Europe, it is a fundamental right to access trustworthy information and be in environments where they are aware of AI exposure.

Furthermore, R2 noted the importance of having accessible procedures for users to contest and report issues. He questioned the effectiveness of current reporting mechanisms on social media platforms, asking, *“Of course you can report that tweet or that post or whatever, but what happens after that, right? Who is looking at that? How serious is that report taken?”* (R2, Appendix I).

Practices for ensuring these user rights overlap with certain practices for the principle of transparency and explainability, which will be presented in the next section. One key practice mentioned was the use of labeling or watermarking to ensure transparency and help users make informed decisions. G2 gave an example of good practice, stating, *“Like if public services and other platforms that have large reach, always and actively tell their consumers and their customers that they're they might be exposed to something”* (G2, Appendix G).

Additionally, G2 discussed a compact made by Microsoft with several private companies to promote responsible watermarking and labeling:

*“I know Microsoft made a did a compact with a lot of private companies to, they made a declaration to work towards responsible watermarking and any kind of labeling. So they brought the companies together and then they that compact, that group of companies, have been going out and educating governments on how easy it is for them to help spot the AI generated material, but very difficult to agree on a specific labeling. How to warn people about it.”* (G2, Appendix G)

D1 provided an example of implementing this practice in user interfaces by displaying information: *“Or we have it in our terms and conditions and we have it on the user interface just like in ChatGPT, where under the input box you have the text saying that ChatGPT can be wrong”* (D1, Appendix E). This practice strengthens user autonomy by displaying information about AI's potential flaws or inaccuracies.

R1 also supported this notion, suggesting that technology could benefit from having a similar practice to nutritional labels on food, which help consumers make informed choices, stating that *“You should be able to see what the system is trying to do to you so you can decide whether to use it or not”* (R1, Appendix H). Another practice highlighted by D1 is the incorporation of user feedback:

*“So we we we are constantly learning from the users and figuring out how to create the interface and the user experience so the users can correct those results or at least notify them that it can make mistake and they can edit it”* (D1, Appendix E).

Here, D1 explains how they continuously refine their systems based on user interactions and feedback to ensure that the AI remains accurate and user-friendly. Lastly, D1 mentioned the use of guidelines as helpful in ensuring the development of safe and user-friendly AI applications:

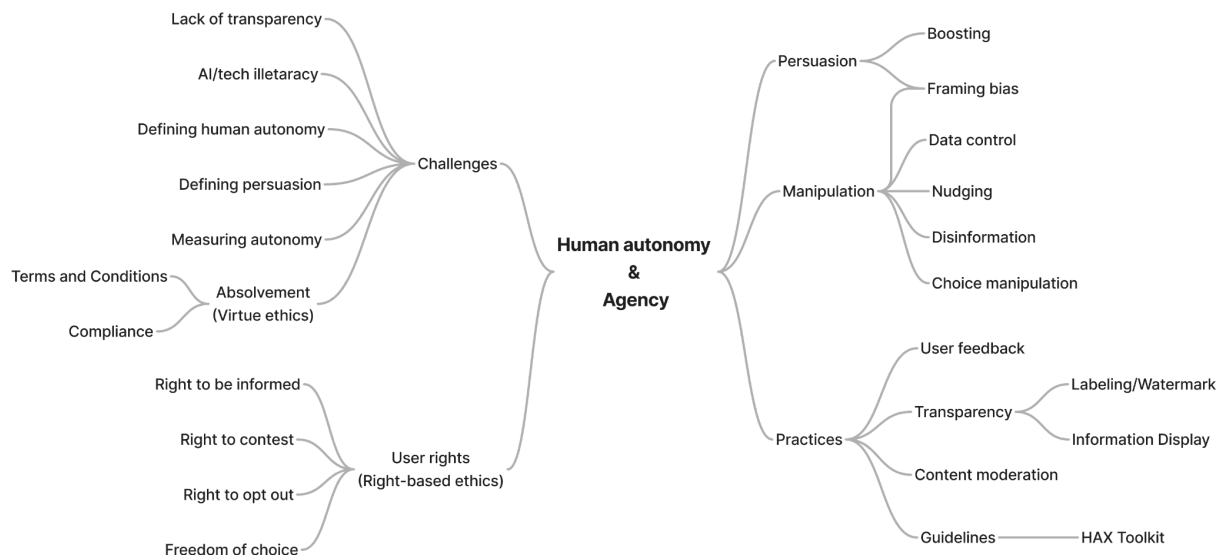
*“Design guideline or like a best practice guideline from Microsoft called HAX Toolkit. [...]. And there are certain things in the guideline, for example, that it is important to explain the users what the AI is capable of, how accurate it is, give them different options to cancel the AI recommendation or notify the users where the source came from. So there are like a lot of techniques that can be used to deal with it”* (D1, Appendix E)

Here, he highlights Microsoft’s HAX tool kit as a good guide for ensuring several of the user rights already discussed.

Despite the many practices mentioned above, ensuring human autonomy and agency in AI technology is challenging. As already mentioned, it can be difficult to know when you are being influenced and to what degree. C1 highlighted this paradox: *“[...] you lose autonomy, so you're not really aware of what decisions you're actually making.”* (C1, Appendix C) These challenges highlight the need for improved transparency and better education on the impact of AI in relation to human autonomy and agency.

**Figure 12**

*Thematic map representing main themes and subthemes associated with human autonomy and agency*



#### 5.2.4. Transparency and explainability

Transparency and explainability have proven to be important principles not only on their own but also in ensuring several of the principles we have already discussed. For fairness and justice, R1 suggested approaching the issue of bias with transparency of the AI system's capabilities rather than trying to correct them. Furthermore, several participants highlighted information transparency and explainability through practices like labeling and information display as crucial aspects in ensuring human autonomy and agency.

The primary issue addressed by the participants regarding transparency and explainability was "the black box," aligning with our findings in the literature review. R1 articulated the complexity of extracting meaningful information from large AI models: *"The problem is that they just do so many things, the models are so big so extracting any meaningful information out of that? No"* (R1, Appendix H). He elaborated, talking about large language models: *"I mean there are efforts in terms of how to apply explainability in certain areas, but what you might do is you might also get a model to produce a reasonably sounding explanation, but whether that's true"* (R1, Appendix H). Furthermore, he noted that even if the models could provide a correct answer to why they made a particular decision, that might not be a sufficient solution either:

*"But asking ChatGPT-4, why it replied what it replied is it is meaningless. I mean, you go in and look at the, what is it, I don't know, 1.7 trillion weights or whatever it has, and then you can see exactly how information propagated, so you can know exactly why it made that decision. It's just not going to be very informative because it's so complex."* (R1, Appendix H)

Reflecting on this issue, C1 pointed to a lack of knowledge from the developer's side as a part of the problem:

*“It's definitely a challenge if we have developers not knowing what is going on, like if they do not understand why a choice has been made by any system. It shouldn't be allowed to go into production. It shouldn't be allowed to reach an end user. I mean that's I think that's pretty standard if the if the person who makes the thing doesn't know how it works, then yeah, maybe we should reconsider.”* (C1, Appendix C)

However, G1 expressed that this might be the reality of today's AI landscape by stating, *“What we're hearing from OpenAI is that they don't understand it”* (G1, Appendix F), and *“So yes, actually I do think that even the people working with it will not be able to explain how it works, how it gives you this or that response”* (G1, Appendix F). D1 also underscored this point, saying, *“So we do have some knowledge about how it works, but we cannot fully explain its behavior and functioning”* (D1, Appendix E). Here, he is talking about the AI model used in his nutrition app. However, it is important to note that he has not developed this model himself but is employing an open model for the app's predictions.

In an attempt to approach the “black box” problem, C2 described how they try to break down the AI system into smaller, more manageable components: *“Yeah, we we try to to break up the, the the different areas where we use AI in as small as possible areas instead of having one big black box, we will have smaller boxes that we combine”* (C2, Appendix D). He also emphasised that when they introduce their AI systems, they try to provide as much information as possible about how it works and what to be aware of. However, not everything can be explained: *“But to a certain limit where we can't explain anymore due to the black box phenomenon that you're not able to explain.”* (C2, Appendix D)

Although the “black box” is a big challenge for transparency and explainability, the participants discussed various approaches to ensuring the principle. As an example from his work, R3 mentioned the creation of AI-generated personalized messages to job seekers that highlight why they would be a good fit for certain roles.

*“Also this I I feel like the personalized explanations that we try to generate in this first project right that fits under transparency or explainability for us, right, we're trying to explain you are getting this recommendation because you have this work experience and for the recruiters what we tried to do is.”* (R3, Appendix J)

Furthermore, R3 discussed the importance of moving beyond simple correlations to understanding causal relationships in AI systems: *“They're also working on causality explaining like a causal relationship, just not just an association that you see in the data”* (R3, Appendix J). However, R4 disagreed with this view, stating that it is the wrong perspective. He emphasized that the focus of explainable AI should be on protecting citizens and serving their rights, such as the right to contest decisions:



*“So what you need to be able to explain is not so much about causality. It's about, you know, which states have you used for training and then afterwards for testing, what is the quality of those data? Have they been tested for bias? Are they likely to result in discrimination? Have they been? How accurate is your system? How did you test accuracy and all of that? These are the kinds of explanation you should be able to give”* (R4, Appendix K)

R2 also underlines the importance of making the training data and the basis for AI decisions more accessible in order to increase transparency:

*“So it's always good to know what the system has been trained on, for example, right. And so you know there is transparency in the sense of what it is that the system can do and on what it is basing its decisions on, right?”* (R2, Appendix I)

He provided an example of a plant detection app where you take a picture of a plant to identify it. To support the app's answer, it can have a bottom panel showing similar images as an explanation as to why the AI came up with this answer, creating transparency about what kind of data it is using. However, when talking about this, it is important to remember that there are different types of AI models, and as R3 noted, *“There are certainly algorithms that are by default more explainable, like decision tree, right? Where you can see how you were going through this decision tree”* (R3, Appendix J). Therefore, this type of explanation cannot be expected to work for all AI.

While transparency is essential, it is not sufficient on its own, as G1 emphasized: *“Even if OpenAI opens up the hood to their software we wouldn't be able to understand it”* (G1, Appendix F). Therefore, practices for ensuring explainability are key, with several of the participants contributing with suggestions and examples, such as G1 noting, *“Funny enough, we need an AI to help us with that. We don't understand what we're looking at.”* (G1, Appendix F)

Although the paradox of AI explaining AI might be a future solution, despite the challenges with this pointed out by R1, other participants highlighted current practices. R2 pointed out the concept of post hoc explainability, which focuses on explaining decisions that have already been made:

*“Well, I recently read something very interesting about explainability which is about that post hoc explainability right that now what we're doing is that most of the time we all are trying to explain a decision that has already been made, right”* (R2, Appendix I).

Furthermore, R3 mentioned another explainability practice, providing an example from Google searches, where the query terms are highlighted in the results to show why certain documents were retrieved:

*“If you search for something, then your query returns will be highlighted so you can see in the results snippets how many of your query terms are like they're marked in yellow or something right in green and you can see oh yeah, this document has a lot of the search terms right” (R3, Appendix J)*

This kind of feature helps users understand the relevance of the search results. R3 also mentioned the use of Shapley values in explainable AI to show which features are most predictive of a particular class:

*“And there's a lot of work on explainable AI and trying to sort of visualize this one thing that you can do is these so-called Shapley values, right, for features where you can say, OK, well, these features are more predictive of this particular class. So that helps as a sort of explanation” (R3, Appendix J)*

C1 emphasized the importance of documentation throughout the AI development process so that the decision-making process of AI systems can be communicated clearly to end-users.:

*“I think it's a lot about actually having documentation alongside the whole process so that you could actually explain to the end user what is going on. And being able to actually do that in a way that people understand” (C1, Appendix C).*

Finally, C1 predicted a future where AI interactions would be more transparent to users:

*“I think that's one of the things I think we're going to see in the coming future, is that every time you are interacting with an AI, it's going to be more transparent and saying this is an AI and this is what it's doing. It's based on this data or whatever it is” (C1, Appendix C).*

This vision aligns with the broader goal of making AI systems more understandable and trustworthy for the public.

Despite these many practices and approaches, there are still challenges in ensuring transparency and explainability. Aside from the major black box challenge, one other issue is explaining AI systems to people with little technical knowledge or those who are not particularly interested in understanding the intricacies of AI. This difficulty was highlighted by G1:

*“But I think it's very difficult because the things, the technical aspects of it, if you're not looking for it and if you're not asking to be explained, I think it's difficult to explain that.” (G2, Appendix G)*

C1 shared a similar view, emphasizing that also not everyone is interested in these technical explanations:

*“Because I think that's also a big problem like it's very difficult to explain to someone what an AI is doing and how it's doing it to someone who just, you know, wants to use something on their iPhone. Yeah, because people are not always interested, actually, in knowing what's going on.” (C1, Appendix C).*

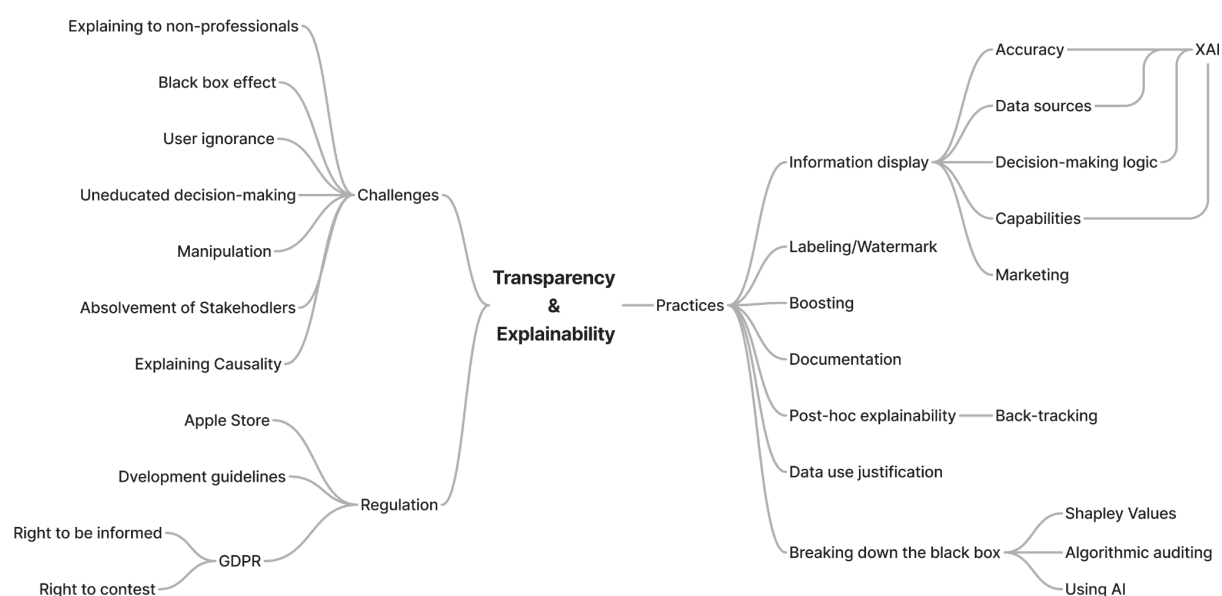
When asked how she would suggest approaching this issue, she highlighted awareness as an important factor:

*“I think maybe it's a very diplomatic and optimistic answer, but I think just having awareness around it, even if there's not a specific good practice, is the best way to go. Like if public services and and other platforms that have large reach, always and actively tell their consumers and their customers that they're they might be exposed to something.” (C1, Appendix C)*

To summarize, transparency and explainability are fundamental principles that intersect with other ethical considerations such as fairness, justice, and human autonomy. Despite the challenges posed by the black-box nature of many AI models, participants emphasized approaches such as breaking down AI systems into smaller components, providing personalized explanations, and documenting the AI development process as essential practices. However, the complexity of AI systems and the lack of interest or understanding among users present ongoing challenges.

