TU WIEN Informatics

Lowering the Threshold for Embedded AI Ethics

Designing Self-Assessment Tools for Stakeholder-Identified Ethical Concerns

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Tobias Hercules Christoph, BSc.
Matrikelnummer 11703823

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ass. Prof. Dr.techn. Cornelis van Berkel, MSc
Mitwirkung: Ass. Prof. Dr.x techn. Katta Spiel, BSc B.A. MSc

Wien, 3. Dezember 2025

Tobias Hercules Christoph	Cornelis van Berkel

TU WIEN Informatics

Lowering the Threshold for Embedded AI Ethics

Designing Self-Assessment Tools for Stakeholder-Identified Ethical Concerns

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Tobias Hercules Christoph, BSc.
Registration Number 11703823

to the Faculty of Informatics

at the TU Wien

Advisor: Ass. Prof. Dr.techn. Cornelis van Berkel, MSc
Assistance: Ass. Prof. Dr.x techn. Katta Spiel, BSc B.A. MSc

Vienna, December 3, 2025

Tobias Hercules Christoph	Cornelis van Berkel

Erklärung zur Verfassung der Arbeit

Tobias Hercules Christoph, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 3. Dezember 2025

Tobias Hercules Christoph

Acknowledgements

I would like to thank the cake meetings and thesis group meetings, and everyone who took part in them during the time I was writing this thesis. Thank you for listening to my (thesis-related or not) struggles, for sharing your own (thesis-related or not) challenges, and for reminding me that having difficulties is something that everyone can relate to.

I also want to thank my friends, who support me no matter what.

And finally, I want to thank the two best advisors anyone could ask for.

Kurzfassung

Technologien die unethisch sind, diskriminieren oder Schaden verursachen (etwa die Privatsphäre verletzen), resultieren häufig aus Entwickler:innen, die Schwierigkeiten haben, ethische Anforderungen im Entwicklungsprozess zu integrieren, zu definieren und zu bewerten. Immer schneller werdende technologische Entwicklungen, der allgegenwärtige Einsatz von KI sowie die unterschiedlichen Interessen aller komplexen Stakeholder verstärken diese Problematik. Diese Arbeit untersucht daher zentrale Fragen zu ethischen Audits von KI-Technologien: Welche Werte und ethischen Risiken sind relevant? Und wie lassen sie sich in eine prüfbare Form verwandeln? Im Mittelpunkt der Arbeit steht insbesondere die Rolle unterschiedlicher Stakeholder bei der Identifikation dieser Bedürfnisse. Des Weiteren werden Tools entwickelt, welche das Erstellen von auditierbaren Artefakten (den Ethical Focus Areas) unterstützen. Der resultierende Workflow sind die Ethics Self-Assessment Tools (ESAT).

Die Arbeit verfolgt einen Mixed-Methods-Ansatz, um die Literatur zu diesen Problemen kritisch zu analysieren, aus bestehenden Ansätzen Lücken herauszuarbeiten sowie einen partizipativen Workflow zum Erheben von ethischen Anliegen zu entwickeln. Durch erste Workshops entsteht eine zugängliche Methode, die ethische Fragestellungen als strukturierte Daten generiert. Ergänzend wird ein einfach nutzbares Webtool entwickelt, welches die Daten in testbare Artefakte verarbeitet. Das ermöglicht es Entwicklungsteams ohne Ethik-Expertise, einen Zugang zu systematischen Bewertungen ihrer Technologien zu erhalten. Fokus dieser Arbeit sind kontextabhängige Technologien wie KI-Beratungsagenten (zum Beispiel auf Large Language Models (LLM) basierende Chatbots). Ebenso wird über die Übertragbarkeit des ESAT-Workflows auf andere Systeme reflektiert.

Die Ergebnisse zeigen darauf hin, dass ein Ansatz, welcher auf den Bedürfnissen von Stakeholdern basiert, kontinuierliche, kontextsensitive Bewertungen von KI-Systemen aufgrund ethischer Aspekte ermöglicht und sich zu einem generellen Audit-Toolkit erweitern lässt.

.

Abstract

Unethical technology, such as systems that discriminate, exploit, cause harm, or violate privacy, often arise from developers' struggles to define, assess and implement ethics during development. The increasing pace of technological advancement, the ubiquitous deployment of AI, and the complexity of shareholder considerations add to this challenge. This thesis seeks to investigate key questions about ethical technology audits: what are the relevant values and ethical concerns involved? And how can these be translated into a testable environment? In particular, it explores the role of stakeholders in identifying relevant values and creates tools that assist in translating them into the auditable artefacts of Ethical Focus Areas (EFAs). The resulting workflow are the Ethics Self-Assessment Tools (ESAT).

A mixed-methods approach is employed throughout the thesis to critically analyse the problem, current approaches and their gaps, and to develop a participatory workflow for eliciting values and concerns. Through trial workshops, a streamlined and approachable method that generates ethical concerns as data is established. Additionally, a straightforward webtool to process and translate this data into testable artefacts is developed, making the ethics assessment more accessible to non-ethical-expert developers. Context-sensitive applications, such as LLM-based advice chatbots, are the focus of this thesis. However, the transferability to other systems is also discussed.

The results indicate that by rooting the concerns directly in stakeholders' desires and expectations, the ESAT workflow builds towards a comprehensive toolkit that integrates auditing procedures early on in an AI's development cycle, enabling a continuous, context-sensitive, and adaptable assessment of a technology's ethical concerns.

Contents

CHAPTER 1

Introduction

While digital technologies appear more in all facets of our lives, their ethical and functional assessment is essential to ensure they operate safely and as intended. Especially due to the rise of Artificial Intelligence (AI) technologies, measures to assure they follow regulations and operate within safe boundaries are required. Even when driven by Silicon Valley-style claims of 'doing good' [39], their broader social impacts are often falling short of consideration by those who build and deploy them.

Many factors contribute to whether technologies work as intended. Their development and the evaluation of their consequences are for instance shaped by financial incentives. Considerations of what parts, how, and by whom technologies should be assessed further influence their design and development. Formal evaluations in the form of audits aim to assist here but are often regarded as obstacles to efficient development and innovation [60]. On top of all that, ensuring that a technology merely 'works as intended' may be deemed an insufficient goal, considering the wide range of societal effects of, for example, general-purpose AI (spreading misinformation [100] or affecting education [171]). Therefore, if the aim is to 'do good', technologies and their impact should also be evaluated along an ethical dimension.

1.1 Motivation

Fundamentally, the motivation for this thesis comes from the observation that many AI technologies are deployed in ways that are not sufficiently reliable or sound [116]. A range of ethical concerns exemplify this:

As many AI systems, especially popular language models, operate as "black boxes", their internal decisions are difficult or impossible to comprehend [83]. This lack of transparency limits users, regulators, or even developers themselves in predicting and correcting potentially harmful outcomes [100]. If not even the creators of a system

are fully able to understand their technology, accountability becomes difficult to assign [18], as it remains unclear who is fully responsible for errors, biased decisions, and consequent harms. A related concern is explainability, which, on top of whether a system can be inspected, refers to whether its reasoning can be articulated in a way that is understandable and meaningful to users, those affected by its decisions, or the ones who assess it [83]. While developers usually provide technical descriptions of how models operate, these explanations fall short in telling users *why* a specific decision was made about an individual. Without accessible explanations, people neither understand an AI's decisions, nor what would need to change for a different outcome [99].

Ethical worries also extend to privacy and user autonomy. Many AI systems rely on extensive collection of data. Individuals often are not fully aware or do not explicitly consent to their data being collected. This personal data can even reveal sensitive information once an AI system is in operation [83]. At the same time, the systems influence what users see and the choices they are offered, thereby influencing the decision-making by subtly *nudging* users towards a certain option [141]. This raises questions about how much control the individuals actually have over their data and the choices they take when interacting with AI technologies.

A further concern is bias in algorithmic systems that often reproduce or amplify inequalities that are reflected in their training data, which leads to systematically different outcomes for different groups [92]. Even if they perform "as intended", AI often inherits and magnifies social hierarchies. Attempts to combat this through fairness and non-discrimination frameworks often focus on preventing discrete instances of bias [7]. They fail to address structural conditions that produce the systematic disadvantage in the first place. This risks that fairness interventions leave the underlying injustices untouched [75]. While model performance can worsen ethical issues, these issues ultimately arise from the society the technologies are embedded in.

Beyond these general concerns, LLMs introduce additional challenges that intensify the need for ethical assessment. Their behaviours are not fully deterministic by explicit programming, resulting in potentially harmful outputs like biased outputs surfacing only in use or over time [126]. One-off and static evaluations are therefore insufficient. As LLMs increasingly appear in interactive systems as advisory chatbots, their ethical impacts go beyond technical performance to how they engage with users in real contexts [141]. This highlights the need for assessment approaches that continuously include people affected by the systems [86] as the ones that can elicit harms that benchmark metrics alone fail to capture [126].

Together, these concerns show the need for ethically grounded assessments of AI systems, but that the barrier in engaging with ethics is high. In this thesis, I intend to contribute to research that helps ensure such technologies are kept in check in ways that lower the threshold to engage in ethics to ensure they operate more predictably and cause less harm.

1.1.1 Problem Statement

A specific incident illustrating ethical shortcomings happened in January 2024 when the Austrian *Public Employment Service* (*Arbeitsmarktservice - AMS*) launched a career guidance chatbot, the *Berufsinfomat* [97]. The AMS, among other responsibilities, supports job seekers with information about potential career opportunities and provides job listings. The tasks of the chatbot were to assist in those functions, but for several reasons (technical issues, a limited knowledge base, and Large Language Model (LLM)-related hallucinations)[25] it failed to work ethically [47].

In general, LLM technologies are increasingly deployed in such advice-giving roles, often incorporated as domain-specific chatbots [32]. By providing advice, these systems shape decisions and directly influence the people they interact with. The result in the case of the Berufsinfomat was a chatbot that sometimes refused to answer, was not always reachable, provided false information, and was discriminatory in its advice [139]. The chatbot seemingly launched without much in-depth assessments; over the first few days, media outlets reported that users tested it with trivial prompts which uncovered many questionable answers [97]. For example, when asking identical questions about possible career options and merely changing the gender of the person in question, the chatbot provided vastly different results [47]. Querying for job options after graduating high school as a women, it would suggest secretarial office jobs or care-work-related careers [25]. These were not provided when specifying jobs for men. In that case, it would respond with listing software engineering as the first option.

The fact that many such examples surfaced immediately after its launch raises the question if the AMS and its developing partners tested their tool thoroughly. It suggests that some rudimentary questioning would have revealed potential concerns that needed the system to be reworked, which could have resulted in a chatbot that works better. The media backlash, and more importantly, a technology that discriminates, therefore seem preventable. This public backlash caused the AMS to respond [168]. While they did excuse the chatbot's behaviour by pointing to LLM-specific limitations and its model training on real-world data, they also ordered their developing partners to apply measures like filtering intended to make the chatbot *less* discriminatory.

According to the AMS chatbot developers [139], the criticism directed at the system was justified, although they viewed their own responsibility within the development, arguing that all AI chatbots would reproduce biases present in their data. Admittedly, if the objective was simply to create a chatbot that provides answers to questions about work information (not regarding whether these answers are incorrect or discriminatory), then the AMS succeeded and created a system that *works*. But the development process is influenced by broader constraints: while developers generally aim to create ethically sound systems [33], they often face structural barriers in doing so [128]. These include narrowly defined contractual obligations, pressure to deliver within strict timelines, limited expertise in ethical impact assessment, and a lack of time or resources for comprehensive testing [61].

Given the numerous reports on flawed AI with seemingly preventable issues [116], one might be inclined to rethink what constitutes 'working' technology. Does working mean that a technology's technical requirements are satisfied? But should systems still be considered functional if discriminatory issues arise, especially if these can be detected immediately after the product's launch?

To combat potential failures, especially when the standards of functionality are higher, technologies are assessed and tested, and typically once they are deployed, *audited*. Technical performance, such as a system's reliability or task-specific competence, can be tested rather clearly through benchmarks and evaluations [105]. The ethical behaviour of these technologies on the other hand, is more difficult to assess [124].

Ethics auditing in particular suffers from unclear goals and a lack of practically applicable methods to test systems. A more thorough description can be found in Sections 2.2 and 2.4, but to summarise the problems in brief, assessing ethics in AI is a challenging topic for several reasons:

- Ethical principles are vague and context-dependent [83]. Their concepts and values like fairness or transparency are often interpreted differently depending on who defines them and in what (e.g. cultural or domain) context. This makes it naturally difficult to transform them into concrete functional requirements of a technology.
- Ethical considerations are difficult to operationalise [124]. Ethical assessments are frequently disregarded during the development or early in product's lifecycle because the existing frameworks are perceived as too abstract to guide practical development.
- Ethics is often framed as a barrier rather than an enabler [60]. In industry settings, ethical considerations are regularly viewed as obstacles to innovation or sources of additional cost [83]. As a result, ethical issues tend to be addressed only as an afterthought – if at all.
- Effective assessments require collaboration with developers [126]. Evaluations cannot be carried out solely from the outside as they depend on access to internal knowledge, system design decisions, and implementation details. At the same time, they require a degree of independence [124] to ensure critical oversight.

The current auditing practices reflect the ethics assessment issues outlined above: audits are typically conducted post-development [126, 153] and framed as a kind of test a system has to pass [124]. Regulations, while important, can reinforce this notion of ethics as a hurdle to overcome [103]. Furthermore, audits are typically conducted by external experts with the aim of being impartial [29]. However, this can result in limited access to key aspects of a technology [124, 126], especially in the case of AI models. A collaborative approach between providers and auditors is therefore crucial, but auditing is still usually conducted with little involvement from the people the technology is designed for [123].

This disconnect can lead to audits missing relevant ethical concerns or failing to reflect the lived realities of users, potentially leading to reputational risks or public backlashes [139]. It consequently raises the issue: whose values are being built into these systems? This culminates in the broader question of *what* values should be considered?

Another challenge lies in embedding ethics in an impactful way throughout the development cycle. While developers often care about the ethical implications of their work [33], they struggle to incorporate ethical thinking into their development workflows [143]. Even when provided with a set of values or principles to follow, or an ethical framework [89], the question of *how* to test these continuously, remains. These challenges point to the broader need for approaches and methods that lower the threshold for developers to engage in ethics assessments.

Consequently, I identify two fundamental problems of AI ethics assessment:

1. The **What** problem: What values should be tested for? What areas of ethical focus matter in a specific context?

2. The **How** problem: How can these values be transformed into testable artefacts? How can the evaluation be designed in an approachable way and applied to ensure continuously repeatable audits?

To address these two problems, I propose an approach that supports developers in integrating ethics, incorporates stakeholder perspectives to define relevant values, and establishes continuously testable criteria for assessment. The following section outlines the goals of this approach in more detail.

1.2 Aim of the Thesis

The outlined problems above highlight the need for practicable ethics assessment approaches usable by developers and the need for a shift from external post-hoc checks towards internal, embedded auditability, where ethics assessment are part of the development process. To achieve this, the research in this thesis investigates current approaches to ethics assessment of AI technologies and seeks to contribute by constructing *Know-How* that can be applied directly. It moves beyond the existing plethora of research and *Know-What* guidance (frameworks, value lists, and high-level principles). The goal is to create a method to go from the abstract to the tangible; from broad principles and good intentions to something concrete and applicable in practice. This means shifting from situations where technology is developed with predominantly just a technical plan and minimal ethical considerations towards one where ethics is embedded in the technology itself. Within the scope of the research, the thesis focuses on advice-giving chatbots.

In particular, **this thesis seeks to address the following goal:**

Develop usable tools that serve as a proof-of-concept implementation of a stakeholder-involved method, enabling developers to continuously self-assess their domain-specific LLM advice chatbot systems. While these tools function as an illustrative application, the underlying approach is designed to transfer beyond this specific case.