**Figure 13**

*Thematic map representing main themes and subthemes associated with transparency and explainability*



### 5.2.5. Accountability and oversight

The principle of accountability and oversight has proven difficult to address with concrete answers or suggested practices. One of the reasons for this is that accountability in AI systems is highly context-dependent. R1 highlighted that the assignment of accountability varies based on the specific scenario and application:

*“I think that in my work, it depends on really on the particular situation and the application and and there are clear cut case where it would either be definitely the underlying system or definitely the human being, and there might be some cases in between as well” (R1, Appendix H).*

This underscores the necessity of evaluating each situation individually and that it can be a system responsibility, human responsibility, or a shared responsibility of the two.

Furthermore, in many situations, the issue of accountability can seem ambiguous. C1 stated *“I think the challenge with it is that it's very difficult sometimes to determine who has the responsibility or who is going to have the oversight of what is going on” (C1, Appendix C).* Underscoring this, R3 expressed uncertainty about who would be accountable if a mistake is made by the recommender system he is working on, questioning whether that would be the CEO, the recruiters, or him and his colleagues. This highlights the complexity of assigning accountability in AI applications, suggesting the need for clear guidelines to navigate these challenges.

In discussing practices for ensuring accountability and oversight in AI systems, the concept of human-in-the-loop emerged as a central theme. R2 emphasized that the goal of AI technology is not to replace humans but to support them and enhance human capabilities through collaboration with AI: *“[...] it is not only the AI and the machine learning that should be in the lead right, it's that basically human AI collaboration” (R2, Appendix I).* He elaborated, *“[...] it's not that we are trying to automate people away. You're trying to augment people right, with AI. We're trying to let people excel through the collaboration with the AI, but still with humans in control” (R2, Appendix I).*

Similarly, R1 stressed the importance of keeping humans responsible for AI-driven decisions:

*“So I think that if we can introduce AI but still keep humans responsible saying like yes, but you cannot say you did that just because the system recommended it you have to look under the hood and and and and really take responsibility for whatever decision is that's taken,” (R1, Appendix H)*

Although he expressed doubt about the feasibility of this ideal, it underscores the importance of human oversight. R4 further highlighted the necessity of human involvement, especially in high-risk scenarios: *“Well, I think that if there if this is a high risk kind of thing, right? Then*

*there should ideally always be a human in the loop that has the end responsibility” (R4, Appendix K).*

C2 provided practical insights into this collaborative approach, mentioning their ongoing experiments with AI and human collaboration: *“So the combination of AI and human being is where we are experimenting at the moment” (C2, Appendix D).* Additionally, R3 is also utilizing this approach, explaining: *“And in terms of autonomy, we’ve also stressed this is human in the loop, right, you’re the end person that decides[...].” (R3, Appendix J)*

Furthermore, having a dedicated team responsible for each AI system was mentioned as a current practice. C2 noted, *“But we always have a team responsible for for each AI instance that we we have” (C2, Appendix D).* This approach ensures that there is always a specific group of individuals accountable for the performance and decisions of each AI system, thereby ensuring clear accountability.

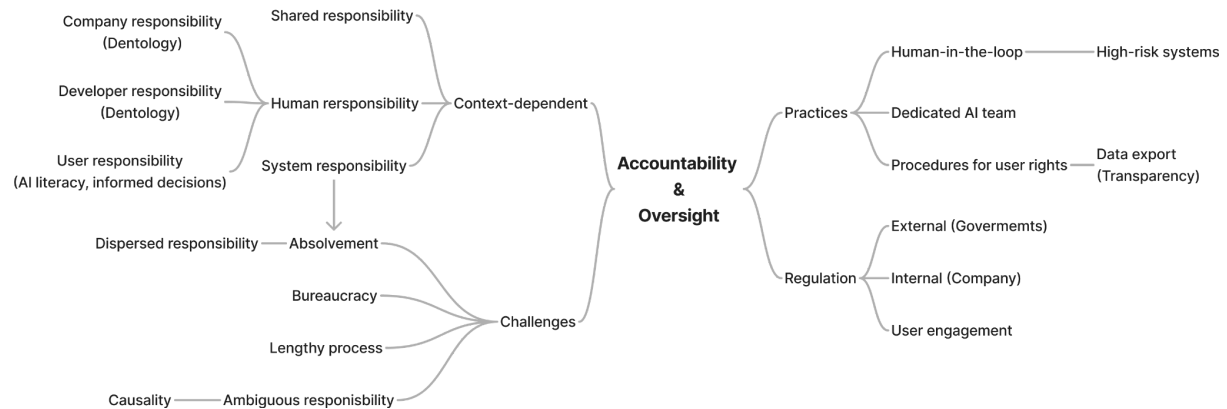
One of the key challenges in ensuring accountability and oversight in AI systems is absolvment, even when practicing human-in-the-loop. R1 described a scenario involving custody cases, where a person might choose to follow the AI recommendation despite disagreeing with it simply because they might think, *“you know, my back is covered if I just go with the recommendation” (R1, Appendix H).* Following the AI recommendation might provide a sense of security, resulting in a situation where *“you sort of remove some responsibility from the human being, taking the decision, right?” (R1, Appendix H).*

Another significant challenge is the complexity and procedural delays associated with ensuring user rights when AI systems make unfair decisions. R3 pointed out that if a person is unfairly treated by an algorithm deciding whether they should get parole or not, then *“Of course, it’s the jurisdictional system that needs to be held accountable to some degree” (R3, Appendix J).* However, he emphasized that addressing such issues involves lengthy bureaucratic processes, highlighting how the system’s complexity can delay justice and complicate efforts of holding someone accountable.

In summary, accountability and oversight in AI are crucial aspects of ensuring ethical and responsible AI technology. We have established that accountability is context-dependent, requiring situational evaluation, and how practices like human-in-the-loop highlight the importance of human oversight. However, challenges such as absolvment and ambiguous responsibility underscore the complexities involved in maintaining accountability. During the interviews, the participants also shared various views on how responsibility should be assigned to the different AI stakeholders. However, this will be presented in a later section, 5.3.5. Responsibility, where we compare what responsibilities are associated with each stakeholder group.

**Figure 14**

*Thematic map representing main themes and subthemes associated with accountability and oversight*



## 5.3. AI Stakeholders

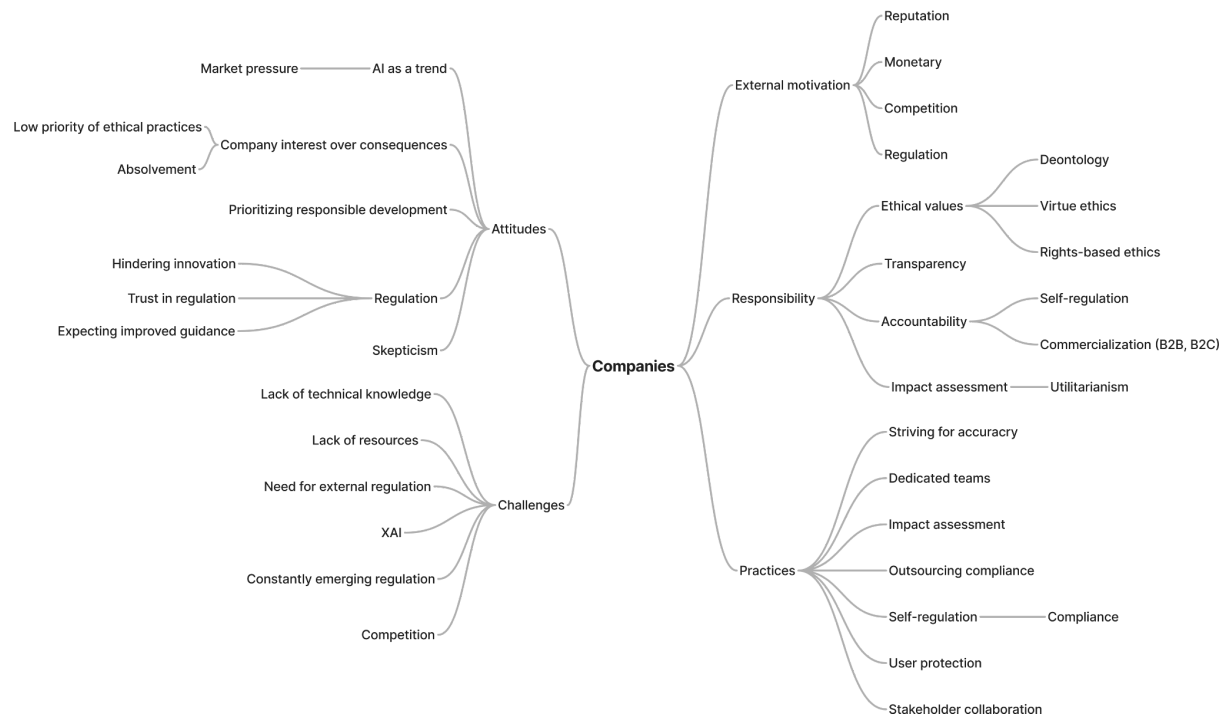
When discussing the ethical principles with the participants, different aspects of their viewpoints on the ethical implications of AI-driven persuasive technology (RQ2) have surfaced as main themes, such as challenges, attitudes, motivations, and practices - both currently applied and suggested. Additionally, in the last part of the interviews, all interviewees were asked about their views on other AI stakeholders, including their responsibilities and attitudes. The stakeholders were specified as the four main groups defined in our project: companies, developers, governments, and researchers.

### 5.3.1. Overview

To have a holistic and comprehensive overview of the themes associated with each stakeholder, we created thematic maps that represent the main themes and subthemes, as well as the relations between these, for each stakeholder based on the collected data. In the following paragraphs, we describe the general outlines of patterns identified in each stakeholder group.

**Figure 15**

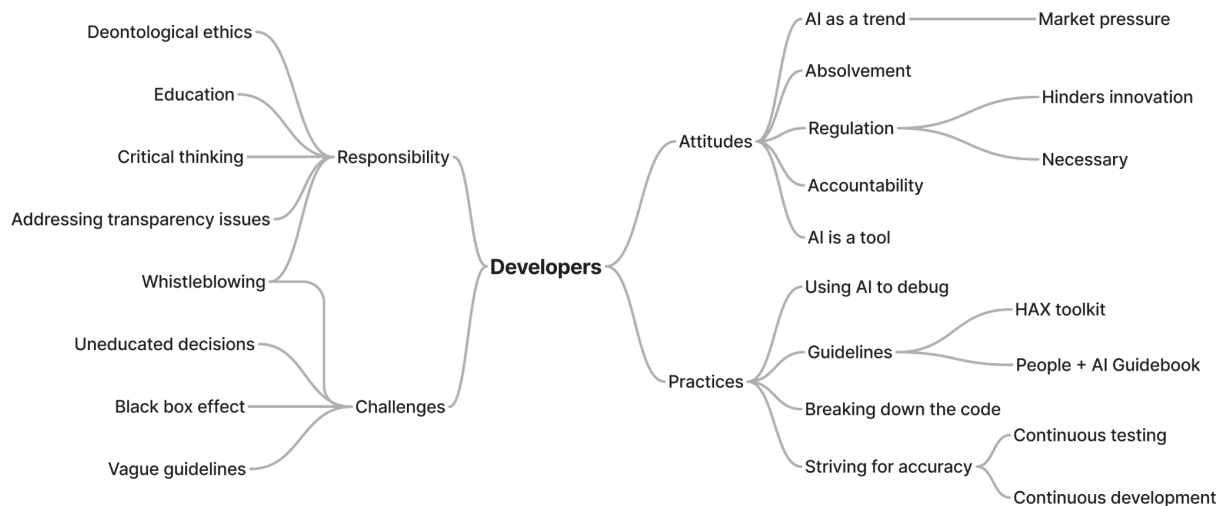
*Thematic map representing main themes and subthemes associated with companies*



**Figure 15** represents the main themes and subthemes related to companies entailing several challenges, including a lack of technical knowledge and resources, the need for external regulation to manage the “black box” effect, constantly emerging regulations, and competition. Companies often view AI as a trend driven by market pressure, leading to a low prioritization of ethical principles and a focus on company interests. Attitudes towards regulation were mixed, with some seeing it as a hindrance to innovation, while others recognized its necessity for reputation and compliance. Participants emphasized that responsibility in AI involves upholding ethical values, ensuring accountability through self-regulation and responsible commercialization, maintaining transparency, and conducting impact assessments with a utilitarian approach. Practices included striving for accuracy, forming dedicated teams, outsourcing compliance, self-regulation, user protection, and stakeholder collaboration. As participants stated, the motivations behind these efforts are driven by reputation, regulatory compliance, monetary incentives, and market competition.

**Figure 16**

*Thematic map representing main themes and subthemes associated with developers*

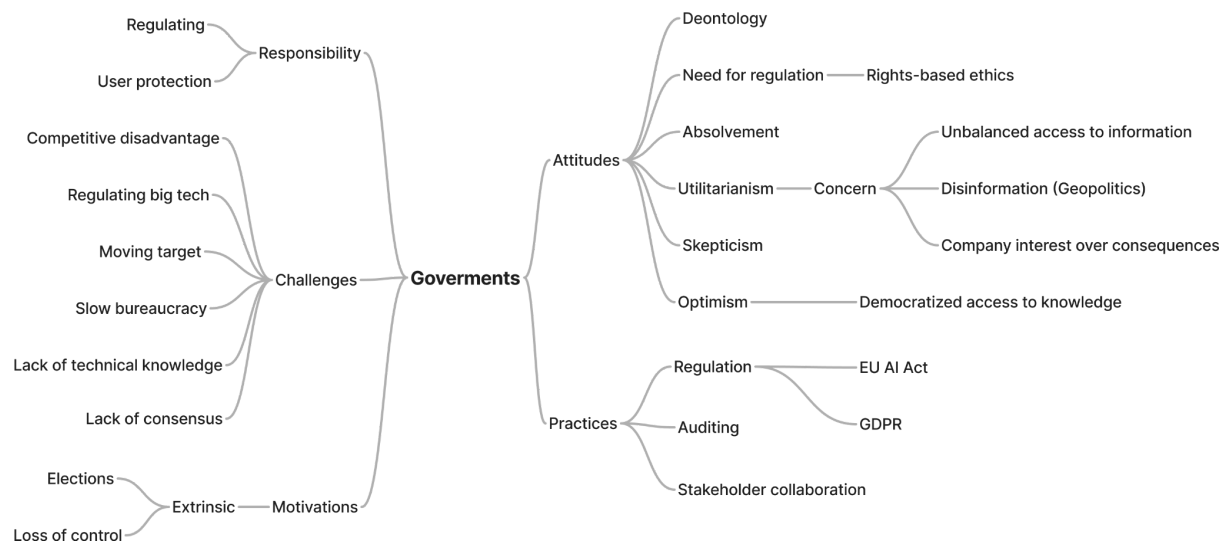


As visualized in **Figure 16**, AI developers face several challenges, including issues related to whistleblowing, making uneducated decisions, the black box effect, and vague guidelines. The theme of attitudes reflects seeing AI as a market-driven trend, acknowledging that regulation, although hindering innovation, is necessary, and viewing AI as a tool while emphasizing accountability. Participants assigned developers certain responsibilities for educating themselves, fostering critical thinking, addressing transparency issues, and whistleblowing. The practices theme entailed subthemes such as focusing on breaking down code, using AI for debugging, adhering to guidelines such as the HAX Toolkit and People+AI Guidebook, and striving for accuracy through continuous testing and development. Finally, the motivations theme included personal moral compass and market pressures as motivating factors.