Existing approaches to AI ethics evaluation often lack the mechanisms for meaningful assessments [10, 128] as they rely on specialised ethics expertise [61], struggle to enable participatory involvement [18, 86, 153], and offer limited guidance on translating values to testable criteria [121, 138]. The envisioned tools therefore have the following goals:

1. They should allow stakeholders to elicit ethically relevant values dependent on the context of the application through participatory workshops.
2. The tools' materials are directly applicable without requiring or becoming a professional in value-based design and ethical theory by providing clear implementation guidance through a script.
3. They allow an easy translation of raw data and identified values into structured data as concrete testable auditing criteria.

Limitations: This thesis and its resulting tools and materials are not a definitive solution to the problem of unethical technologies. Rather, this research seeks to provide practical guidance that supports and enables more responsible development and assists in identifying ethically relevant issues (ethics in the scope of this thesis is further discussed in Section 2.2). The tool represents ongoing work towards a more developed AI auditing toolkit. It, and any similar toolkit or framework, should not be considered an attempt to operationalise ethics or suggest that ethics can be reduced to merely following a checklist.

1.2.1 Research Question

The challenges and goals imply the core research question that guides this thesis. Sub-questions are identified to narrow the focus of the research:

RQ What tools can developers employ to effectively translate stakeholders' ethical values and expectations towards an AI technology into testable items?

- **RQa** How can a workshop and a webtool that processes the workshop data be designed to empower developers to facilitate ethics throughout their AI development process, minimising the need for them to become or employ ethical experts?
- **RQb** To what extent can the designed materials be transferred to other AI systems or domains which require ethical assessment?

1.3 Contributions

The thesis provides two primary contributions. First, it outlines the conceptual idea of the meta-method that argues for lowering the threshold for ethical self-assessment in AI development. At its core is the use of participatory practices and script-based, readily applicable materials that also enable reflexive engagement with the ethical concerns that arise. This approach addresses the identified gaps in the ethics assessment literature between high-level principles and practical evaluation methods. It enables developers to collaboratively involve stakeholders in identifying context-specific values and ethical concerns, and it translates these into auditable structures by proposing **Ethical Focus Areas (EFAs)** as the central conceptual artefacts.

Second, the thesis provides a practical contribution through the development of the **Ethics Self-Assessment Tools (ESAT)** as a proof-of-concept implementation of the meta-method. ESAT consists of:

(1) a participatory workshop format and its materials,
(2) a webtool that assists in transforming workshop output data into structured EFAs,
(3) and a supporting script that guides the user through the entire process.

Workshops were carried out and their data were processed using the tools, demonstrating the feasibility of the approach. ESAT is further conceptualised as an initial workflow toward a more comprehensive and marketable ethical auditing toolkit (Section 7.1.1), and its broader applicability beyond advice-giving LLM chatbots is discussed in Section 6.2.1.

1.4 Scope & Structure

The focus of this thesis is on the development and investigation of a meta-method that lowers the threshold for self-auditing in development processes. A full-scale validation through deploying it in a potential industry setting, e.g. through a case study, is beyond the scope of this thesis. Nonetheless, it describes the structure of such a study, potentially as a subsequent research project, in Section 7.1.2.

The research development process is situated within a Western (European) ethical context, where existing regulatory and value frameworks inform some of the workshop materials (e.g. the selection of values on value cards). As explained in the following, the ESAT workflow is developed with chatbots as the example technology to ground and test the conceptual foundation of the research in a use case.

1.4.1 Why Chatbots?

Practical and conceptual considerations have motivated the decision to focus on advice-giving chatbots in this research: practically, these systems are widely deployed across

diverse sectors. Examples can be found in career counselling [77], customer service [51], and healthcare [14, 107]. This makes them highly relevant AI systems for studying the considerations that users face. While the domain is varied, the technical specifics of chatbots remain sufficiently concrete to develop a structured and actionable tool, without risking overgeneralisation. Their context-sensitive behaviour makes them an ideal testbed for use-based stakeholder auditing approaches. In general, while much of the recent auditing literature focuses on general-purpose LLMs, domain-specific, use-oriented systems like advice-giving chatbots remain underexplored.

Conceptually, advice-giving chatbots feature some characteristics that make them suitable for ethical assessments. They are interactive, adaptive, and normative systems; their outputs directly influence users and their environments and by naturally involving guidance, the act of giving advice is inherently ethically charged [132]. Additional background information on chatbots and further technical details are described in Section 2.1.

1.4.2 Structure of the Thesis

The thesis is structured further into six chapters:

Chapter 2: Background introduces the foundational literature on chatbots, ethics, regulations, and development-centred approaches. It provides an overview of auditing in the academic literature and briefly introduces its approaches, namely internal or external, ethical, and participatory auditing. This chapter situates this literature as a the backbone of evaluating AI technologies in this thesis.

Chapter 3: Methodological Overview outlines the research approach to develop the ESAT workflow and describes the applied methods and phases of the development.

Chapter 4: Insights from Literature and Practice presents insights from critically analysing existing auditing methods and conversations with experts to identify clear gaps. Here, I define Ethical Focus Areas (EFAs) as the conceptual core of the ESAT tools.

Chapter 5: Designing the Toolkit presents the design process of the components of the ESAT workflow. In this chapter, I outline the planning and first versions, to playtesting and adjustments, to the finalisation of the tool. Decisions behind the changes I made to the workflow are also discussed here and the workshop materials and the web-based tool are described.

Chapter 6: Results and Discussion describes the findings from the tool's development and illustrates their intended use by showing example data. Here, the research questions are revisited by discussing how ESAT addresses them. This chapter summarises the tangible and conceptual contributions that were produced during the research and discusses their implications. Reflections on the tools and their transferability to other technologies and domains are also outlined.

Chapter 7: Conclusion and Future Work summarises the thesis' contributions and outlines directions for extending the ESAT workflow. A structure for conducting a full

case study is provided and the development into a full-fledged auditing framework is discussed.

The **Appendix** features supplementary material such as workshop timelines or example and raw data.

1.5 Positionality Statement

As established, technologies are launching without rigorous considerations of their impacts, often resulting in unethical systems (i.e. technologies that produce avoidable harms, are discriminatory, produce misleading outputs, or have other socially undesirable consequences). Recognising these problems and finding ways that address them was also the driving factor behind a research project at the TU Wien on ethical auditing of AI-based services (see Appendix III. Project Description for the description of the research project). This thesis is embedded in this project and documents the development of AI ethics assessment tools. A full case study of the AMS chatbot as initially described in the project description is not in the scope of this work.

I am employed as a student researcher in this project and have been selected for my background in the relevant fields: a background in *traditional* computer science with a focus on Generative AI provides for a strong technical foundation and understanding of machine learning. Additionally, courses on Philosophy of Digitalisation, Science and Technology Studies, Critical Theory of Informatics, and Technology Impact Assessment reflect my general interest in connecting technology with its societal impact and provide me with ethical perspectives that are valuable to this research. Furthermore, experience as a workshop instructor for schoolchildren where I taught informatics concepts with unplugged explorative materials equipped me with an ability to communicate complex topics effectively and allowed me to develop my own workshop exercises. I also interned as a test engineer in a software development company where I gained insights into the practical side of technology development and software testing. These varied experiences contribute to my ability to execute this project effectively.

Additionally, while there is ongoing discussion about the merit of reflexivity statements [152], I consider them warranted as they allow for reflection on one's own privileges and provide readers with insights into the researcher's position and its limitations. As the researcher conducting this work, I approach this research from within a Western, first-world context, specifically Austria. From a member state of the European Union, software development is in this context strongly shaped by the EU's regulatory landscape and embedded in capitalist structures. Considering the political and legal discourse, industry actors whose products have wide-ranging impacts are, to varying degrees, required to engage in accountability practices and comply transparently with regulations such as those governing data protection or AI governance. This perspective informs my beliefs that governance should serve the public interest through rights-based approaches, and that auditing AI technologies is an important mechanism to support this. However, I acknowledge that the regulatory context present in the EU is not universal, and that

assumptions shaped by this environment may not generalise to other regions. I recognise that a financially stable socioeconomic background and a funded research position with budget and access to resources influence my abilities to design and conduct participatory research greatly. Additionally, being a white cis man is an aspect of my identity that affords me certain privileges and influence the assumptions I bring to the research, and how I am perceived by and interact with the stakeholders in workshops.

CHAPTER 2

Background

In this chapter, I discuss the research that informs this thesis. Ethics assessment for AI lies at an intersection of multiple disciplines and requires literature on theoretical perspectives, normative frameworks, and technical approaches. Topics in informatics, ethics, law, and social sciences are therefore outlined in the following sections. I provide an outlook on the relevant literature on chatbots, regulations, and auditing. While this chapter introduces the general concepts, a more comprehensive analysis of specific auditing and assessment approaches and methodologies is provided in Chapter 4.

2.1 Chatbots

As established in Section 1.4.1, this thesis focuses on LLM-based advice-giving chatbots. To create a method that assesses this technology holistically, a brief overview of this development and its applications is required.

Chatbots can broadly be defined as computer systems which are designed to interact with humans through natural language with the goal of simulating a conversation [38, 52]. This usually happens either in written or spoken form where they represent a key application of natural language processing (NLP) and natural language understanding (NLU) [1], the field of artificial intelligence concerned with enabling algorithms to understand, generate and interact with human language. Chatbots are applied in a wide range of domains, from customer service and education to healthcare and creative writing, and can vary from simple rule-based systems to advanced generative AI models [38].

Al-Amin et al. [4] summarise the history of chatbots extensively. They trace the early origins of chatbots back to the 1960s when *ELIZA*, a rule-based programme that relied on pattern-matching to simulate therapeutic dialogue, was developed. While early systems like ELIZA or *PARRY* demonstrated the potential of conversational interfaces and established first conceptions of human-like technology, they were still limited in

their flexibility and depth. Through the 1990s and early 2000s, chatbots evolved into retrieval-based systems. These systems give selected predefined responses from large repositories, as seen in products like *SmarterChild* on AOL/MSN Messenger or even early customer service bots. A little later, voice-based assistants such as Siri or Alexa brought AI chatbots into everyday life and sparked an interest through enabling spoken interaction. By becoming part of home-automation systems and entertainment devices [4], *intelligent* chatbots introduced the notion of receiving advice from a computer in the form of natural language. Such voice-based systems were important milestones, yet their conversational abilities remained narrow and highly domain-specific.

Al-Amin et al. describe that a significant leap occurred with advances in machine learning and deep learning at the turn of the millennium. This allowed NLP models to incorporate much larger datasets and capture statistical regularities in language. However, the most profound transformation arrived with the advent of generative AI and *Large Language Models* (LLMs) in the late 2010s. By training on massive bodies of natural text and using an innovative application of high-dimensional vector representations that capture semantic meaning, LLMs could now generate fluent, human-like responses to an actually open-ended range of inputs [4]. With the release of ChatGPT, public interest skyrocketed. The company behind it, *OpenAI*, placed general-purpose conversational AI into the hands of millions of users [112] by offering their services without monetary payment.

The current chatbot technologies can be broadly grouped into four categories: rule-based systems [4] for more scripted conversations, retrieval-based systems [112] which select from pre-written responses, generative systems [112] such as LLMs producing novel text, and hybrid systems that most prominently use *Retrieval-Augmented Generation* (RAG) [8]. The latter combine the generative capabilities with access to external knowledge sources. RAG-based chatbots are particularly relevant for domain-specific applications, as they allow their responses to be grounded in knowledge bases which are then open to be modified (e.g. when data needs updating or to have externally verified sources), rather than relying solely on static training data and black-box systems. Such systems are now widely deployed in fields like healthcare, finance, education, and technology development [8]. There, they are used to support tasks ranging from coding, answering domain-specific questions, and summarising text to assisting decision-making, detecting harmful content, or providing writing support [8].

This rapid adoption of chatbots into various technologies has been driven by their scalability, cost-effectiveness, and capacity for personalisation [112]. While individuals also use them for productivity, learning, and creative exploration, chatbots are often employed on websites where the companies benefit from being able to provide around-the-clock service delivery [81]. At the same time, their limitations, especially with LLMs, remain [8]: they can generate hallucinations, struggle with reasoning, inherit biases from training data, and raise concerns around trust, transparency, and accountability. These issues have called for a growing focus in the research on human-centred chatbot design, highlighted by the already 8th iteration of the Conversations workshop [52] which took place in 2024. Dedicated to chatbots and human-centred AI, this workshop shows the

increasing academic interest in aligning chatbot design with user needs and societal values. Scholars also work to align these language models and taxonomise their biases [58, 87].

Furthermore, as these LLM-based chatbots are increasingly deployed in advice-giving roles [8], their influence on decision-making and everyday life has become more direct. This makes it essential to refine their technical capabilities and to critically address the ethical implications of their use. The importance of such ethical and human-centred considerations becomes especially clear when chatbots are placed in positions of authority, as exemplified by Albania's government initiative on its chatbot *Diella* [73], initially a public service portal to assist citizens in answering questions and linking relevant government services [147]. In September 2025, a new, anthropomorphised iteration of the chatbot was launched. On top of featuring a voice interface and an avatar depicting a woman in traditional Albanian clothing, it also represents the world's first 'virtual AI minister'. Arguing that AI, unlike humans, cannot be coerced through bribes or threats, its goal is to combat corruption in the country. With the update, the AI's tasks will now expand to decisions on public tenders. Notably, the government did not address any potential risks, possibilities of manipulation, or mechanisms of oversight [147]. Initially designed as an advice-giving chatbot, Diella has evolved into a governing actor, potentially reflecting a shift in accepting AI systems in roles that go beyond *indirect* influences. This is happening while the ethical and governance challenges of them have not even been resolved [8]. Ideally, the assessment and auditing practices for such technologies, and whatever they are transformed into, should be designed to apply to all levels of power they hold.

2.2 Ethics

When chatbots are trained on vast amounts of natural language data, they inevitably absorb human biases [4]. Biases can also be introduced through design, such as how a model is trained, fine-tuned, or instructed, leading to outputs that include misrepresentation, stereotyping, or derogatory language [58]. Hence, assessing chatbots requires an approach that moves beyond functional performance and considers how these systems may produce harm or inequality [126]. Hereby, I consider ethics as a *lens* or way to approach this problem, and not just the abstract philosophy. Fundamentally concerned with what is good or harmful in human life, and with the question of how the world ought to be, ethics provides a practical way [124] of examining the social, political, and cultural impacts of technology. As the concept of ethics is highly abstract, some clarifications are necessary. Fundamentally, and applied to this research, I understand it as a means to investigate how systems like chatbots reflect societal values, shape interactions, and influence broader structures.

Many existing fairness frameworks in AI, however, are defined narrowly and often focus on the notion of individual 'bad actors' or single identity categories, which then prioritise limited metrics like access or rights [75]. This approach risks overlooking deeper,

structural and intersectional injustices that data-driven systems, such as the chatbots, can reinforce. Ethics, in this sense, requires moving beyond the quick technical fixes towards investigating and engaging with systemic and societal issues.

One approach that addresses this is multi-scale ethics [158] proposed by Smallman. Their framework analyses technologies across different levels, from individuals and communities to institutions, nations, and global systems, and also over time. Ordering ethical effects in such a way makes it possible to outline the impacts across their different scales and include multiple stakeholder perspectives. For chatbots, this means considering not only the individual user interactions, but also, ideally, their effects on labour markets, institutions, long-term societal transformations and global concerns such as environmental impacts. Accordingly, a methodology for assessing ethics thereby should also be capable of eliciting and addressing these broader social and structural issues without being limited to technical performance or usability considerations (see Insight 4 of Section 4.1.3).

Another relevant approach is *embedded ethics* [117]. By not just adding ethical reflection after a product has launched, embedded ethics integrates the considerations of ethicists throughout the design and implementation process [118]. This prevents reactive, ad hoc solutions and seeks to ensure that ethical considerations are documented and auditable (see Section 2.4.6). It allows the designers to anticipate social and ethical challenges and supports bridging regulatory gaps around AI systems, and thereby embeds the ethics.