**Figure 17**

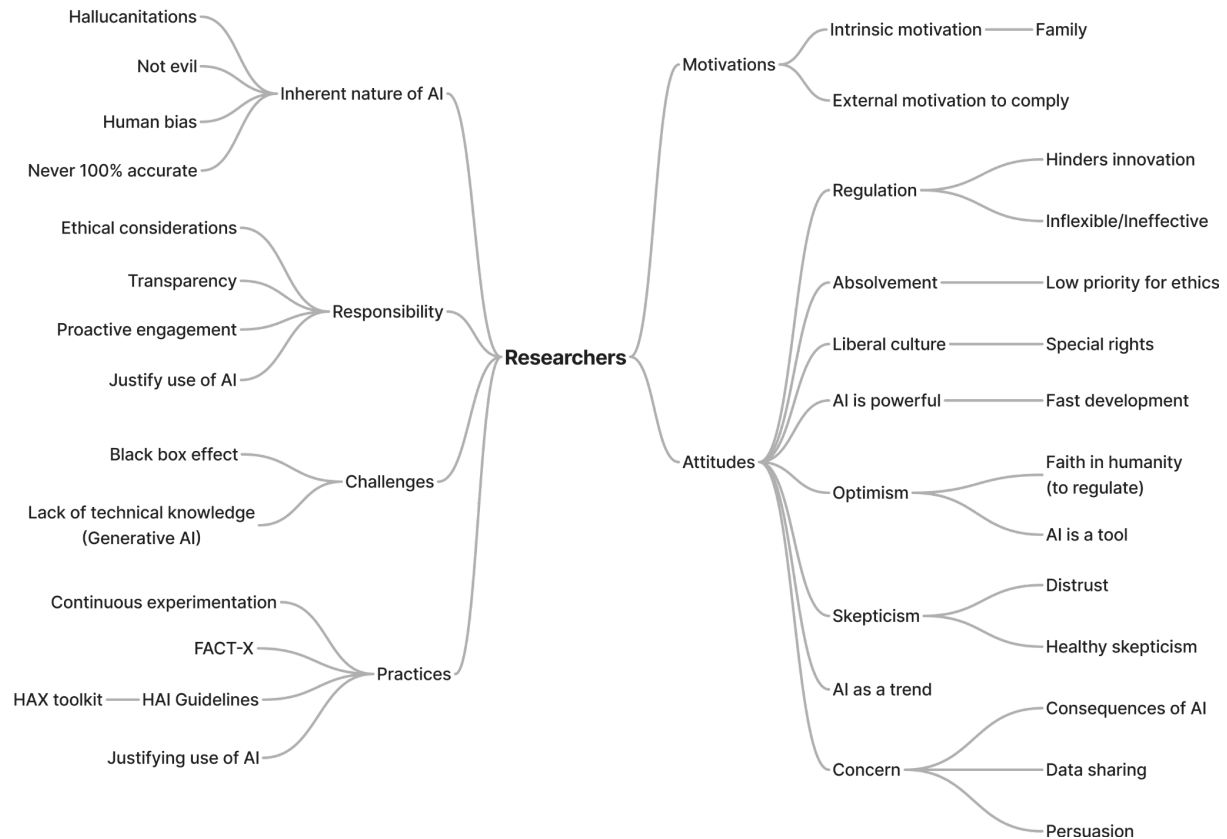
*Thematic map representing main themes and subthemes associated with governments*



For governments as stakeholders, the theme challenges in regulating AI contained subthemes such as competitive disadvantage, regulating big tech, dealing with a constantly evolving field, slow bureaucracy, lack of technical knowledge, and lack of consensus, as **Figure 17** shows. Attitudes included recognizing the need for regulation, seeking absolvment, and having concerns about unbalanced access to knowledge, disinformation, and companies prioritizing their own interests over consequences. On the other hand, attitudes also embodied subthemes such as optimism about democratizing access to knowledge and skepticism in relation to persuasive AI technologies. Participants expressed opinions according to whom, governments are responsible for regulating AI and protecting users. Their practices involved subthemes such as implementing regulations like the EU AI Act and GDPR, conducting audits, and fostering stakeholder collaboration. The motivations theme embraced motivations such as elections and the potential loss of control.

**Figure 18**

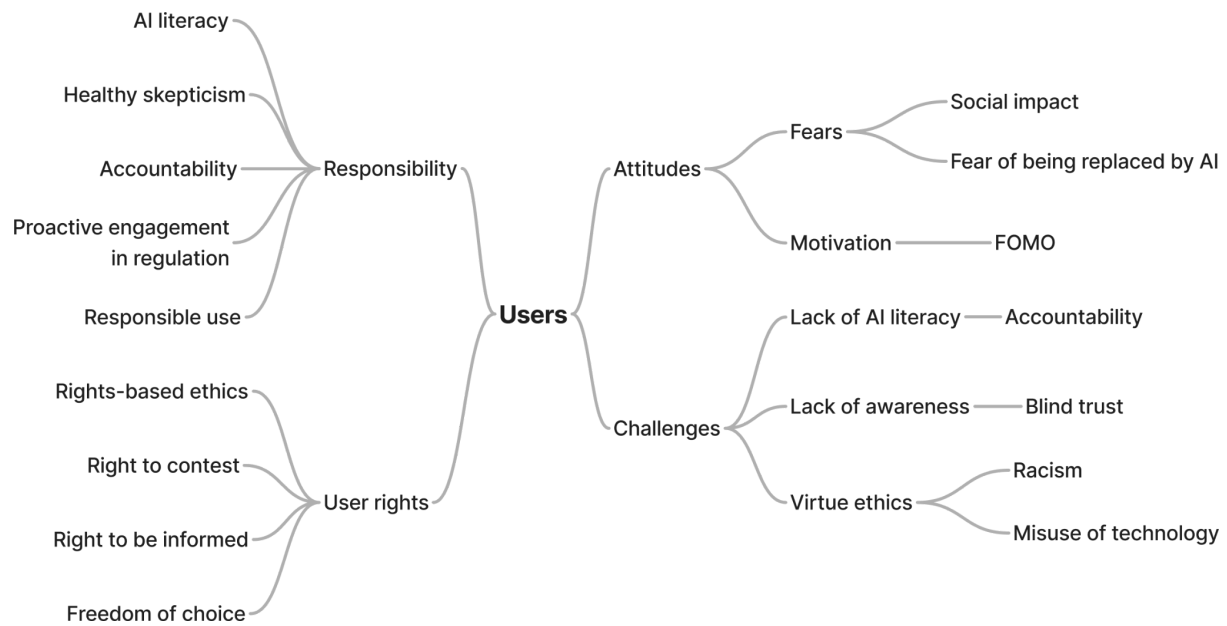
*Thematic map representing main themes and subthemes associated with researchers*



**Figure 18** represents the thematic map created to interpret themes associated with researchers. The challenges theme encompassed subthemes such as the “black box” effect and a lack of technical knowledge in generative AI. Attitudes included viewing regulation as a hindrance to innovation and as inflexible or ineffective, with some participants seeking absolvment and placing a low priority on ethics. Participants of the researcher group perceived AI as a trend, powerful, not inherently evil, but always biased and never 100% accurate. They expressed both optimism and skepticism, with concerns about the unethical sharing of data and the negative consequences of persuasive AI. The practices theme associated with researchers included continuous experimentation, following guidelines, and justifying the use of AI. Motivations entailed factors such as compliance, a moral compass, and the safety of their families.

**Figure 19**

*Thematic map representing main themes and subthemes associated with end-users*



Moreover, an additional fifth stakeholder group emerged as a main theme, as the participants representing the initial four stakeholders highlighted the role end-users play in ensuring the ethical and responsible use of persuasive AI technology. **Figure 19** represents the thematic map created to visualize the themes associated with the end-users, specifically challenges, attitudes, responsibility, and user rights.

Subthemes of the challenges faced in relation to ethical AI in persuasive technology are racism, misuse of AI and lack of AI literacy of users, which affects the sense of accountability, cultivating blind trust in technology, and uneducated decisions. Attitudes entailed fears of being replaced by AI, concerns about the social impact of AI, and the motivation driven by fear of missing out (FOMO). Participants emphasized that users have responsibilities to improve their own AI literacy, maintain a healthy skepticism, hold themselves accountable, proactively engage in regulation, and use technology responsibly. Furthermore, the user rights theme emerged, entailing the right to contest AI decisions, the right to be informed, the right to opt out of AI systems, and the freedom of choice regarding AI usage, as discussed in relation to human autonomy and agency.

In the following sections, we use a comparative approach to understand the main themes and subthemes associated with the stakeholders. This allows us to identify potential patterns of similarities and differences among stakeholder viewpoints regarding the ethical use of AI in persuasive technology.

### 5.3.2. Challenges

The first main theme that emerged from the thematic analysis, as summarized in **Table 7**, is the challenges that participants experienced or observed in relation to ensuring ethical AI development, especially when it comes to persuasive technology.

**Table 7**

*Summary of challenges experienced by stakeholders in relation to ensuring ethical use of AI in persuasive technology*

Companies	Developers	Governments	Researchers
Lack of technical knowledge	Uneducated decisions	Lack of technical knowledge	Lack of technical knowledge (GenAI)
AI as a trend	AI as a trend	AI as a trend	AI as a trend
XAI (“Black box” effect)	“Black box” effect	Moving target + slow bureaucracy	“Black box” effect
Constantly emerging regulation	Vague guidelines	Lack of consensus	Lack of resources
Lack of resources	Whistleblowing	Regulating big tech	
Competition		Competitive disadvantage	
Need for external regulation			

The first prevalent subtheme when it comes to challenges is the lack of knowledge or education about AI technologies. Participants associated this theme with each stakeholder group, and it was often paired with AI being a trend rather than a necessity, for instance:

*“they will have golfed with one of their other CEO buddies, and their CEO says bragged saying, yeah, we're doing AI now, right. And he has no idea what they're doing, right. But he knows that they're doing something. And then the other ones like, shit, we're falling behind and we need to do something as well when he comes in the Monday morning says we need to do AI. And then people in the development department... What the fuck are you talking about? What? What is it that we need to do? [...] For what? And so a lot of this is also pushed without a need sometimes.”*  
(R3, Appendix J)

The statement suggests that AI being a trend is a challenge for developers and researchers as well since they might be pushed to implement AI even if it is not necessary. This pattern was also identified in the context of governments, with G1 stating: *“I’m not going to be one of the tech guys to compare it to magic, but we don’t know how it works.”* (G1, Appendix F) and *“9 out of 10 politicians do not care about GenAI unless the population do.”* (G1, Appendix F). These statements suggest that politicians, despite not understanding AI, address it because it is popular. They also suggested that researchers *“definitely have something to write about”* and that there is a lack of researchers who understand the newest AI technology because the few people that do are hired by private big tech companies, such as the ones in Silicon Valley:

*“We also have a generation of academics that don’t understand vector models and GenAI. No one does maybe [...] less than 100 people in the world, and none of them are in academics. All of them are in Silicon Valley making millions.”* (G1, Appendix F)

This also suggests that academia and governments lack the resources to attract people with such knowledge and expertise, which could aid in the advancement of regulation and the formulation of guidelines.

Lack of knowledge manifested also in relation to developers leading to uneducated decisions when C1 stated, *“I think there’s something dangerous about people not knowing what is actually like the theory behind everything, just implementing things”* (C1, Appendix C). This concern has also been justified by D1 saying that *“[...] we are using an external AI service. So we do have some knowledge about how it works, but we cannot fully explain its behavior and functioning”* (D1, Appendix E), suggesting that some developers indeed implement the technology without fully comprehending how the models work.

One reason stakeholders struggle to understand the technology is the “black box” effect, which eight participants out of nine (C1, C2, D1, G1, R1, R2, R3, R4) mentioned. While G1 opposed calling the technology magic, D1 made the comparison, *“AI is kind of like a magic black box, which can tell you certain results even though it doesn’t know the context”* (D1, Appendix E). It also seemed to be a common comparison among researchers, according to R1: *“I know that many I researchers say it’s a black box [...]”* (R1, Appendix H). When it comes to companies, C2 stated that there is a *“certain limit where we can’t explain anymore due to the black box phenomenon that you’re not able to explain”* (C2, Appendix D), and G1 expressed concern by saying, *“What we’re hearing from OpenAI is that they don’t understand it”* (G1, Appendix F). There is also uncertainty around the degree to which big tech companies understand the models behind generative AI:

*“And we see OpenAI Superalignment Team being fired this week and they’re saying we don’t know how our own tool worked and then OpenAI said, yeah, because you were in the wrong department. [...] the C-Suite executive didn’t know how it worked, there’s not a lot of people in the company that knew how it worked”* (G1, Appendix F)

While governments are struggling to understand the technologies and, at the same time, find ways to protect the end-users, they face a big challenge when it comes to the pace at which the technology is developed, as represented by the “moving target” subtheme. Accompanied by slow bureaucracy, mathematically, it seems like an impossible task, as revealed by G1 and G2: “[...] the technology and the pace that we're developing and putting out services and products is so much faster than we could even think” and “[...] we see new technological increments every week in Silicon Valley, and it takes two years to build solid EU legislation, 2 years more to then have it into effect”(G2, Appendix G). Companies also seem to recognize the difficulty of regulating a moving target: “It's also very difficult for them to make a law about something that keeps changing all the time”(C1, Appendix C).

A second great challenge when it comes to regulation is the lack of consensus between governments and academia regarding regulations and guidelines. According to G2, “the approach to AI is very fragmented. There's many companies that have very well defined practices and then if you don't look at that specific company there's other other practices. So it's very fragmented” (G2, Appendix G), “there's no general rule book to look towards”(G2, Appendix G), and “the AI Act, [...] it's very broad. It's not very specific. There's a lot of questions that haven't been answered.”(G2, Appendix G). Researchers have also confirmed this issue by saying that “it's very much [...] trying to use a very blunt tool to regulate [...] it's ineffective”(R1, Appendix H), “the Human AI guidelines from Microsoft or the Google + AI playbook and so on [...] those tools are not straightforward to use [...] because when you design AI-infused systems, then there is that basically unpredictability of the system [...]”(R2, Appendix I), and “there's a lot of criticism that the EU AI act is too watered down, that it doesn't really regulate enough”(R3, Appendix J). C1 added that companies recognize these regulations are ineffective because they are difficult to implement: “even though I think it's great that we're getting legislation and different standards that we can use, I think it's very, very difficult to just apply it to all systems” (C1, Appendix C), while D1 expressed they didn't use guidelines or regulation, except for those required to publish an application on the App Store and the HAX Toolkit by Microsoft. Instead, he has been relying to some extent on personal judgment, stating that “it feels natural for us that users would expect certain requirements” (D1, Appendix E). This indicates that some developers might use a combination of utilitarian and rights-based approaches to creating ethical digital products.

When it comes to compliance, participants mentioned that employees might face challenges regarding safe whistleblowing, providing an example from Boeing, where a “whistleblower was recently found dead in apparent suicide by gunshot”(R1, Appendix H). Moreover, D1 also expressed that they “do not believe that businesses can regulate themselves”(D1, Appendix E), bringing up the example of the increasing number of Boeing jet accidents despite doing standard checks, leading to the theme of “need for external regulation.”

Another challenge related to regulation is the subtheme of competitive disadvantage, which many argue the EU is now facing in the global landscape of innovation, a criticism that has also been highlighted in the literature review. Participants (R1, R2, R3, D1) have associated regulation with competitive disadvantage in comparison to other states, such as the USA and

China, with them stating, “*They shouldn't smother innovation, and that is a tricky tightrope to walk*” (R3, Appendix J) and “[...] *in America, things are happening much faster because there is less conversation and [...] less regulation around AI*” (D1, Appendix E). It has also been mentioned on the national level in a sense that “*all this regulation may be a hindrance to AI uptake, and that may [...] mean that a company has closed or they [...] have to leave the market due to the competition*” (R4, Appendix K). This aspect overlaps with the theme of stakeholder attitudes and beliefs, which we will examine further in the following section.

### 5.3.3. Attitudes and beliefs

When analyzing the attitudes and beliefs presented by the participants outlined in **Table 8**, we employ ethical theories, such as deontology, utilitarianism, virtue ethics, and rights-based ethics, to identify the approaches they take to address ethical concerns in relation to persuasive AI technology.

**Table 8**

*Summary of stakeholder attitudes in relation to ethical use of AI in persuasive technology*

Companies	Developers	Governments	Researchers
Company interest before consequences	Concerns for negative consequences of AI - social isolation and polarization	Concerns of unbalanced access to knowledge and concerns of disinformation	Concerns for negative consequences of AI, data sharing, persuasion
Low priority of ethical principles	Low priority of ethical principles	Company interests over consequences	Low priority of ethics
Absolvement	Absolvement	Absolvement	Absolvement
Responsible development	Accountability		Liberal culture (Special rights)
Regulation hinders innovation	Regulation hinders innovation		Regulation hinders innovation and is inflexible/ineffective
Need for regulation	Need for regulation	Need for regulation	AI is powerful, not evil, always biased and never 100% accurate
Expecting improved guidance	Optimism (Democratization of knowledge)	Optimism (Democratization of knowledge)	Optimism
Skepticism		Skepticism	Skepticism

Each stakeholder type expressed concerns regarding the negative consequences of AI, suggesting that they have a utilitarianist approach that focuses on maximizing overall happiness and minimizing suffering for the greatest number of end-users. Other than the risks already discussed in relation to persuasive technology affecting individuals and society, some participants (G1, G2, R3) emphasized concerns relating to companies, specifically that *“the economic competition and gaining market access and selling most and becoming the main source of some specific service, can overshadow their wants to make it ethical”*(G2, Appendix G), leading to low prioritization of ethical innovation. Additionally, according to R3, the majority of companies handle sanctions imposed by regulators with the goal to solve them *“as quickly as possible with as less resources and make as much money as possible”*(R3, Appendix J). In other instances, low prioritization of some ethical aspects can be due to a lack of time or monetary resources and justified by their low impact, as described by D1:

*“[...] it's early stage startup, and we are trying to more focus on features that would help the users, we don't have that much time to explain the transparency and making sure that the application tells the users what they should expect [...] small businesses, they don't have such impact”* (D1, Appendix E)

The example of the early startup scenario above suggests that there is a difference in attitudes between big and small companies, as the latter have more resources to address ethical concerns. However, they still tend to deprioritize them due to market competition. When asked about the attitudes of companies in relation to regulation, G1 and G2 stated that there are big organizations, AI developers, and researchers that prioritize responsible development and stakeholder engagement:

*“I see the other companies that are trying to move slower but build more responsibly and involve more governments and NGOs the Anthropic, the Hugging Face, Google actually as well. Meta going open source and just being... Asking the world can you fuck around with our code? Then we also have people who can fix it. I like that approach a bit more.”* (G1, Appendix F)

*“[...] a lot of these large AI developers, researchers, companies do show up in forums that are only about ensuring ethical AI. They're they're sitting there, they're sending people, top people from their companies into some kind of conferences or summit for ensuring safe AI, so they are interested.”* (G2, Appendix G)

However, challenges to these organizations persist, as *“even the platforms that do think twice and do put in energy, money and labor in ensuring ethical solutions they might feel like they are pushed into putting out products that are not safe to keep up with the competition”* (G2, Appendix G). While other big tech companies, such as OpenAI, *“just get to see it's working in so far that it's marketable”* (G1, Appendix F) and have attitudes such as *“We don't have any responsibility because you can't prove that our system is persuasive”*(G1, Appendix F), which suggests that they prioritize making their products marketable over ensuring they are



safe to use. According to another viewpoint of G1 - *“I don't think that a lot of national politicians will do anything outside of saying... Someone should control that.”*(G1, Appendix F) - some government officials manifest it by acknowledging the importance of regulating AI; however, they point at other representatives when it comes to taking action.

In comparison, D1 exhibited some level of absolvment of responsibility by saying that *“I don't have the same responsibility as the person that creates the AI model”*(D1, Appendix E) and that he tended to use the terms and conditions as a “safety net,” reasoning that *“in the worst situation, we don't want to take the blame”*(D1, Appendix E).

Moreover, some researchers expressed that they consider ethical principles only to some extent in their research. This is an important first step, however, they *“leave that kind of ballpark to the ethics guys and to dig deeper into that”*(R2, Appendix I) when it comes to thorough ethical considerations, as they focus on developing and testing their solutions first. This indicates some level of absolvment of self, which has been identified as a common attitude relating to each stakeholder group. Self-absolvment and opposition in researchers might stem from previous academic privileges and liberal culture, as described by R4, *“In research, you can actually bypass much of the GDPR because you have special rights to access to data as a research.”* (R4, Appendix K) and *“the data collection and and data use has been the regulated to a very limited degree”*(R4, Appendix K). With recent regulations, however, academia will be more regulated, and R4 emphasized that many researchers consider it a significant barrier to innovation:

*“There would be a strong consensus among researchers that regulation is really a problem, and blocking AI development because there is a culture, there's a very liberal culture, where they have had extensive access to data for many years”* (R4, Appendix K)

As previously mentioned, participants (R1, R2, R3, R4, D1) expressed their beliefs about the regulation hindering innovation in both academia and the tech industry by stating that *“[...] they just wanted they focus on regulation and [...] we also have to be on the forefront and we have to be competitive and so forth, which is a shame in terms of research”* (R1, Appendix H), *“most of the industry they consider regulation a barrier”* (R2, Appendix I) and *“among researchers, I think regulation is considered a significant barrier to AI development and AI uptake”* (R4, Appendix K). On the other hand, C1 emphasized the need for regulation and expressed optimism and trust in governments in regulating big tech companies:

*“[...] I'm pro-law. I think it's important that we do set up some laws about this because I mean we have laws to kind of define what kind of society that we want to live in. And I think with AI it's very, very important that we start to define how we want to use it in society and how we don't want to use it especially.”* (C1, Appendix C)

*“I'm very hopeful that something is going to change. And I think it's going to be much easier for especially the EU to actually make changes in regards to the big tech companies” (C1, Appendix C)*

The same attitude pattern emerged among researchers when R1 stated that they think *“it is very important to try to build up these principles”*(R1, Appendix H), *“[...] humanity will overcome this as well as we learn more about these technologies and mature understanding of them.”* (R1, Appendix H), and *“we're gonna look back at that as being like there was very much the Wild West of data”*(R1, Appendix H). These statements indicate an optimistic attitude and trust in humanity to learn to regulate emerging AI technology.

Similarly, G2 emphasized that *“there is a need for regulation, even amongst AI developers asking for regulation or some kind of guidelines to have in place before they start developing”*(G2, Appendix G) because *“a lot of very, very important questions that haven't been answered with the EU AI act or any regulation”*(G2, Appendix G). The need for bilateral discussions about regulation has also been highlighted by G1, especially on the global landscape of emerging AI technologies:

*“[...] the bilateral discussions between the US, where a lot of this tech is being developed, and the EU where a lot of this tech is being adopted needs to be better [...] it needs to include China as well because they're also doing shit [...] If the EU isn't having this as their core task over the next couple of years, also during the Danish presidency, we're going to be so far behind.”* (G1, Appendix F)

Both D1 and G1 expressed a certain degree of optimism in relation to having a tool that democratizes knowledge. As described by D1 *“Steve Jobs [...] predicted this, that every person would have in on in their hand like a way to communicate with Socrates”*(D1, Appendix E), hinting at a utopian scenario where everyone can access education for free. In contrast, G1 discussed possible dystopian outcomes of not democratizing and ensuring the accuracy of such knowledge:

*“Can we then... argue that not everybody should have access to all of the knowledge? That is a really hard philosophical argument to hold that like, no, I don't trust those people with all of knowledge. Which people? The Jews, the Palestinians, the Russians, the Ukrainians? Like who do you not trust? People who are not educated, people who are not rich people who are needing to steal to feed their family? OK, like now we're just making a medieval class-based society with more steps. [...] if we ask them to control who has access to all the knowledge in the world is really fucking scary.”* (G1, Appendix F)

Moreover, G1 compared the boom of generative AI technology with the time when the first printing press or the internet was invented, saying that *“Maybe this is our generation's Gutenberg moment.”* (G1, Appendix F), hinting at the need for collaborative discussions about its regulation and democratization.

Lastly, a key subtheme of attitudes has been skepticism, or “healthy skepticism,” as referred to by C2, explaining that *“if you are talking to a computer, I would, as a human, always be skeptical of the answers I get”*(C2, Appendix D). This indicates that it is a crucial responsibility of the end-user to question the validity of responses generated by a computer, whether it is regulated or not, because *“we cannot trust everyone who is making AI systems.”*(C2, Appendix D). G1 had a similar viewpoint on persuasive technology, especially artificial agents, emphasizing that *“we should be very scared of them, and we should be very mindful of who's controlling the chatbots that will control us”*(G1, Appendix F) and *“nothing worth billions of dollars is objective”*(G1, Appendix F). Researchers have also emphasized the responsibility of end-users, which is closely related to AI literacy:

*“And there should also be some AI literacy. You should also understand, OK, these systems, what kind of data is it predicting on, right? This is a prediction model, right? Should I trust this? Like when I look at this, does this pass the smell test, sort of? Humans are also responsible to be critical towards this”* (R3, Appendix J)

#### 5.3.4. Motivation

When it comes to the subthemes that emerged in relation to the stakeholder motivation theme, it is important to distinguish between intrinsic and extrinsic motivations to be able to examine whether their behavior is self-determined, as outlined in **Table 9**.

**Table 9**

*Summary of stakeholder motivations in relation to ensuring ethical use of AI in persuasive technology*

Companies	Developers	Governments	Researchers
AI as a trend /Market pressure (Extrinsic)	AI as a trend /Market pressure (Extrinsic)	AI as a trend (Extrinsic)	AI as a trend (Extrinsic)
Regulation (Extrinsic)	Compliance (Extrinsic)	Elections (Extrinsic)	Compliance (Extrinsic)
Monetary: Sanctions, Profits (Extrinsic)	Internalization of ethical principles (Extrinsic)	Loss of control (Extrinsic)	Moral compass (Intrinsic)
Reputation (Extrinsic)	Amotivation		Safety of family (Intrinsic)
Amotivation			Amotivation

Based on the analysis so far, the AI boom created a trend among all four stakeholders, implementing and motivating them extrinsically in different ways. The data suggests that it is a form of introjection based on SDT, a somewhat extrinsic motivation for companies, governments, and researchers. This leads to less self-determined behavior, where actions are performed to gain approval from oneself or others (Ryan & Deci, 2000). As G1 described, “9 out of 10 politicians... Their main task is to get elected, so they will do what is popular” (G1, Appendix F), indicating that some government officials are motivated by gaining the approval of others, namely citizens, and winning elections. The latter can be described as an external motivation, where individuals are motivated by gaining an extrinsic reward, entailing monetary and prestigious benefits. Similarly, some companies are motivated by “*becoming the main source of some specific service*”(G2, Appendix G), combined with “*gaining market access and selling most*”(G2, Appendix G), implying monetary rewards. Finances are also evoking non-intentionality of companies to make ethical decisions, leading to what SDT describes as “amotivation”, based on a statement by C1: “*they want to have customers, they want to earn money and I think that's that's probably hindering them a lot from actually making some responsible choices in the development.*” (C1, Appendix C)

On the other hand, some companies are motivated to comply with regulations in order to avoid punishment, which is another form of entirely extrinsic motivation that leads to non-self-determined behavior. This aspect has been indicated by statements such as “*we can't afford to be off or or have any type of case where we are not compliant*”(C2, Appendix D) and “*Otherwise we will become obsolete and then we will go out of business*” (C2, Appendix D). Moreover, R3 expressed their views on the companies being amotivated to invest in a more ethical implementation due to lack of sanctions by describing:

*“One of my colleagues used to work for Yahoo for many years. And I asked them once, so what would they say if someone raised any ethical concerns? and then their bosses always say "as long as it's legal". As long as it's not illegal then we can do it.”*  
(R3, Appendix J)

Monetary sanctions and loss of customers have been a prevalent subtheme of extrinsic motivation for both companies and developers: “*Industries are not interested in regulating themselves. They will only do this if their customers abandon them because they realize that they fucked up, or because it costs them money.*”(R3, Appendix J) and “*there are certain decisions that requires us to think about the business first because, at the end of the day, if you don't make money, then your business's not going to survive*” (D1, Appendix E). The subtheme of reputation and market competition, as an external motivator, is common among companies and developers, according to D1, who is also an entrepreneur: “[...] *an example of this is Facebook. If the people are not believing in the app or they think that it's an app that would harm them, they wouldn't consider using it and giving their data, which is bad for the business.*”(D1, Appendix E) and “*if more users are trusting your AI solution, it can be reasons for the users to use your app instead of a competitor*”(D1, Appendix E). Moreover, D1 indicated a tendency to internalize the ethical guidelines, pointing to him identifying with

these and a value-based approach when stating that *“it feels it feels natural for us that users would expect certain requirements”*(D1, Appendix E).

Similarly, researchers are, on one hand, motivated extrinsically by the requirement to comply, supported by the following affirmations: *“if you're not writing that part in the paper, then you're not getting accepted”*(R2, Appendix I) and *“I'm forced to reflect on it because otherwise my work will not be will not be published or cannot be seen”*(R2, Appendix I); where the word ‘forced’ underlines nonself-determined behavior. However, as discussed in relation to attitudes, they reflect on ethical aspects only to the degree that it is required, as they focus on developing and testing the product first. Furthermore, R1 underscored that it depends on the individual how motivated they are to consider how ethical their research is by stating:

*“I work very close to researchers very concerned about ethics and raised the point all the time and also have other people who, you know, don't... other colleagues who are not concerned at all about what they're doing.”* (R1, Appendix H)

A similar case of amotivation has been identified in relation to developer attitudes, according to which *“it's some engineers that do not necessarily care much about how it is used and the the use of friendliness, but just get the technology straight”*(R2, Appendix I). This indicates that in the developer community and academia, it depends on the individual whether they are motivated or amotivated to consider ethical principles.

On the other hand, researchers R1 and R3 described motivations where protecting and supporting their family played a key role in ensuring the ethical use of AI in persuasive technology:

*“I have two kids [...] I try to limit it to limit the the the type of content that they get so they just don't get these empty calories but but they have something that's more substantial [...] What the hell are we doing? [...] our kids!”*(R1, Appendix H)

*“I have a daughter [...] Girls could be just as good ninjas as boys [...] I also have that in mind in, in the back of my head, right, that doesn't have to do anything with age discrimination or ethnicity or something.”*(R3, Appendix J)

These examples demonstrated the integration of extrinsic motivations and an internal locus of causality. Researchers as parents have assimilated with ethical principles, placing emphasis on parental duty and the morality of actions. Another aspect of the deontological approach to ethics has been described by R3:

*“I also believe that we are in a privileged position, right, because we live in a country where we're not hungry, we have a roof over our head. We have a good life, right. So that also means that you have a responsibility to help people that are less fortunate.”*(R3, Appendix J)

Other than duty-based ethics as a motivator, R2 mentioned a different type of extrinsic motivation, where they identify with the values of other researchers, with an expression like *“Emily Penn [...] those kind of voices motivate me to also echo this kind of ideas, both in teaching to some degree and more and more from year to year and also within research in that sense”* (R2, Appendix I). This type of extrinsic motivation has a somewhat internal locus of causality, as they consciously value and endorse the goals of an activity, which, in this context, is ethical development.