2.2.1 Assessing AI Ethics

The ethics-based assessment literature further expands on how ethics principles apply to AI. AI systems raise a wide range of ethical concerns, for instance through risking discriminatory outcomes, manipulation in user interactions, opacity of algorithmic processes or privacy erosion, environmental burdens and even disruptions to labour and social structures [99]. These highlight the need for systematic ways to identify and address potential harms. In their review of how the AI auditing literature conceptualises ethics, Laine, Minkkinen, and Mäntymäki [99] identify seven widely recurring principles: justice and fairness, transparency, non-maleficence, responsibility, privacy, trust, beneficence, and freedom or autonomy. These principles provide the shared vocabulary necessary for articulating the ethical concerns and serve as the foundation for ethics approaches in this domain. Building on this work [100], the authors focus on generative AI in a scoping review and identified six definitive ethics principles: respect for intellectual property, truthfulness, robustness, recognition of malicious uses, sociocultural responsibility, and human-centric design. This illustrates the persistence of principlism as a dominant paradigm in AI ethics [16, 100]. Principlism refers to an ethics approach that identifies a set of high-level principles as the basis for evaluating the right and wrong [121]. In the context of AI, such principles arguably provide a useful starting point for articulating values and guiding reflection. However, as the principles often fail to translate into actionable practices [121], they by themselves are not enough and require integration in AI development. This thesis responds directly to two of their proposed future directions

[100]: (1) the development of complementary ethical approaches to principlism, and (3) the inclusion of end-user perspectives on the ethics of generative AI.

2.2.2 Ethical Theories

Ethics principles, particularly in Western traditions, are derived from the dominant moral philosophies [115, 127]. In their book "Towards a Code of Ethics for Artificial Intelligence" [19], Boddington describes that the Utilitarian view situates values as goals that are justified by the extent to which they maximise the *overall well-being* or minimise harm. It focuses on the consequences. The author further outlines the Kantian Deontological perspective, where values are *universal duties* that must be upheld regardless of the outcomes. These could be the normative principles that one *ought* to follow. In contrast, Virtue Ethics are the *qualities of character* which guide the judgement and moral behaviour over time. These therefore differ in the nature of the values, whether they are conceived and must be followed or achieved by considering the overall goals. Additionally, they are also shaped by the societal and cultural norms and communities [16, 19]. But as values, no matter the philosophy or socio-technical framing, are not fixed or unambiguous, applying them to real systems requires negotiating the tensions between them [49]. How this is addressed in the scope of this research is described in Section 6.3.

Understanding these perspectives is an important foundation in creating a method that assists in assessing chatbots on their ethics [56]. For the materials developed in this thesis, these ideas justify an approach that assesses chatbots for the ethical personal and societal impacts, instead of just their performance. However, since no theory sufficiently accounts for assessing the diverse values involved in sociotechnical systems such as chatbots, this thesis does not commit to any single predominant ethical theory. Therefore, ethically reflexive design is important to prompt users to express their values and expectations naturally, for example, by negotiating them through interaction with others [40].

2.3 Regulations

At the policy level, a range of Responsible AI frameworks provide ethical guidelines. These include the OECD's AI principles [133], UNESCO guidelines [163], and various ISO Standards on responsible AI [79]. These frameworks, much like ethical principles, outline the high-level goals towards concepts like fairness, privacy, or safety, but often remain abstract. They lack in specifying *how* such values should be put into practice, measured, or enforced as concrete development practices or during evaluation processes. One reason for that is that policymakers also face structural challenges in regulating such a rapidly evolving field as AI [12]. The technological innovations outpace the ability for regulatory frameworks to adapt sufficiently [173]. Additionally, with over one hundred different policy and regulatory initiatives existing [157], such an amount of official documents only adds to the complexity of these problems. Similarly to policymakers, developers and providers of AI systems also struggle in navigating the requirements. Regulation is therefore often implemented superficially and considered as a hindrance to innovation

[103]. Rather than as a useful mechanism for responsible development, developers employ the regulatory demands in a check-box manner [55].

The *Center for AI and Digital Policy* (CAIDP) [23] is a non-profit organisation that assesses the countries subject to the vast amount of policies. Their Artificial Intelligence and Democratic Values Index (AIDV) [53] evaluates how countries' AI policies align with principles like transparency, accountability, and fairness, covering over 120 countries in its 2025 edition. In over 1500 pages, the countries are compared in categories like their national AI strategy, data protection, or algorithmic transparency. While the report does not provide concrete measures or binding legislation, its evaluations provide a foundation for the local implementations of regulations and challenges countries face. The CAIDP ranks Japan as the highest-scoring country of its AIDV index [53]. They highlight that Japan adopts a human-centric, multi-stakeholder approach to AI governance which emphasises ethics, risk management, and societal benefits through its AI Guidelines for Business and ongoing updates to its national AI Strategy, most likely the outcomes of hosting recent G20 and G7 meetings. The report notes that during these meetings, Japan focused on AI principles in 2019 and the Data Free Flow with Trust framework in 2023, placing the country at the forefront of AI regulation. Nonetheless, the report also highlights concerns about their unregulated technologies, such as facial recognition.

On an international level, the European Union's Artificial Intelligence Act [44] is among some of the more concrete approaches. Adopted in 2024, it is a step towards binding governance of AI in a risk-based approach. With August 2025, several clauses apply that target providers of general-purpose AI (GPAI) systems (e.g. ChatGPT and other technology that integrates LLMs into their products). For instance, key requirements are the obligation to maintain extensive technical documentation (Article 53) and to comply with detailed liabilities towards transparency and copyright. Technical documentation obligations, detailed in Annex IV [43], specify that the providers must include detailed descriptions of the system, its technical specifications (information on data and training) and its intended purposes. Furthermore, risk management procedures and human oversight mechanisms must be provided. The failure to comply is enforced through penalties (Chapter XII of Annex IV [43]). These provisions represent the most comprehensive international regulatory effort to date, but current AI regulations still face significant challenges in translating the principles and even binding laws into action.

Moreover, these regulatory approaches remain largely Western-centric, even if the technologies are available in other parts of the world. Considering the global scale, the regulatory landscape is described as 'chaos', driven by divergent cultural philosophies [172]. Comparative analyses by Hasan et al. [72] show that regions such as South Asia and Africa lag behind considerably in developing coherent AI governance structures. Countries face additional challenges from a lack of standardisation and insufficient public trust. Further comparing the United States' market-driven approach, the EU's rights-based regulatory model, and Asia's state-centric governance, the authors highlight the differences and difficulties in regulatory and ethical coherence. Scholars therefore advocate for more cooperative approaches among governments, the private sector, and the

civil society, which ideally can be applied globally [95]. Mechanisms, such as algorithmic impact bonds (based on social impact bonds where payments depend on achieving ethical outcomes), ethics stress-testing, and modular legislation have been proposed to address these issues [95].

At the intersection between regulation and practice lie the mechanisms that ensure compliance and accountability. These mechanisms are often embodied by auditing processes and oversight bodies who assess whether AI systems meet the regulatory and ethical standards set out in such frameworks.

2.4 Assessments and Audits

Although not all assessments are audits, the extensive literature on auditing provides a structured foundation to investigate many existing methods for evaluating technology. Therefore, a large amount of research for this thesis went into studying the practice of auditing. While the following sections also mention some specific approaches for auditing and related work, a deeper analysis of current methods and their gaps is discussed in Chapter 4.

2.4.1 Auditing

Auditing is considered the main structural mechanism to ensure that systems and organisations follow compliance and accountability. Teck et al. [161] describe that with roots in the financial and business sector, auditing has been established as a means to verify a company's records and detect any discrepancies between declared and actual data. The authors outline that traditional auditing relies heavily on the analysis of logs to track the data and has established methods for their assessment and verification. Over time, these approaches have expanded beyond the finance and business domains to technical domains and their related fields [123, 154], e.g. data protection and other legal systems, algorithmic governance, and even environmental management.

Considering AI, auditing emerged as the central mechanism for evaluating the regulatory performance [18, 126]. It seeks to oversee the enforcement behind regulations and aims to work towards the defined goals by assessing the current state of a system. As technologies, especially with the advent of more advanced AI systems, grow in complexity, auditing frameworks have to adapt [124]. Similarly, and particularly through the introduction of LLMs, the societal reach of those technologies also grows. This results in audits also having to assess the ethical dimensions alongside the established legal and technical ones. Mökander et al. [126] propose layered models, tackling the model, system and governance aspects. They argue that while traditional approaches captured important flaws in data records, an evaluation of ethics and broader societal needs demands that auditing cannot be limited merely to code inspection or log reviews. Bias detection and procedural or institutional accountability must also be introduced. The AI ethics auditing literature

identified overlapping themes [99], such as fairness, transparency and accountability, but is still subject to a significant lack of shared definitions and methodologies [100].

Auditing has also been mentioned more directly in AI regulations or policies, for example by Sloane and Wüllhorst's systematic review of legislations [157]. They note that New York City law[1] requires annual bias audits. According to the authors, what constitutes a bias and how principles are interpreted is defined by the US anti-discrimination laws in employment. This demonstrates how audits function as instruments that translate legal and ethical expectations into measurable checks. Because audits typically rely on immediately observable outputs and system logs, they can only evaluate impacts that the system records internally [50], whereas many harms (ethical, legal, and societal) occur outside of a system's data trace, for example, shifts in power relationships or changes in behaviour through exclusion. Therefore, strictly log-based auditing alone is insufficient and must be complemented by assessment approaches that account for social impacts, value tensions, and situated forms of harm [105].

2.4.2 Impact Assessments

Impact assessments represent a complementary approach to audits that focus more on anticipating the potential societal, environmental and ethical consequences of technologies, instead of verifying their compliance [120]. In Austria, the *Institute of Technology Assessment* (ITA) applies these practices to advise the parliament and inform the public about effects of technologies [131]. It provides evidence-based democratic oversight by conducting interdisciplinary studies that assess both the intended and unintended consequences. Such institutions are in place in different countries. In an OECD Policy Paper, Robinson et al. [145] demonstrate the increasing relevance of technology assessment (TA) and argue that, with innovation accelerating, TA must provide "strategic intelligence" by anticipating technical consequences across multiple levels. Building on this argument, the authors identify eight dimensions that TA should incorporate as inclusive practices to better respond to systemic uncertainties: *Fit-for-purpose, legitimate and trustworthy, clear granularity and scope of TA focus, smart and inclusive participation, interdisciplinary, explicit in terms of values, frames, and biases, anticipatory and managing uncertainty well, and producing useable intelligence.*

Building on general technology assessments, *Algorithmic Impact Assessments* (AIA) [120] focus on data-driven and automated decision systems. Their purpose is to also go beyond just identifying potential harms and provide accountability mechanisms. Metcalf et al. [120] argue that impacts are not pre-existing, objective facts that simply require to be detected. Instead, they consider impacts as co-constructed through the social, political, and institutional contexts that make certain harms visible and actionable in the first place. To clarify this, the authors introduce the "Assessor's Regress" as a recursive challenge: assessing harm requires a definition of impacts, but the impacts can only be

[1] https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page

defined through an assessment. Additionally, the authors point out that impacts are co-constructed through social components and the relationships of accountability among stakeholders. The authors conclude that while AIA can never capture all potential harms, they still consider them actionable within the established institutional boundaries and see them as measurable. Furthermore, they argue that effective AIA combine the technical documentation and organisational mechanism for keeping technology in check but also rely on input from the affected communities and operational accountability methods. Importantly, according to Metcalf et al., the assessment process should be directly linked to governance or implementation; AIA should influence real actions and not just produce report documents.

2.4.3 Who Audits? Internal, External, User-Involved

AI systems can be audited in several ways, and each has their benefits and disadvantages [123]. Internal audits are conducted by an organisation or company itself and thereby benefit from deep knowledge of their system. This can enable very detailed assessments, and in the context of AI, grant the internal auditors an upside by directly understanding the workflows and data of the technology. But internal audits might also face conflicts of interest.

External auditing is performed by independent parties [123]. These could be regulators, consultancies, or even civil societies. They provide objectivity and credibility, especially in communicating the findings to the public, but often have restricted access to internal processes and limited understanding over exact workings [157].

Thirdly, participatory processes involve stakeholders or users of the technology. This collaborative effort can happen through structured feedback, co-design workshops, or citizen panels. It seeks to reveal the real-world impacts in systems [100]. Hybrid approaches attempt to combine the benefits of internal expertise and external oversight. Additionally, there has been a rise in automated audits which continuously monitor the compliance metrics, which are increasingly often AI-assisted. I provide a brief critical survey of such attempts in Section 4.1.3.

The different nature of these approaches suggests that effective AI auditing is multi-layered [105]. Especially for adaptive systems like LLMs, where ongoing model training complicates the traditional assessments, auditing as a combination of the above might be sensible [126].

2.4.4 When to Audit?

Considering the development and circulation of technologies, audits can take place at different points in a system's *lifecycle* [123]. The prevalent and traditional approach is ex-post audits [166], which are conducted when a system has already been deployed. They can evaluate its live performance and real-world impacts. Ex-ante audits, on the other hand, are conducted before deployment [99, 123]. They try to anticipate potential risks and harms by investigating a system's biases or structure. While both approaches

provide important insights, the literature increasingly advocates for continuous auditing [123, 165, 166]. It monitors systems on an ongoing basis and combines the technical monitoring, periodic reviews, and ideally feedback from users. This allows companies and organisations to detect a larger amount of risks before or as soon as they emerge. Additionally, it can track the long-term impacts of a technology and, in the case of AI, keep a record of changes in behaviour over time [50]. Similarly to the question of who audits and relying on a combination of internal or external auditing, using audits at multiple points in time or continuously helps to cover a broader range of risks.

2.4.5 Ethics Audits

In an attempt to understand the AI auditing landscape, Li et al. [105] identify five types of audits: empirical, data, governance or compliance, technical, and *ethics-based auditing*. Ethics auditing in AI emerged alongside literature on algorithmic accountability [109] where it is argued that the ethical principles must be made auditable through systematic evaluations and documentation [123]. Similarly to impact assessments, ethics audits extend the traditional auditing approaches of compliance and performance metrics to address moral and societal dimensions of AI systems [154]. It aims to assess if systems align with ethical principles, (e.g. fairness, transparency, accountability) [49]. In that sense, ethics audits provide the means to examine how values are embedded in a technology and enacted within the socio-technical systems. Existing approaches [126] are often frameworks which inspect systems and models (e.g. through assessing bias) and organisational assessments that look into the governance structures and documentation. Furthermore, social evaluation through stakeholder or public engagement has also emerged [134]. Recent approaches apply the ethics auditing to levels of the model, the system, and the governance, seeking to make assessments of larger systems like LLMs more comprehensive [126]. Unlike financial or legal audits, quantifying and standardising compliance of ethics faces a challenge. On top of measurements, context-dependent principles and judgement might be necessary [99, 123].

Sloane et al. [157] outline that the growing institutionalisation of ethics audits expands the application of auditing, but it also carries the risk of reducing ethics to a mere checklist. This would undermine the critical and reflexive intent behind assessing them [125]. Sloane et al. argue, if ethics audits seek to be meaningful, they must go beyond just verifying if a system conforms to *defined* norms. They should support continuous ethical reflexivity by allowing organisations and companies to investigate how their designs and deployment shape society and its long-term consequences. This means providing mechanisms and tools that ensure accountability, ideally through tracing systems evaluations and consequent decisions, and examining how values are understood in their applying context. At the same time, ethics audits have proven valuable in making AI accountability more systematic: they translate high-level ethical commitments into processes that can be examined and improved [123] and therefore provide the foundation upon which this thesis builds.