Finally, when it comes to government officials, the subtheme of losing control over the country has been emphasized by G1 as a key external motivator: *“[...] politicians, well... If you want to have any control over your country, over your region... You need to have control over this.”* (G1, Appendix F) and *“If you don't at least keep tabs on OpenAI, GenAI, you're not gonna be able to keep the reins, keep like the hands of the wheels in regards to the country that you hope to govern one day”* (G1, Appendix F). This type of extrinsic motivation leads to non-self-determined efforts to regulate emerging persuasive AI technology, which, as G1 stressed, has the potential to have catastrophic consequences, stressing that *“We've got to get really fucking smart, really fucking fast”* (G1, Appendix F).

### 5.3.5. Responsibility

Deontology, sometimes referred to as duty-based ethics, is a crucial aspect of ethically implementing AI in persuasive technology. In this section, we examine the subthemes identified (shown in **Table 10**) in relation to the responsibilities of each stakeholder group when it comes to ethical AI development and regulation.

**Table 10**

*Summary of stakeholder responsibilities in relation to ensuring ethical use of AI in persuasive technology*

Companies	Developers	Governments	Researchers
Accountability	Accountability	Accountability	Ethical considerations
Transparency	Addressing transparency issues	User protection	Transparency
Self-regulation	Education	Regulating	Proactive engagement
Impact assessment	Critical thinking		Justify use of AI

Based on the qualitative data, the responsibility of stakeholders comes in many different forms. However, each stakeholder group, including end-users, has a key role and responsibility to engage in discussions about the regulation and ethical use of AI as G2 emphasized, “[...] a lot of stakeholders that should be responsible, not only the services like the AI services, whether it be a company, a platform [...]”(G1, Appendix F). They also distinguished between private sector actors and types of users by describing that *“the responsibility of the private sector, AI developers is one thing, but also the ones that use and buy AI to optimize their products and their systems. [...] That specific area of not being an AI developer but being an AI consumer, but not being a private person”* (G2, Appendix G). These snippets of information entail that within the tech industry, it is vital to consider both B2B (business-to-business) and B2C (business-to-consumer) companies when discussing ethical implementation and use of persuasive AI technology. This presupposes the need to address scenarios holistically when it comes to B2B2C (business-to-business-to-consumer) AI-driven products.

When it comes to technology companies, G2 underlined that they have a great share of responsibility as they have the best understanding of how their product works: *“people that are creating a service, developing it from the bottom and then putting it out in the world and selling it and making money off of it should have more responsibility than, let's say, civil society”*(G2, Appendix G). The sense of responsibility was reinforced by C1 affirming that *“we have to be accountable for what we're making”*(C1, Appendix C). Moreover, D1 acknowledged the responsibility of companies and developers and added the governments as being responsible for protecting users from manipulation:

*“[...] it's also the responsibility of the state or [...] the company to be responsible for their users because they might create a certain type pattern of behaving which nudge the person to do something not in their best interest [...] it's the responsibility of us, the company or the developer of the app to take the blame [...]”* (D1, Appendix E)

The same argument was supported by G2, emphasizing that *“governments have a regular accountability to ensure that their citizens are protected from being manipulated”*(G2, Appendix G). The user protection subtheme was also identified when G1 stated that it is *“Margrethe Vestager's task to save us from all of our information being curated through OpenAI”* (G1, Appendix F), placing responsibility on the European Commission for ensuring user data protection on EU level. Meanwhile, on a national level, G1 underscored that *“there is a choice to be made about integrating GenAI”*(G1, Appendix F) in public services. Furthermore, G2 described the importance of assigning governments responsibility considering the contrary: *“not placing any responsibility upon the governments will just make everybody be in the hands of the developers and services and then not having anywhere to go to make the wrong right”* (G2, Appendix G). Therefore, governments are a third actor whose role is crucial in protecting the well-being of individuals and society, as there is a need for an external regulator end-users can reach out to for regulation.

The subtheme of accountability as a researcher's responsibility related to ethical considerations being a part of every research project. According to R2, showing awareness about the ethical implications of a project is a minimum requirement when publishing research papers, however, researchers are not required to solve these ethical issues, which is reflected in the following statement:

*“[...] if you publish at conferences and journals [...] you always reflect on what you're doing in some kind [...] section on ethics about what you have been doing [...] you are not like trying to solve that that issue, but you're at least showing awareness. And I think that's the first step [...] that's why everyone has his/her own area of expertise.”*

(R2, Appendix I)

Another common subtheme among participants in relation to stakeholder responsibility was transparency and addressing transparency concerns. R3 stressed the responsibility to be transparent among researchers and developers – *“we need to be transparent”*(R3, Appendix J) – while G1 pointed out that *“some people would need to write code, I guess, or vectors to neutralize that in the chatbot but that is why OpenAI pays their software engineers half a million dollars a year”* (G1, Appendix F), emphasizing that the few people who understand the emerging and pervasive AI technology have a responsibility to address explainability and fairness issues. However, the responsibility to be transparent was mainly associated with companies which they also pronounced: *“We also need to make sure that we're informing the people using an AI system of what is happening”*(C2, Appendix D) and *“[...] they have a responsibility to the users of their systems and that they have a responsibility to actually ensure that there is transparency and everything that they do every step”*(C1, Appendix C). Moreover, R3 also supported this argument, attesting that *“companies [...] have a big responsibility in making sure that their use of AI is fair, accountable, transparent [...]”*(R3, Appendix J).

The theme of company responsibility leads to the subtheme of self-regulation. C1 specified that *“it's important that each organization is very, very clear about what their like ethics are”* (C1, Appendix C), and C2 described AI development as *“a continuous balance between research and experimentation and making sure that the the end-users or people in using it doesn't get into any kind of harm”*(C2, Appendix D). Moreover, C1 expanded on the idea by stating that *“it's very important that the companies are also, from the beginning, aware of what they are doing that can ensure that it's not happening, but also be realistic about if it happens, [...] what can we actually do about it”*(C1, Appendix C) pointing to the significance of initial impact and risk assessment when developing a new product or feature.

As an addition to the responsibility of developers, two subthemes emerged, entailing education and critical thinking. The importance of education has been emphasized by both companies and researchers, affirming that *“there's a responsibility in being educated about what ethical AI is”*(C1, Appendix C) and *“they also need to learn about these technologies,*



*even if you just want to become a programmer and you just want to program” (R3, Appendix J). Moreover, R2 mentioned that “the overall challenge is [...] people are not using [...] the tools that are already out there”(R2, Appendix I). R3 expanded on the argument expressing that education and critical thinking are closely related:*

*“It's not just data science, that has this course at it. There's also reflections on computer science, right. If you just do computer science, all these programs have reflections on whatever it is that they're doing, which is where you should learn these things, right? You should learn to ask these kinds of critical questions.”(R3, Appendix J)*

When it comes to the subtheme of critical thinking, R3 emphasized that developers:

*“have a responsibility to push back against their CEOs for this kind of bullshit, like we “need to put AI in everything”, and “large language models are what we need”, and “can't we solve this with ChatGPT?”(R3, Appendix J) and they “need to be critical, especially here and say no, this is not going to work or this is bullshit or let's do this”(R3, Appendix J).*

Therefore, critical thinking is intrinsically linked with critical decision-making, and developers have a key responsibility to argue against unjustified or unethical use of AI. A similar pattern has been identified among researchers, pointing to them having the responsibility to use AI in a justified and ethical way:

*“researchers have been so obsessed with whether we could do something with machine learning then not really have looked at the question of is it actually that helpful and are there other ways of doing it, or so do we really need to do that?” (R2, Appendix I)*

Finally, proactive engagement has been a prevalent subtheme in relation to stakeholder responsibility of preventing manipulation in persuasive AI technology as, according to G2, *“not only the developers and the services, but also the governments have a responsibility to spread awareness and take down, literally take down or stop any situations where people are being manipulated”(G2, Appendix G). G2 also emphasized that “all stakeholders, even academia and civil society, has a role to play as long as it is cross-border cross nature” (G2, Appendix G) and “if not a responsibility to ensure ethical AI, responsibility to engage in the conversation, at least or be in the rooms where these things are discussed” (G2, Appendix G).*

As previously mentioned, the qualitative data has revealed a significant theme concerning end-users as a stakeholder group, particularly their responsibility in ensuring the ethical use of AI in persuasive technology, as summarized in **Table 11**. Therefore, this section will also discuss users' responsibility.

**Table 11**

*Summary of end-users as stakeholders in relation to their responsibility in ensuring ethical use of AI in persuasive technology*

Challenges	Motivation	User rights	Responsibility
Lack of AI literacy	FOMO (Motivation)	Right to be informed	<b>AI literacy</b>
Lack of awareness (Blind trust)		Right to contest	<b>Healthy skepticism</b>
Misuse of technology		Right to opt out	<b>Responsible use of technology</b>
Racism		Freedom of choice	<b>Proactive engagement in regulation</b>

This important aspect of responsible AI has been revealed, for instance, by G2, who described how governments are expected to constantly formulate new rights, absolving the end-users from bearing responsibility:

*“There's been a lot of criticism around if you keep wanting states to be responsible in every single aspect and look at it like they have the obligation to protect you that might be creating new human rights every day, and that's not very functionable. At one point, when does the person themselves be accountable? (G2, Appendix G)*

The subthemes of lack of AI literacy (or tech literacy) and lack of awareness have been prevalent in relation to participants mentioning user responsibility. For example, R3 highlighted that even if users are presented with information about how their data would be used, *“they don't know what this information can be used for. You know, right, because you're taking this degree, and I know because I'm a researcher, and it's my job to know this.”* (R3, Appendix J), underlining that if end-users are not educated in the technology field, they might not understand how their personal data could be exploited. Moreover, the participant added: *“If you give them complete control over this, they will share too much. So people have a really hard time in handling this kind of [...] consent.”* (R3, Appendix J), which indicated that users tend to blindly trust technologies with their data.

Another researcher added that *“most often they don't even realize that they are being influenced. Especially [...] in the case of social media influencers and such.”* (R1, Appendix H), emphasizing that lack of awareness can lead to being persuaded in subtle ways. Moreover, C1 pointed out that in the case of social media, some end-users are motivated extrinsically to keep using the platforms despite being aware of its possible harms,

identifying with them: “[...] if I was the only one not having social media, you know, it's kind of it's all the feeling left out and not knowing what's happening” (C1, Appendix C). This attitude indicates fear of missing out (FOMO), which is present, especially among young adults who want to keep up with what everyone else is doing, as also suggested by G1: “Every young person today, everybody between 15 and 25 is on a social media” (G1, Appendix F). This activity implies both extrinsic and intrinsic motivations, such as introjection, where users focus on the approval of others or inherent satisfaction by the enjoyment of the content recommended by social media platforms.

Another challenge of user responsibility is the subtheme of misuse of technology and racist attitudes, as identified in explanations by G1 in the context of generative AI. He stated, “The problem is not that someone learned how to build a bomb. The problem is that someone wanted to build a bomb” (G1, Appendix F), and that:

*“[...] people found out fairly fast that people could use Midjourney or Google Gemini to produce racist outputs [...] racism is a human problem. Solving human problems with technology, I've always been very iffy of that because the core of the problem is that some people will do racism. That's a problem with the racists. It's not a problem with the tools they use for racism.”* (G1, Appendix F)

To address these ethical challenges, all stakeholder types (C1, D1, G2, R1, R2, R3, R4) exerted a rights-based and deontological approach by unveiling subthemes such as the right to be informed, the right to contest, the right to opt-out and the freedom of choice. Ensuring transparency is a key aspect of regulation and safeguarding human autonomy in the EU, as described by G2:

*“[...] in Europe, it's the right to access trustworthy information and be in a safe environment where you either know that you might be exposed to AI or that you can actively choose to go into another platform or into another area where you actively choose”* (G2, Appendix G)

Companies and researchers have acknowledged the importance of preserving the right to be informed for the users: “ [...] it's important that people are actually aware of what they're interacting with”(C1, Appendix C) and “they should be informed about the content”(R1, Appendix H). The right to contest has been emphasized by researchers affirming that it should be easy for users to “reach out if they see themselves being mistreated”(R2, Appendix I), they “should have a way of contesting these results”(R3, Appendix J), and they have a right to state their “opinion on the automated decision-making and to contest decisions”(R4, Appendix K). The right to opt-out subtheme, however, was more prevalent among companies and developers, emphasizing that it is crucial to provide the users “the opportunity to say no and opt out of whatever that AI is doing”(C1, Appendix C), including the use of personal data and there is a “need to have a way for the users to delete their account”(D1, Appendix E). Finally, R1 argued that users should be given the freedom to choose, even if they decide to choose the option that might be harmful to them: “it's sort of like the same process as food

*and smoking and so forth went through like and I think that to a large extent, people should still be allowed to do what they want to do” (R1, Appendix H)*

The freedom of choice subtheme is closely linked with personal responsibility, as indicated by D1, stating that *“whatever you do on the Internet, it's on your own response responsibility, and you have the choice to act on it or not”*(D1, Appendix E). It is also linked with the subtheme of the right to be informed, as highlighted by R1: *“[...] you should be able to see what the system is trying to do to you so you can decide whether to use it or not”* (R1, Appendix H).

However, according to the participants, informing the users and providing them with the freedom of choice is not sufficient to ensure the ethical use of AI technologies. There are additional angles to it, such as AI literacy, healthy skepticism, responsible use of technology, and proactive engagement in regulation. Seeking education and applying critical thinking is not only the responsibility of the developers, as R3 described:

*“[...] there should also be some AI literacy. You should also understand, OK, these systems, what kind of data is it predicting on, right? This is a prediction model, right? Should I trust this? Like when I look at this, does this pass the smell test, sort of? Humans are also responsible to be critical towards this.”*(R3, Appendix J)

In addition to AI literacy, users have the responsibility to search for the objective truth behind the outputs presented to them by AI, as the subtheme of healthy skepticism suggests.

This viewpoint has been highlighted by companies and government representatives through statements like *“I would, as a human, always be skeptical on the answers I get”* (C2, Appendix D) and *“we should be very mindful of who's controlling the chat bots that will control us”* (G1, Appendix F). Moreover, G1 raised an interpretative question – *“Is it the phone in itself or is it the phone used by the user - me?”*(G1, Appendix F) – and argued that *“the usage of tools are not the responsibility of the developers”*(G1, Appendix F). C1 also supported this aspect by stating that their customers in relation to their technology, *“have to be accountable for what they're going to use it for”*(C1, Appendix C). These aspects point out that end-users, be they private individuals or organizations, have the responsibility to make informed, responsible decisions when using persuasive AI technology.

Finally, as all other stakeholders, end-users have a responsibility to proactively engage in conversations about regulation and ethical use of AI, based on the explanation of G2:

*“[...] civil society may not have the tools to be able to ensure responsible AI. They can scream at the top of their lungs to developers all they want, but they might not have the last say, but engaging all of them and putting in forums, platforms, whatever, where all stakeholders engage and spread awareness and keep on looking at the difficulties and when does it go wrong and share good experiences is the right way forward”* (G2, Appendix G)

### 5.3.6. Practices

The practices theme entails a multitude of viewpoints on the practices associated with each stakeholder, combining current practices, recommended practices, and participant opinions about what stakeholders from a different stakeholder group should implement.