2.4.6 Auditability and Testability

The concept of auditability conceptualises the extent to which a system (its design, data, decision-making processes and governance) can be examined [123]. Without the ability to inspect and evaluate how a system functions, assessing it, on its ethics or other measures remains symbolic. Auditability is therefore a direct prerequisite for accountability [109]. Closely linked to it is also *testability*, which strictly speaking refers to determining if a system works as intended. In contrast, auditability assesses if the means to even trace and verify what happened exist. Together, they define the foundations for meaningful AI assessment and audits [157].

Several factors can influence auditability. For instance, complex model architectures or proprietary systems (software or hardware systems that are controlled by a single company which restricts their access and modification) can limit the technical opacity of technologies. This is particularly the case in large AI systems with opaque black-box models [20]. Organisational barriers (intellectual property claims and fear of losing trade secrets) can also restrict access to a company's inner workings [122]. Additionally, even when transparency is theoretically provided, a technical documentation of a system might be unintelligible to outsiders seeking to assess it [21]. Moreover, ethical concepts such as harm, fairness, or dignity are difficult to be reduced to numbers or indicators without losing important details [99]. Because of this, ethics audits attempt to use indirect measures, for instance, representing the explainability of a model [85]. However, as Krishnan [94] points out, "interpretability" and "explainability" are often ill-defined, and focusing on them may not address the underlying ethical issues. Relying on such measures can miss larger patterns of harm, especially for groups affected in multiple or overlapping ways [93]. For example, a tool might seem fair if you look at women and people with disabilities separately, but it could still disadvantage someone who is both a woman and has a disability. Ethical concerns are difficult to be observed directly in system outputs as they sometimes only become visible in real-world use and long-term social dynamics [100]. This means that a system could pass predefined tests while still causing harm once deployed.

Limits of relying only on benchmark metrics [106, 129] point to the importance of combining them with qualitative approaches. Expert assessments, feedback from those who are affected, and participatory evaluation processes are necessary [86]. Furthermore, in their review of 117 research articles and 23 policies, Li et al. [105] identified the need to establish approaches that assist and empower the developers as an additional key requirement in auditability. How these considerations influence my approach in this thesis is discussed further in Section 4.1.3.

2.4.7 LLM Auditing

The aforementioned three layers by Mökander et al. [126] apply particularly to LLMs: In technical audits the training data, model architecture and metrics that continuously assess a system's usage and performance can be evaluated by measurable criteria. On a system

level, the user interaction (whether usability is provided), terms of services (e.g. who is responsible for the generated output), and deployment and access (are there restrictions in using the system?) are evaluated. Governance audits assess the documentation and institutional oversight through checking the regulatory compliance [157]. Nonetheless, assessing LLMs and other adaptive AI presents specific challenges due to their complexity and broad societal reach, and as systems learn and evolve based on new data or user interactions [126]. Weidinger et al. [167] characterise six specific risks: discrimination due to biases, information hazards through leaks of sensitive information, misinformation by hallucinating, malicious use through users with bad intentions (like using LLMs to generate large amounts of fake data), environmental concerns (because of power-intensive training) and operational costs, and harms in human-computer interaction (e.g. where users are misled into believing an anthropomorphic agency in the systems).

Beyond the conceptual, attempts to assess the biases in LLMs emerged since the systems gained popularity [11, 58, 62, 63, 92]. But since LLMs might generate output that is not specifically encoded during training, monitoring bias through one-time metrics is complicated, and as mentioned previously, asks for ongoing and periodic assessments [126]. Layering the approaches assists in capturing some of these challenges. Moreover, if systems that integrate LLMs, like advice-giving RAG-based chatbots, are hard to assess because of a lack of transparency, they benefit from a participatory and multi-stakeholder-involved approach [86, 141]. Incorporating user feedback and the affected communities can assist in uncovering ethical harms which the technical metrics overlook [126]. The next section highlights some of the efforts that introduce participatory approaches to democratise ethics auditing.

2.4.8 Participatory Approaches in AI Ethics Assessment

Unlike traditional audits that are often performed solely by internal or external experts, participatory approaches seek to democratise assessments by involving those affected by AI systems: stakeholders [30]. Its origins stem from *Action Research* [55] and participatory design [86]. They emphasise the co-production through cycles of reflection, action, and evaluation conducted *with* participants instead of only on or for them. Participatory ethics assessment is also inherently reflective and should encourage stakeholders and developers to examine a technology together. Frauenberger et al. [55] argue that it is not enough to be anticipatory by trying to foresee ethical issues through merely being inclusive and inviting diverse participants. Ethical participation also requires questioning empathy critically by actively engaging with others' experiences and by accepting these experiences as valid without assuming that one has to or can fully understand them.

Participatory approaches can be co-design sessions, citizen panels, crowdsourcing feedback, or structured workshops [30]. In them, users and stakeholders articulate their ethical values, concerns, or expectations. They can inform the design of a technology or the policy surrounding it. Participatory approaches align with a general movement towards human-centred [52, 108] and value-sensitive design [56] where people's perspectives are ideally included throughout a technology's lifecycle. In them, the participant's *lived*

experiences [37, 141] are centred: the sum of their personal experiences and the knowledge that has been gained through them. Lived experiences represent how people perceive and act within society and towards technologies, and what might be the background for how they perceive it; i.e. what they *bring* to the participatory setting. All these approaches bridge a gap between the technical auditing and the real-world implications and ethics of AI systems. And as ethical issues rarely present themselves as universal rules that apply uniformly across all contexts and depends on people's situated experiences, participatory approaches can be utilised in capturing what is considered harmful or fair.

2.4.9 Development-Focused Approaches

When discussing technologies, it is essential to consider the entire lifecycle of a product [105]. The lifecycle refers to all stages from the initial conception and early ideas, through design and development, to a product's market presence or use, and finally to its eventual dissolution once it is no longer in use. Traditional auditing methods typically focus on the period after a product has been designed and marketed [166]. However, the ethical evaluation should extend across the entire lifecycle to enable the anticipation and mitigation of potential issues before they happen in practice.

One approach within this perspective is *Constructive Technology Assessment* [144]. It is described as 'constructive' as it seeks to participate in shaping both development and how technologies become embedded within society. When applied to ethics, the ethical values should be an active concern and *embedded* throughout the development process [59, 66]. Relevant applications include *Ethics-by-Design* and *Value-Sensitive Design* [160]. Ideally, these approaches also incorporate stakeholder values into design decisions. They can be further combined with citizen science approaches [22] to ensure broader societal input.

Mökander et al. [123] argue that audits contribute to several upsides. From an organisational and economic point, addressing ethical or quality concerns during development is both more effective and cost-efficient than retrospectively. Further advantages the authors describe include are, an enhanced consumer trust, a stronger company reputation, alignment with emerging laws and policies, and an increased confidence among their employees and shareholders. The described approaches acknowledge that AI systems and their ethical dimensions are not static. This means that as products evolve, their ethical expectations are also dynamic and must adapt accordingly which also necessitates the continuous mechanisms of assessment and reflection throughout development.

2.4.10 Human-Centred Approaches

As ethical impacts are ultimately experienced by people in specific real-world contexts, human-centred approaches argue that those affected by AI systems must play an active role in assessing them [108]. DeVos et al. [37] showcased this by emplying interviews, diary studies, and workshops. The authors conclude that "participants showed great ability to detect and reason about potentially harmful algorithmic biases". They further emphasise

that participants were capable of identifying and interpreting harm in different ways, e.g. as misrepresentation, annoyance, anger, offence, or the perpetuation of misleading stereotypes. Adding to this growing body of work, Deng [35] has made prominent contributions to this body of work by exploring the values and challenges of user-driven auditing through interviews with practitioners of these approaches. Their research also introduced WeAudit [34], a specific method for *user audits*. Furthermore, Deng was among the organisers of the HEAL@CHI workshop series on Human-Centered Evaluation and Auditing of Language Models [108]. In the third iteration of this workshop, the growing and active community around this field was showing. I participated in this workshop by contributing a position paper [26] that introduces a method of eliciting Ethical Focus Areas (EFAs) (see Chapter 4 for EFAs) that adds to the ongoing development of user-centered auditing approaches. Discussions in the workshop highlighted an emerging consensus that what is now most needed are practical, applicable methods rather than even more conceptual papers that are merely calling for changes. The wide range of publications addressing this demonstrates that the research community is indeed making progress in advancing the field.