As previously mentioned, stakeholder collaboration is a key practice, indicating the need for a shared commitment to engaging with diverse groups to promote ethical AI deployment. R2 emphasized this, stating, “*You need to have interdisciplinary collaborations that basically ensures that those different kind of values are put into the system and not forgotten*” (R2, Appendix I). However, differences in approaches become evident in how each group handles self-regulation and striving for accuracy, as summarized in **Table 12**.

**Table 12**

*Summary of stakeholder responsibilities in relation to ensuring ethical use of AI in persuasive technology*

Companies	Developers	Governments	Researchers
Stakeholder collaboration	Stakeholder collaboration	Stakeholder collaboration	Stakeholder collaboration
Self-regulation (Dedicated teams, Outsourcing compliance)	Guidelines (HAX Toolkit, People+AI Guidebook, App Store)	Regulation (EU AI Act, GDPR)	HAI guidelines, Hax Toolkit, FACT-X, GDPR
Striving for accuracy	Striving for accuracy (Continuous testing and development)	Auditing	Continuous experimentation
Impact assessment	Breaking down the code		Justifying use of AI
User protection	Using AI to debug		

Based on the practices highlighted when analyzing views in previous sections about the five ethical principles, companies could self-regulate to form dedicated teams and outsource compliance to meet ethical standards (C1), while developers like D1 follow specific guidelines such as the HAX Toolkit and the People+AI Guidebook to guide their ethical practices, and are regulated by the platforms on which they deploy their software, i.e., Apple App Store.

Governments (G1, G2) take a regulatory approach, exemplified by the EU AI Act and GDPR, enforcing strict compliance and accountability. Researchers have a combined approach,

focusing on GDPR (R4) and on frameworks like HAI guidelines (R2) and the FACT-X toolkit (R2, R3) to maintain ethical standards. In terms of accuracy, companies and developers both emphasize continuous testing and development, with C2 recommending more detailed methods, such as breaking down the code to address explainability issues. Governments should utilize auditing (R3), while developers and researchers should engage in continuous experimentation (C2, R2) to ensure the reliability and ethical integrity of AI technologies.

When it comes to assessing impact and ensuring user protection, distinct subthemes of methodologies emerged. Companies should perform impact assessments (C1) to understand and mitigate the effects of AI, whereas companies and researchers emphasize breaking down code (R1, C2) robustness and reliability. Another method has been suggested by G1, according to whom using AI should be considered for addressing explainability issues.

Finally, researchers have emphasized the need to justify the use of AI (R2, R3) by continuously evaluating its applications and outcomes, ensuring that ethical considerations are consistently met. This comparative overview highlights the multifaceted strategies employed by different stakeholders, each playing a critical role in the ethical implementation of AI in persuasive technology.

## 5.4. Summary of Results

The analysis has addressed our second research question, investigating *what are the viewpoints of AI stakeholders about current practices for addressing the ethical implications of AI-driven persuasive technology (RQ2)*. In the first section of the analysis, focusing on persuasive AI technology, the main themes identified were defining persuasion, techniques and technologies, as well as associated risks. Persuasion was distinguished from manipulation, emphasizing its cognitive and open nature, unlike the hidden influence of manipulation. The analysis highlighted industries like social media and journalism, which use algorithms to optimize user engagement and identified persuasive strategies such as recommender systems, nudging, boosting, gamification, emotional framing, and priming. Concerns were raised about the hidden strategies behind these technologies, with potential risks including psychological impacts like addiction and loneliness, as well as broader societal issues such as political polarization. The transition to generative AI in artificial agents further emphasized the need for transparency and ethical considerations to mitigate manipulation and ensure responsible use.

The main part of the interview focused on the views and experiences of the participants in relation to the five ethical AI principles: (1) beneficence and non-maleficence, (2) fairness and justice, (3) human autonomy and agency, (4) transparency and explainability, and (5) accountability and oversight. Participants discussed the challenges and practices associated with these principles, highlighting the complexities of defining and implementing them in AI development. Key themes included the necessity of clear definitions and boundaries,

stakeholder engagement, and continuous monitoring to ensure ethical standards. Participants expressed varying levels of uncertainty about how to measure and achieve these principles, particularly in areas like fairness, where biases are subjective and context-dependent. Transparency and user rights, such as the right to be informed and to contest AI decisions, were emphasized as crucial for maintaining human autonomy. Despite several practices in place, the principle of accountability and oversight was noted as challenging due to context-dependent responsibility and procedural delays. Overall, the interviews revealed a need for a critical dialogue on the ethical use of AI, with participants stressing the importance of questioning when and where AI should be used.

The analysis revealed significant challenges, attitudes, motivations, and practices related to the ethical implications of AI-driven persuasive technology among companies, developers, governments, and researchers. Key challenges include a pervasive lack of technical knowledge, the "black box" effect, and difficulties in regulation due to rapidly evolving technology and fragmented guidelines. Companies often prioritize market pressures over ethical principles, while developers and researchers struggle with vague guidelines and the need for transparency. Furthermore, governments face competitive disadvantages and slow bureaucracies in regulating AI. Despite these challenges, there is a consensus on the need for regulation to protect users and ensure accountability, although opinions vary on its impact on innovation.

Motivations driving ethical considerations in AI are predominantly extrinsic, such as market pressures, compliance, and reputation, with intrinsic motivations like moral compasses playing a lesser role. All stakeholders recognize the importance of accountability and transparency, with companies emphasizing self-regulation and impact assessments, developers focusing on continuous testing and adherence to guidelines, governments enforcing regulations like the EU AI Act and GDPR, and researchers justifying AI use through continuous experimentation.

Finally, end-users emerged as a central stakeholder group, with responsibilities to improve AI literacy, maintain skepticism, and engage proactively in ethical AI use and regulation. Overall, stakeholder collaboration is crucial for promoting ethical AI deployment, highlighting the need for shared commitment across diverse groups.

## 6. Discussion

In the previous section, the thematic analysis, we examined the viewpoints of AI stakeholders about current practices for addressing the ethical implications of AI-driven persuasive technology, thereby addressing RQ2. Based on the findings from the analysis, this section will address RQ3: *What should be the focus areas of stakeholders in future efforts to ensure ethical AI in persuasive technology?* In doing so, we propose eight focus areas for the four stakeholders to work towards the ethical use of AI in persuasive technology: (1) clear

definitions and boundaries, (2) stakeholder engagement and collaboration, (3) continuous monitoring and transparency, (4) addressing the “black box” challenge, (5) regulation and ethical standards, (6) education and AI literacy, (7) justified use of AI, and (8) balancing innovation and regulation.

This section summarizes the standpoints of the responsible AI stakeholders, showcases their attitudes toward prioritizing ethics in persuasive AI technology, and provides a similar overview of the various persuasive techniques and practices identified in the analysis. This serves as a conclusion to RQ2 and lays the foundation for the subsequently proposed focus areas for future research, addressing RQ3.

## 6.1. Standpoints of AI Stakeholders

### **Companies**

Companies face numerous challenges, including a lack of technical knowledge and resources, the need for external regulation, and the pressure of constant competition. Their attitudes towards AI are primarily driven by market pressures, often leading to a low prioritization of ethical principles. Companies typically view AI as a market trend, and their motivations are largely extrinsic, focusing on reputation, regulatory compliance, monetary incentives, and market competition. Practices within companies emphasize self-regulation, forming dedicated teams, outsourcing compliance, and conducting impact assessments with a utilitarian approach to maximize user protection and market benefits.

Despite mixed attitudes towards regulation, with some seeing it as a hindrance to innovation, others recognize its necessity for maintaining a good reputation and compliance. Companies acknowledge their responsibility to ensure transparency, accountability, and responsible commercialization of AI technologies. They strive for accuracy through continuous testing and development, reflecting a commitment to maintaining user trust and protecting their interests.

### **Developers**

Developers share some challenges with companies, such as the “black box” effect and vague guidelines, but they also face issues like making uneducated decisions and concerns about whistleblowing. Their attitudes reflect the view of AI as a market-driven trend, acknowledging the necessity of regulation despite its potential to hinder innovation. Developers bear significant responsibilities for educating themselves, fostering critical thinking, addressing transparency issues, and whistleblowing when necessary.

Developers' practices include breaking down code to enhance transparency, using AI for debugging, adhering to guidelines, and striving for accuracy through continuous testing and development. Motivations among developers are a mix of intrinsic factors, such as a moral



compass, and extrinsic pressures, including market demands and compliance requirements. This balance highlights the need for developers to critically evaluate their work and ensure it aligns with ethical standards while meeting market expectations.

## **Governments**

Governments encounter specific obstacles like competition disadvantages, overseeing major technology corporations, managing a rapidly changing sector, sluggish bureaucracy, and a shortage of agreement and technical expertise. Their outlook on AI includes both hope for making knowledge more accessible and doubt about the influence of persuasive AI technologies. Governments have the duty to oversee AI, safeguard users, and uphold ethical guidelines by enforcing rules such as the EU AI Act and GDPR, performing evaluations, and encouraging collaboration among stakeholders.

The main reasons governments act are external, as they aim to keep authority, safeguard citizens, and maintain public confidence. Highlighting the need for regulation underlines the significance of developing a fair strategy that promotes creativity while guaranteeing ethical behavior. Governments have a vital role in establishing the guidelines for ethical AI usage and ensuring the appropriate supervision to uphold these regulations.

## **Researchers**

Researchers encounter challenges such as the black box phenomenon and insufficient technical expertise in generative AI. They have conflicting views on regulation, seeing it as a barrier to innovation and essential for upholding ethical standards. Research shows that AI is viewed as a potent yet inherently biased tool that necessitates careful thought and ongoing testing to guarantee its ethical application.

Among researchers, common practices include ongoing experimentation, adherence to guidelines, and providing thorough ethical justifications for the use of AI. Motivations are a mix of internal factors, such as ethics and self-preservation, and external factors, such as rules and job demands. Researchers underline the importance of transparency, accountability, and continuous communication to tackle ethical issues and enhance understanding of AI.

## **Users**

End-users of AI systems have surfaced an essential group of stakeholders tasked with enhancing AI understanding, fostering a cautious attitude, and actively participating in ethical AI utilization and oversight. Difficulties faced by users encompass a lack of understanding of AI, over-reliance on technology, and uninformed choices. Their attitudes reflect concerns about the social impact of AI, but they continue engaging with the technologies motivated by the fear of missing out on information and experiences.

User rights consist of the right to receive information, the right to challenge AI decisions, the right to opt out of AI systems, and the freedom to choose how they utilize the technology.

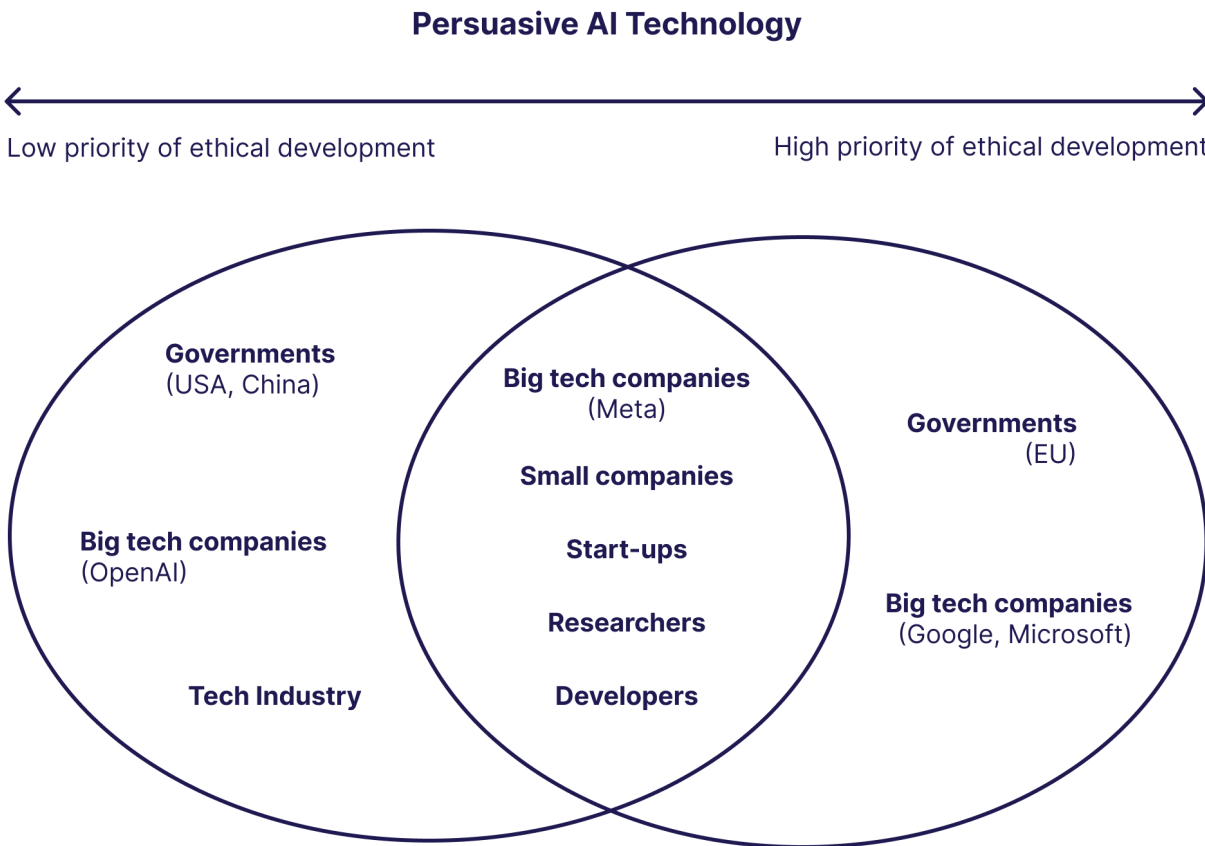
Furthermore, it is their duty to utilize technology responsibly and participate in discussions regarding regulation. To make sure AI is used ethically, end-users need to be knowledgeable and discerning about the technology they use, playing an active role in establishing ethical norms and procedures.

Summary of standpoints

Inspired by the interpretations of an interview participant, we summarized the standpoints of stakeholders in **Figure 20**, categorizing various stakeholders based on their prioritization of ethics in the use of AI technology. This diagram visualizes the overall landscape of stakeholder standpoints in relation to ethical AI development, concluding our investigation for RQ2.

Figure 20

*Categorization of stakeholder attitudes in relation to prioritizing ethics for persuasive AI technology*



On the left, indicating a lower priority for ethics, are stakeholders such as the governments of the USA and China, big tech companies like OpenAI, and the broader tech industry. These entities are positioned here, suggesting they place less emphasis on ethical considerations in their AI practices, possibly driven by competitive and market-driven motivations.

In the center, where priorities are mixed, are big tech companies like Meta, small companies, start-ups, researchers, and developers. These stakeholders demonstrate a more balanced approach, occasionally prioritizing ethical considerations while focusing on innovation and market competitiveness. On the right, indicating a high priority of ethics, are the EU governments and big tech companies like Google and Microsoft. These entities emphasize stronger adherence to ethical standards in their AI development and deployment, reflecting a commitment to maintaining ethical integrity in their operations.

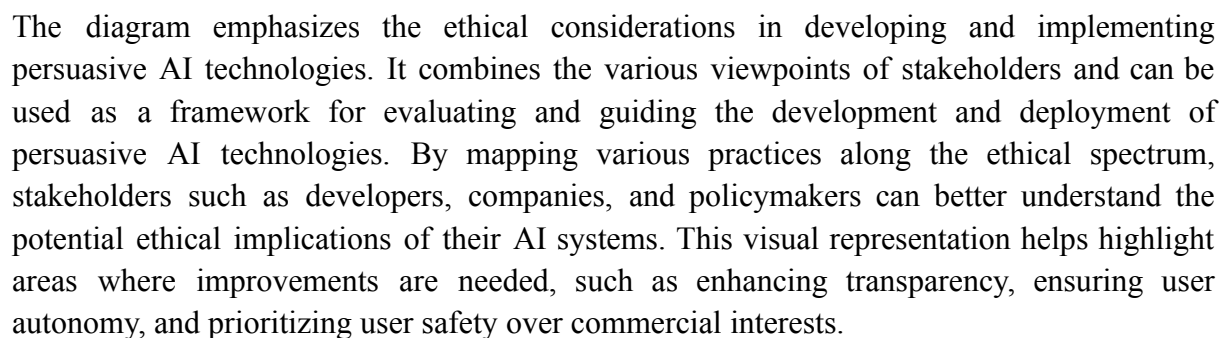
**Figure 20** highlights the diverse approaches stakeholders take toward ethics in persuasive AI technology, influenced by varying motivations and regulatory environments. The proposed diagram can be used as a tool to identify and analyze the ethical priorities of various stakeholders in the use of persuasive AI technology.

## 6.2. Ethical Use of Persuasive AI Technology

Similar to the diagram on stakeholder attitudes in relation to prioritizing ethics, we propose an evaluation of techniques and practices in persuasive AI technology. This is based on the literature review and the recommendations of the interview participants, as visualized in **Figure 21**. It depicts various practices that hinder or promote ethical development through a categorization of less ethical to more ethical. This diagram summarizes the different viewpoints of stakeholders in relation to what makes AI-driven persuasive technology ethical or less ethical based on current and recommended practices.

The left circle, representing the least ethical practices, includes activities like nudging, gamification, prioritizing commercialization over user safety, hidden terms and conditions, unjustified use of AI, data exploitation, and manipulation. The right circle, representing the most ethical practices, includes XAI (Explainable AI), boosting (as a more ethical alternative to nudging), labeling/watermarking, documentation, use of guidelines, impact assessment, justified use of AI and data, and compliance with regulations like GDPR and the EU AI Act. The overlapping section lists neutral or context-dependent practices such as persuasion, nudging, recommender systems, gamification, and generative AI (GenAI) as a new type of persuasive technology.

### *Categorization of ethical use of persuasive AI technology with key techniques and practices*



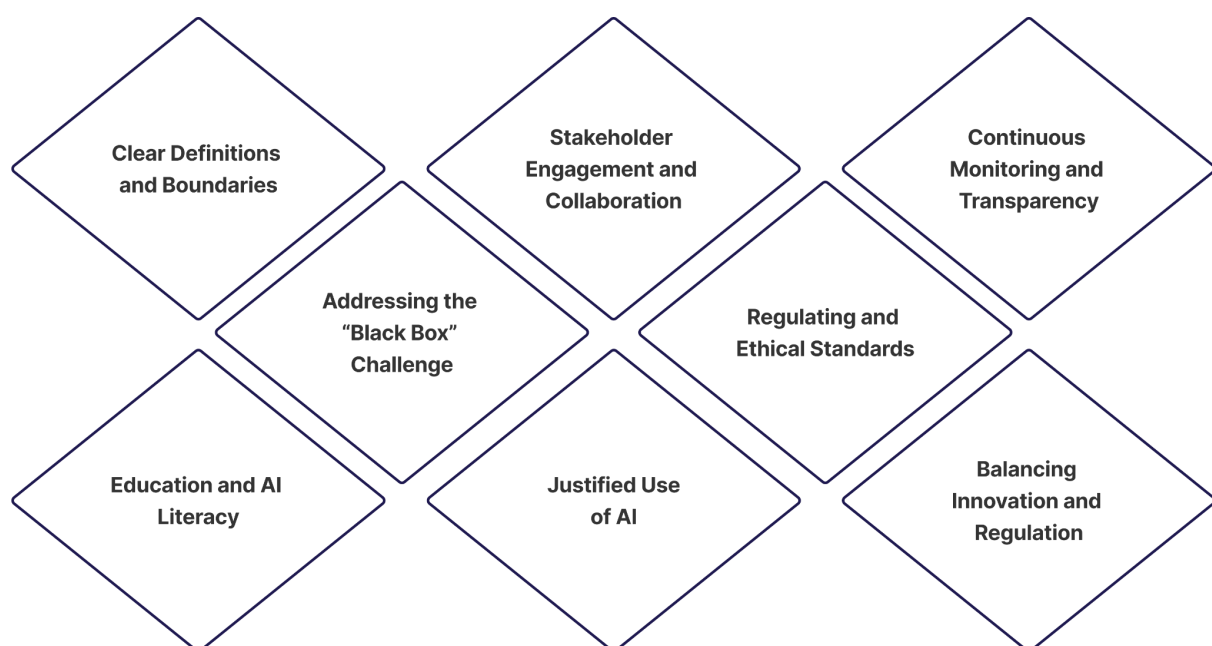
120

## 6.3. Focus Areas for Future Efforts

To answer RQ3, we propose eight focus areas for future efforts of stakeholders to ensure the ethical use of AI in persuasive technology, also shown in **Figure 22**: (1) clear definitions and boundaries, (2) stakeholder engagement and collaboration, (3) continuous monitoring and transparency, (4) addressing the “black box” challenge, (5) regulation and ethical standards, (6) education and AI literacy, (7) justified use of AI, and (8) balancing innovation and regulation. These have been developed based on our findings from the analysis and are central to developing a holistic approach to addressing the ethical challenges associated with persuasive AI. By understanding and addressing these areas, stakeholders can better ensure the ethical development and deployment of AI technologies.

**Figure 22**

*The eight focus areas for future efforts of stakeholders in ensuring the ethical use of AI in persuasive technology*



### Clear definitions and boundaries

A recurrent theme in the analysis was the need for clear definitions and boundaries for ethical principles such as beneficence and non-maleficence, fairness and justice, and human autonomy and agency. Participants emphasized the challenge of defining these principles and the compromises required when applying them. In upcoming endeavors, it is important for stakeholders, especially researchers and policymakers, to work together to create clear ethical guidelines. This cooperation can facilitate the establishment of a mutual comprehension and uniform implementation of ethical principles across various AI systems and situations.

Furthermore, determining the meaning of beneficence and non-maleficence involves a multidisciplinary method, incorporating perspectives from philosophy, sociology, and technology.

### **Stakeholder engagement and collaboration**

Engaging stakeholders is essential to ensure diverse perspectives are taken into account in the ethical advancement of AI. The study highlighted the significance of actively involving all interested parties, such as end-users, in discussions regarding ethical AI. In order to develop frameworks that meet the needs and values of various groups, companies, developers, governments, and researchers need to collaborate. This teamwork method can assist in detecting ethical concerns at an early stage of the development process and guarantee that AI systems prioritize the needs of end-users. Future focus should be on creating channels for consistent involvement and input from stakeholders.

### **Continuous monitoring and transparency**

Continuous monitoring of AI systems is essential to ensure they adhere to ethical standards over time. Participants stressed that monitoring and auditing contribute to explainability and transparency, supporting user rights to be informed and to contest AI decisions. Future efforts should focus on developing robust monitoring mechanisms that can track the performance and impact of AI systems. Transparency can be enhanced through practices such as documentation, labeling, and explainability measures. Companies and developers should implement transparent practices that allow users to understand how AI systems operate and make decisions. This includes using clear and accessible language to explain AI processes to non-technical users.

### **Addressing the "black box" challenge**

The lack of transparency and explainability in numerous AI models creates obstacles for accountability and oversight. Future attention should be directed towards creating techniques that increase the interpretability and explainability of AI systems. Researchers and developers need to dedicate resources to developing AI models that offer clear and significant explanations for their decisions. Methods such as post hoc explainability and simplifying complex models into smaller parts can aid in solving this issue. Furthermore, it is important to educate both developers and end-users about the limitations and operations of AI systems.

### **Regulation and ethical standards**

Regulation and ethical standards were a key focus of the analysis. Governments and regulators have a responsibility to create and implement rules that promote the ethical use of AI. The EU AI Act and GDPR were mentioned as instances of regulatory frameworks that can lead to ethical AI advancement. Future endeavors should prioritize improving these rules to tackle new ethical dilemmas and ensure they can adjust to evolving technological innovations. Governments need to address bureaucratic delays in order to keep up with the rapid advancements in AI technology.

## **Education and AI literacy**

Increasing understanding of AI among all parties is essential to promote ethical AI. Attendees observed that insufficient technical expertise can result in uninformed choices and difficulties in applying ethical guidelines. Future focus should be placed on education and training initiatives that improve comprehension of AI technologies and the ethical considerations involved. Companies and educational institutions need to work together to create curricula that address the technical, ethical, and societal dimensions of AI. Improving AI understanding can enable individuals to make educated choices and support the ethical advancement and utilization of AI.

## **Justified use of AI**

The participants shared the view that AI is often pursued as a trend, with companies and researchers frequently motivated to implement it simply to keep up. However, researchers and developers have stressed that the use of AI, especially large language models, is often unjustified and can negatively impact individuals and society. Risks associated with the use of persuasive AI technology include exploitation of personal data, impact on individuals' decision-making, and political polarization. Moreover, the development and deployment of AI require significant computational resources, leading to increased energy consumption and electronic waste, resulting in negative environmental effects.

Therefore, it is crucial that stakeholders, especially developers, conduct an impact assessment before developing such technologies and provide justifiable reasoning for using them. Furthermore, governments could also focus on setting boundaries for contexts in which the use of AI is justifiable.

## **Balancing innovation and regulation**

One main difficulty in ensuring ethical AI development is finding a balance between the importance of innovation and the essentiality of regulation. Participants were worried that too much regulation could hinder innovation, whereas too little regulation might result in ethical failures. Future endeavors should strive for a balance that enables creative advancements while guaranteeing ethical guidelines are upheld. Continual communication is needed among regulators, companies, and researchers in order to establish a regulatory framework that encourages ethical advancements.

To summarize, guaranteeing ethical AI in persuasive technology involves adopting a comprehensive strategy that includes defining boundaries, involving stakeholders, ongoing monitoring for transparency, tackling the “black box” issue, implementing regulations, educating, and finding a balance between innovation and regulation. By concentrating on these domains, stakeholders can strive to create AI systems that are efficient and also uphold societal values in an ethical manner.

## 7. Conclusion

After going through the various stages of this exploratory research, we are now able to address the problem statement presented in the introduction: What are the ethical implications of using AI to influence user behavior, and how can responsible stakeholders work towards ensuring the ethical use of AI in persuasive technology?

The review of related work provided a comprehensive understanding of the key ethical considerations associated with using AI to influence user behavior. Through a synthesis of the literature, we identified five key ethical principles: beneficence and non-maleficence, fairness and justice, human autonomy and agency, transparency and explainability, and accountability and oversight. Additionally, we identified persuasive AI technology as a primary theme in AI influencing user behavior and recognized companies, developers, governments, and researchers as the key responsible stakeholders in AI.

To investigate the viewpoints of AI stakeholders on current practices for addressing the ethical implications of AI-driven persuasive technology, we conducted a series of interviews with representatives from the four stakeholder groups. The analysis of these interviews revealed various themes, including the views of participants on the five ethical principles and the associated challenges and practices. Additionally, we conducted a comparative analysis of the different stakeholder groups, focusing on their challenges, attitudes and beliefs, motivations, responsibilities, and practices in relation to ethical AI in persuasive technology.

Based on the analysis, we identified eight focus areas for stakeholders in future efforts to ensure ethical AI in persuasive technology. These include clear definitions and boundaries, stakeholder engagement and collaboration, continuous monitoring and transparency, addressing the “black box” challenge, regulation and ethical standards, education and AI literacy, justified use of AI, and balancing innovation and regulation. By addressing these areas, stakeholders can better ensure the ethical development and deployment of AI technologies. Furthermore, we proposed an evaluation of techniques and practices in persuasive AI based on their ethical implications, as well as a spectrum of stakeholder attitudes in relation to prioritizing ethics for persuasive AI technology.

Overall, this study has emphasized the complex ethical landscape of AI-driven persuasive technology. It has underscored the need for a multifaceted approach involving all key stakeholders to navigate the global digital challenges and ensure that AI systems are designed and used ethically. Continued dialogue, research, and collaboration will be essential in advancing ethical AI practices, protecting user rights, and promoting trust in AI technologies.



## 8. Reflection and Future Work

This exploratory study has provided insights into the ethical implications of using AI to influence user behavior and how different stakeholders work to ensure ethical design in AI-driven systems. However, several methodological strengths and weaknesses must be acknowledged.

One of the primary strengths of this study is its qualitative nature, which allowed for rich, in-depth insights into the perspectives of various AI stakeholders. Embracing a constructionist view, we regarded the interview subjects as co-creators of knowledge, emphasizing the dynamic and collaborative nature of data production (Roulston, 2010, as cited in Brinkmann & Kvale, 2018). This approach facilitated a nuanced understanding of the complex ethical landscape surrounding AI technology.

Nevertheless, the qualitative nature also entails limitations in the generalizability of the findings. Since the study only included a small number of participants, they are not necessarily representative of the broader population of each stakeholder group. Additionally, the participants were all selected from Danish organizations or institutions, which limits the applicability of the findings in other contexts, such as global settings.

Furthermore, from an interpretivism perspective, the sample of the study presented some challenges and potential biases. Notably, three out of the four researchers interviewed had connections to the same university, potentially leading to a homogeneity of views. Additionally, there was an uneven number of participants from each stakeholder group, resulting in the perspective of researchers being overrepresented while the perspective of developers was underrepresented. Moreover, there was a lack of gender diversity, with only 2 out of the 9 participants being female. Hagendorff (2020) highlights the underrepresentation of women in the tech industry and AI ethics, a trend mirrored in our sample. This lack of diversity might have skewed the findings, limiting the representation of differing perspectives within our sample.

The interviews also faced practical challenges, such as internet connection issues, which occasionally disrupted the conversations. These disruptions may have affected the clarity of the recordings and, subsequently, the accuracy of the transcriptions. Despite efforts to ensure the reliability of the data, some nuances and details might have been lost or misinterpreted.

Given these limitations, future research should aim to address these gaps by employing different methodologies, expanding the sample size, and including a more diverse range of participants. Furthermore, we have presented focus areas for future efforts of stakeholders to ensure the ethical use of AI in persuasive technology, serving as suggestions for future research within this field. In addition to these, the field of ethical AI would benefit from further research on defining the ethical principles within specific contexts.

# References

- Banovic, N., Yang, Z., Ramesh, A., & Liu, A. (2023). *Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust*. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–17. <https://doi.org/10.1145/3579460>
- Bélisle-Pipon, J.-C., Monteferrante, E., Roy, M.-C., & Couture, V. (2023). *Artificial intelligence ethics has a black box problem*. AI & Society, 38(4), 1507–1522. <https://doi.org/10.1007/s00146-021-01380-0>
- Benner, D., Schöbel, S., Janson, A., & Leimeister, J. M. (2022). *How to Achieve Ethical Persuasive Design: A Review and Theoretical Propositions for Information Systems*. Association for Information Systems Transactions on Human-Computer Interaction, 14(4), 548–577. <https://doi.org/10.17705/1thci.00179>
- Bingley, W. J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S. A., Ko, R. K. L., Steffens, N., Wiles, J., & Worthy, P. (2023). *Where is the human in human-centered AI? Insights from developer priorities and user experiences*. Computers in Human Behavior, 141, 107617-. <https://doi.org/10.1016/j.chb.2022.107617>
- Botes, M. (2023). *Autonomy and the social dilemma of online manipulative behavior*. AI and Ethics (Online), 3(1), 315–323. <https://doi.org/10.1007/s43681-022-00157-5>
- Braun, V., & Clarke, V. (2012). *Thematic Analysis*. In H. Cooper (Ed.), APA handbook of research methods in psychology (pp. 57-71)
- Bremmer, I., & Suleyman, M. (2023). *The AI Power Paradox: Can States Learn to Govern Artificial Intelligence—Before It's Too Late?* Foreign Affairs. Retrieved May 24, 2024, from <https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox>
- Brinkmann, S., & Kvale, S. (2005). *Confronting the ethics of qualitative research*. Journal of Constructivist Psychology, 18(2), 157–181. <https://doi.org/10.1080/10720530590914789>
- Brinkmann, S., & Kvale, S. (2018). *Doing Interviews* (Second edition, Vol. 2). SAGE Publications, Limited. <https://doi.org/10.4135/9781529716665>
- Bryman, A., & Bryman, A. (2016). *Social research methods* (5. edition.). Oxford University Press.
- Buhmann, A., & Fieseler, C. (2021). *Towards a deliberative framework for responsible innovation in artificial intelligence*. Technology in Society, 64, 101475-. <https://doi.org/10.1016/j.techsoc.2020.101475>

- Burr, C., & Floridi, L. (2020). *Ethics of Digital Well-Being: A Multidisciplinary Approach* (1st Edition 2020, Vol. 140). Springer Nature.  
<https://doi.org/10.1007/978-3-030-50585-1>
- Burtell, M., & Woodside, T. (2023). *Artificial Influence: An Analysis Of AI-Driven Persuasion*. <https://doi.org/10.48550/arxiv.2303.08721>
- Chignell, M., Wang, L., Zare, A., & Li, J. (2023). *The Evolution of HCI and Human Factors: Integrating Human and Artificial Intelligence*. ACM Transactions on Computer-Human Interaction, 30(2), 1–30. <https://doi.org/10.1145/3557891>
- Clarke, V., & Braun, V. (2017). *Thematic analysis*. The Journal of Positive Psychology, 12(3), 297–298. <https://doi.org/10.1080/17439760.2016.1262613>
- Coeckelbergh, M. (2020). *AI ethics* (1st ed.). The MIT Press.  
<https://doi.org/10.7551/mitpress/12549.001.0001>
- Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & de Oliveira, N. (2023). *Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance*. Patterns (New York, N.Y.), 4(10), 100857–100857.  
<https://doi.org/10.1016/j.patter.2023.100857>
- Cronin, Ryan, F., & Coughlan, M. (2008). *Undertaking a literature review: a step-by-step approach*. British Journal of Nursing (Mark Allen Publishing), 17(1), 38–43.  
<https://doi.org/10.12968/bjon.2008.17.1.28059>
- Deci, E. L., & Ryan, R. M. (2000). *The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior*. Psychological Inquiry, 11(4), 227–268.  
[https://doi.org/10.1207/S15327965PLI1104\\_01](https://doi.org/10.1207/S15327965PLI1104_01)
- Dehnert, M., & Mongeau, P. A. (2022). *Persuasion in the Age of Artificial Intelligence (AI): Theories and Complications of AI-Based Persuasion*. Human Communication Research, 48(3), 386–403.
- Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). *Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation*. Information Fusion, 99, 101896–. <https://doi.org/10.1016/j.inffus.2023.101896>
- Dix, A., Finlay, J. E., Abowd, G. D., & Beale, R. (2004). *Human-computer interaction*. (3. ed.). Pearson Education.
- Duarte, E., & Baranauskas, M. (2016). *Revisiting the Three HCI Waves: A Preliminary Discussion on Philosophy of Science and Research Paradigms*. ACM International Conference Proceeding Series, 128046, 1–4.  
<https://doi.org/10.1145/3033701.3033740>
- Eitel-Porter, R. (2021). *Beyond the promise: implementing ethical AI*. AI and Ethics (Online), 1(1), 73–80. <https://doi.org/10.1007/s43681-020-00011-6>

- Endsley, M. R. (2023). *Ironies of artificial intelligence*. *Ergonomics*, 66(11), 1656–1668. <https://doi.org/10.1080/00140139.2023.2243404>
- Fogg, B. J. (2003). *Persuasive technology using computers to change what we think and do*. Morgan Kaufmann Publishers.
- Gao, B., Wang, Y., Xie, H., Hu, Y., & Hu, Y. (2023). *Artificial Intelligence in Advertising: Advancements, Challenges, and Ethical Considerations in Targeting, Personalization, Content Creation, and Ad Optimization*. *SAGE Open*, 13(4). <https://doi.org/10.1177/21582440231210759>
- Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). *Trust in Artificial Intelligence: A Global Study*. The University of Queensland and KPMG Australia. <https://doi.org/10.14264/00d3c94>
- Gutierrez, C. I. (2023). *Uncovering Incentives for Implementing AI Governance Programs: Evidence From the Field*. *IEEE Transactions on Artificial Intelligence*, 4(4), 792–798. <https://doi.org/10.1109/TAI.2022.3171748>
- Hagendorff, T. (2020). *The Ethics of AI Ethics: An Evaluation of Guidelines*. *Minds and Machines* (Dordrecht), 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Hindocha, S., & Badea, C. (2022). *Moral exemplars for the virtuous machine: the clinician's role in ethical artificial intelligence for healthcare*. *AI and Ethics* (Online), 2(1), 167–175. <https://doi.org/10.1007/s43681-021-00089-6>
- Hoxhaj, O., Halilaj, B., & Harizi, A. (2023). *Ethical Implications and Human Rights Violations in the Age of Artificial Intelligence*. *Balkan Social Science Review*, 22(22), 153–171. <https://doi.org/10.46763/BSSR232222153h>
- Ibáñez, J. C., & Olmeda, M. V. (2022). *Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study*. *AI & Society*, 37(4), 1663–1687. <https://doi.org/10.1007/s00146-021-01267-0>
- Klenk, M. (2020). *Digital Well-Being and Manipulation Online*. In Burr, C., & Floridi, L. (2020). *Ethics of Digital Well-Being* (Vol. 140, pp. 81–100). Springer International Publishing AG. [https://doi.org/10.1007/978-3-030-50585-1\\_4](https://doi.org/10.1007/978-3-030-50585-1_4)
- Koniakou, V. (2023). *From the “rush to ethics” to the “race for governance” in Artificial Intelligence*. *Information Systems Frontiers*, 25(1), 71–102. <https://doi.org/10.1007/s10796-022-10300-6>
- Kreps, S., George, J., Lushenko, P., & Rao, A. (2023). *Exploring the artificial intelligence “Trust paradox”: Evidence from a survey experiment in the United States*. *PloS One*, 18(7), e0288109–e0288109. <https://doi.org/10.1371/journal.pone.0288109>
- Laato, S., Tiainen, M., Najmul Islam, A. K. M., & Mäntymäki, M. (2022). *How to explain AI systems to end users: a systematic literature review and research agenda*. *Internet Research*, 32(7), 1–31. <https://doi.org/10.1108/INTR-08-2021-0600>

- Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction* (Second edition). Morgan Kaufmann Publishers, an imprint of Elsevier.
- Li, Y., & Hilliges, O. (Eds.). (2021). *Artificial intelligence for human computer interaction : a modern approach*. Springer (pp. 5-16)
- Lorente, A. (2022). *Setting the goals for ethical, unbiased, and fair AI*. In *AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI* (pp. 13–54). <https://doi.org/10.1016/B978-0-32-391919-7.00014-7>
- Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). *The potential of generative AI for personalized persuasion at scale*. *Scientific Reports* (Nature Publisher Group), 14(1), 4692. <https://doi.org/10.1038/s41598-024-53755-0>
- Ministry of Foreign Affairs of Denmark (2021). *Strategy for Denmark's Tech Diplomacy 2021-2023*. Office of Denmark's Tech Ambassador. Retrieved May 24, 2024, from <https://techamb.um.dk/-/media/country-sites/techamb-en/media/strategy-for-denmarks-tech-diplomacy-2021-2023.ashx>
- Ministry of Foreign Affairs of Denmark (2024). *The Ministry of Foreign Affairs Strategy for Tech Diplomacy*. Office of Denmark's Tech Ambassador. Retrieved May 24, 2024, from <https://techamb.um.dk/strategy>
- Mittelstadt, B. (2019). *Principles alone cannot guarantee ethical AI*. *Nature Machine Intelligence*, 1(11), 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Mougan, C., & Brand, J. (2023). *Kantian Deontology Meets AI Alignment: Towards Morally Grounded Fairness Metrics*. <https://doi.org/10.48550/arxiv.2311.05227>
- Müller, Vincent C. (2022), *The history of digital ethics*, in Carissa Véliz (ed.), *Oxford handbook of digital ethics* (Oxford: Oxford University Press).
- Oinas-Kukkonen, H., & Harjumaa, M. (2009). *Persuasive Systems Design: Key Issues, Process Model, and System Features*. *Communications of the Association for Information Systems*, 24, 485-500. <https://doi.org/10.17705/1CAIS.02428>
- Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotko, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., ... Xu, W. (2023). *Six Human-Centered Artificial Intelligence Grand Challenges*. *International Journal of Human-Computer Interaction*, 39(3), 391–437. <https://doi.org/10.1080/10447318.2022.2153320>
- Palladino, N. (2023). *A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices*. *Telecommunications Policy*, 47(5), 102479-. <https://doi.org/10.1016/j.telpol.2022.102479>

- Panchal, M. H., & Panchal, S. D. (2023). *Challenges and future work directions in artificial intelligence with human-computer interaction*. In Scopus. Elsevier Inc. <https://doi.org/10.1016/B978-0-323-99891-8.00006-1>
- Pickard, A. J. (2013). *Research Methods in Information* (pp. 5-24). Facet.
- Prabhakaran, V., Mitchell, M., Gebru, T., & Iason Gabriel. (2022). *A Human Rights-Based Approach to Responsible AI*. arXiv.Org. <https://doi.org/10.48550/arxiv.2210.02667>
- Richards, K. A. R., & Hemphill, M. A. (2018 ). *A Practical Guide to Collaborative Qualitative Data Analysis*. Journal of Teaching in Physical Education. Retrieved May 21, 2024, from <https://doi.org/10.1123/jtpe.2017-0084>
- Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of Behavioral Research: Methods and Data Analysis* (pp. 1-34). McGraw Hill.
- Rowley, & Slack, F. (2004). *Conducting a literature review*. Management Research News, 27(6), 31–39. <https://doi.org/10.1108/01409170410784185>
- Ryan, M. (2020). *In AI We Trust: Ethics, Artificial Intelligence, and Reliability*. Science and Engineering Ethics, 26(5), 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>
- Ryan, R. M., & Deci, E. L. (2000). *Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions*. Contemporary Educational Psychology, 25(1), 54–67. <https://doi.org/10.1006/ceps.1999.1020>
- Sadeghian, A. H., & Otarkhani, A. (2023). *Data-driven digital nudging: a systematic literature review and future agenda*. Behaviour & Information Technology, 1–29. <https://doi.org/10.1080/0144929X.2023.2286535>
- Sharp, H., Rogers, Y., & Preece, J. (2019). *Interaction design : beyond human-computer interaction* (Fifth edition). John Wiley & Sons.
- Shin, D. (2023). *Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm*. Journal of Information Science, 49(1), 18–31. <https://doi.org/10.1177/0165551520985495>
- Shneiderman, B. (2020a). *Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems*. ACM Transactions on Interactive Intelligent Systems, 10(4), 1–31. <https://doi.org/10.1145/3419764>
- Shneiderman, B. (2020b). *Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy*. International Journal of Human-Computer Interaction, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Sigfrids, A., Leikas, J., Salo-Pöntinen, H., & Koskimies, E. (2023). *Human-centricity in AI governance: A systemic approach*. Frontiers in Artificial Intelligence, 6, 976887–976887. <https://doi.org/10.3389/frai.2023.976887>



- Smuha, N. A. (2021). *From a “race to AI” to a “race to AI regulation” : regulatory competition for artificial intelligence*. *Law, Innovation and Technology*, 13(1), 57–84. <https://doi.org/10.1080/17579961.2021.1898300>
- Stahl B. C. (2021). *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. *Concepts of Ethics and Their Application to AI* 19–33. [https://doi.org/10.1007/978-3-030-69978-9\\_3](https://doi.org/10.1007/978-3-030-69978-9_3)
- TED (2024, April 22). *What Is an AI Anyway? | Mustafa Suleyman | TED* [Video]. YouTube. [https://www.youtube.com/watch?v=KKNCiRWd\\_j0&t=18s&ab\\_channel=TED](https://www.youtube.com/watch?v=KKNCiRWd_j0&t=18s&ab_channel=TED)
- Thiebes, S., Lins, S., & Sunyaev, A. (2021). *Trustworthy artificial intelligence*. *Electronic Markets*, 31(2), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- Todorova, C., Sharkov, G., Aldewereld, H., Leijnen, S., Dehghani, A., Marrone, S., Sansone, C., Lynch, M., Pugh, J., Singh, T., Mezei, K., Antal, P., Hanák, P., Barducci, A., Perez-Tellez, F., & Gargiulo, F. (2023). *The European AI Tango: Balancing Regulation Innovation and Competitiveness*. *ACM International Conference Proceeding Series*, 2–8. <https://doi.org/10.1145/3633083.3633161>
- Toloka Team. (2023, Jun 26.). *The history, timeline, and future of LLMs*. Toloka. Retrieved 30 May, 2024, from <https://toloka.ai/blog/history-of-llms/>
- Torning, K., & Oinas-Kukkonen, H. (2009). *Persuasive system design: state of the art and future directions*. *ACM International Conference Proceeding Series*, 350, 1–8. <https://doi.org/10.1145/1541948.1541989>
- Vallor, S. (2016). *Technology and the virtues : a philosophical guide to a future worth wanting*. Oxford University Press.
- Wang, G., Liu, X., Wang, Z., & Yang, X. (2021). *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 762–766). ACM. <https://doi.org/10.1145/3443467.3443850>
- Warwick, K. (2016). *Homo Technologicus: Threat or Opportunity?* *Philosophies* (Basel), 1(3), 199–208. <https://doi.org/10.3390/philosophies1030199>
- Wenar, L. (2023) *Rights*. *Stanford Encyclopedia of Philosophy*. Retrieved May 24, 2024, from <https://plato.stanford.edu/archives/spr2023/entries/rights/>
- Wulf, A., J., & Seizov, O. (2020). *Artificial Intelligence and Transparency: A Blueprint for Improving the Regulation of AI Applications in the EU*. *European Business Law Review*, 31(Issue 4), 611–640. <https://doi.org/10.54648/EULR2020024>

# Appendixes

(Attached separately)

Appendix A - Literature Search

Appendix B - Interview Guide

Appendix C - Coded transcript for Company 1 (C1)

Appendix D - Coded transcript for Company 2 (C2)

Appendix E - Coded transcript for Developer 1 (D1)

Appendix F - Coded transcript for Government 1 (G1)

Appendix G - Coded transcript for Government 2 (G2)

Appendix H - Coded transcript for Researcher 1 (R1)

Appendix I - Coded transcript for Researcher 2 (R2)

Appendix J - Coded transcript for Researcher 3 (R3)

Appendix K - Coded transcript for Researcher 4 (R4)

Appendix L - Literature list and supervisor approval