2.5 Summary of Challenges and Gaps

Across all the different approaches outlined in this chapter, a number of overarching challenges can be identified: First, auditability and testability remain limited in practice as opaque model architectures and limited access to proprietary systems or inadequate documentation frequently prevent meaningful evaluations. Another challenge is that traditional auditing practices risk reducing ethics to compliance checklists. This overlooks situated, long-term social impacts and the evolving nature of AI systems. Next, while continuous auditing is increasingly recognised as necessary, established mechanisms to support ongoing evaluation, especially for adaptive systems such as LLMs, are still underdeveloped. Human-centred and participatory approaches that are currently emerging begin to address these limitations, but still often struggle to capture the lived experiences of the many groups affected by AI. Additionally, only focusing on end-users makes particularly harms that are intersectional, contextual, or only visible in use over time unattainable. So far, they lack standardised processes which offer guidance in translating inputs by the stakeholders into audit outcomes and can be resource-intensive for organisations to implement.

Together, these gaps indicate that a more robust approach is needed. Ideally, such a method combines the systematic structure of auditing with meaningful stakeholder involvement and would support developers in responding to ethical concerns throughout the system's lifecycle. While this chapter introduced and discussed the relevant fields for this thesis and outlined general approaches within the domain, a more focused and critical examination of specific methods follows in Section 4.1.2. There, analysis informs insights and ultimately the concrete base of this research and the subsequent development of a participatory ethics assessment methodology.

CHAPTER 3

Overview of Methods

At the core of this thesis is the iterative development of the Ethics Self-Assessment Tools (ESAT) and its workflow. The research process follows a *Design-Based Research* approach [142]. It emphasises an iterative development through reflective practices and refining any created practical tools in real-world contexts. This means that rather than starting from predefined hypotheses, the research takes an exploratory and emergent stance that allows concepts to develop from the data and design process itself. Through that, it supports ongoing reflection and avoids premature assumptions during development. In practice, this means that the ESAT's form and content emerge through cycles of data collection, analysis, and revision. Each new insight, whether from literature or policy texts, conversations with experts or developers, or workshop trials, allows for immediate revisions to both the tools' concepts and its accompanying artefacts.

The methodology is structured in two overlapping parts: the **assessment of current methods and practices**, and the **design of the tools**. The two parts also inform each other continuously, and neither was paused when the other was developed. This allows for a dynamic and adaptive approach. Figure 3.1 shows the graphical representation of the work, and the following sections describe the applied methods of either part in more detail.

To support the development of the tool, I will draw on critical theory [45] that does not treat ethics as a fixed checklist of values but takes a critically informed perspective in this thesis, where I view ethical values as situated and contested in specific socio-technical contexts. Such a perspective enables reflection on which interests and assumptions ultimately shape the development of LLM-based systems and how ethical claims are negotiated in practice. This directs a considerate development of the tools and is further discussed in Section 6.3.

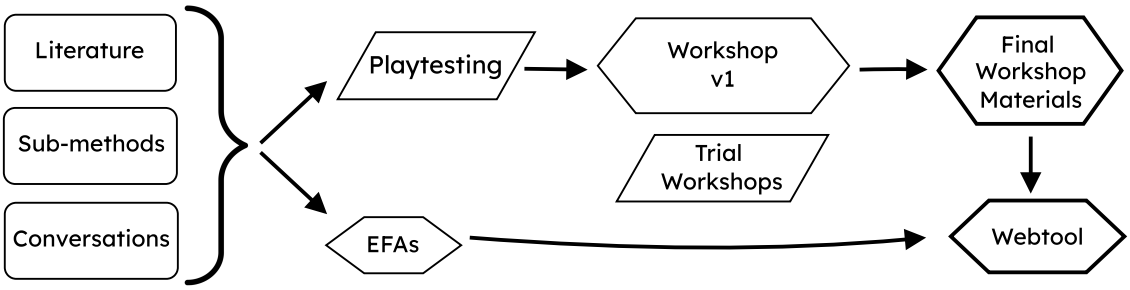


Figure 3.1: Workflow of the methodology to develop the ESAT workflow. Parallelograms indicate participatory sessions to refine the materials. The specific components are described in Section 3.2.

3.1 Objectives

The insights from the literature and practice inform the first outcome of this thesis: a structured datatype conceptualised as *Ethical Focus Areas (EFAs)*, presented in Section 4.1.4. The tools themselves can be considered the main material objective. The core components are the workshop (and all-encompassing materials) and a webtool to process data into EFAs, showcased visually in Figure 4.1. Section 3.3 describes the methodology behind their creation briefly, while the more detailed documentation of my research can be found in Chapter 5.

3.2 Components

The insights that inform the results of this thesis can be broadly structured into three components: the existing related literature, identifying and analysing auditing and assessment approaches on their feasibility to be integrated into an auditing toolkit, and what I describe as *conversations*. They are represented to the left of Figure 3.1, but their influence extend throughout the entire thesis. For instance, insights from the literature naturally also shape the final workshop materials.

3.2.1 Literature

The first component is the literature I base my research on. While systematic reviews of literature are a good approach to a structured overview of a field, its time intensiveness and the number of existing literature reviews on related literature of this thesis argue against it [105, 115, 153]. On top of that, the relevance of systematic literature reviews has also been questioned, as they might lead to research waste [137]. Therefore, my analysis of the literature is more in line with critical scoping reviews. I investigate the body of work introduced in Chapter 2 and approach the identified challenges structurally to identify gaps in the existing approaches.

Such an approach entails a critical engagement with current research practices to uncover and address potential shortcomings. Drawing inspiration from feminist content analysis, this means not only mapping themes in the existing literature but also examining whose interests, experiences, and perspectives are centered, and whose may be overlooked. By looking for power asymmetries and stakeholder gaps in how ethical and socio-technical issues of AI advice systems are discussed [102], the analysis provides ongoing input into the development of ESAT. Concretely, this goes beyond examining what auditing methods propose, to also question how they distribute responsibility. For example, this means evaluating how existing approaches recognise the challenges that developers face in commercial environments which are influenced through time-sensitive delivery dates and financial incentives. Additionally, evaluating whether they offer practical, empowering mechanisms that support responsible design without shifting responsibility away from those in the positions of control. Such insights these directly inform the design of ESAT. ESAT aims to be usable in real development contexts while also staying attentive to stakeholder perspectives at the same time.

To navigate the expansive auditing literature, I first identified key papers using the **ACM Digital Library** with the keywords *AI Auditing*, *Ethics Audits*, *LLM Assessment*. To expand on this literature with papers that might not appear in the journal's database, I used the search engine **Google Scholar** and the *Find papers* feature of the AI tool **Elicit**. I identified works by Mökander [123, 126] to be the core of recent auditing literature. Using Google Scholar's 'cited by' feature assisted in finding further literature, and the tool **Connected Papers** allowed me to explore similar papers graphically, as shown in Figure 3.2.

This results in 38 papers linked to the assessment auditing of AI, including both meta-analyses and applied methods, which served as the starting point of the research. With the tool **Miro**, I spatially arrange and cluster similarities between the papers. They can be distinguished along an axis of literature that researches auditing practices more generally (e.g. investigating requirements for auditability [105] or accountability [109]), towards research that proposes and evaluates auditing methods and applies their practices (e.g. prompting techniques [11, 129] or user-involved approaches [34, 37]). These papers were tagged with post-its of their approaches, methods and results as shown in Figure 3.3.

The literature research also expanded into related topics relevant to this thesis, including *chatbots* (and *NLP and linguistics*), *ethics*, *stakeholder involved methods*, *workshops*, and *value-driven design*. It was structured further into an assessment of current regulations and frameworks and general related approaches. I examined some concrete applied methods in auditing to conclude with an outline of what the research gap is and what this thesis aims to tackle in Chapter 4.

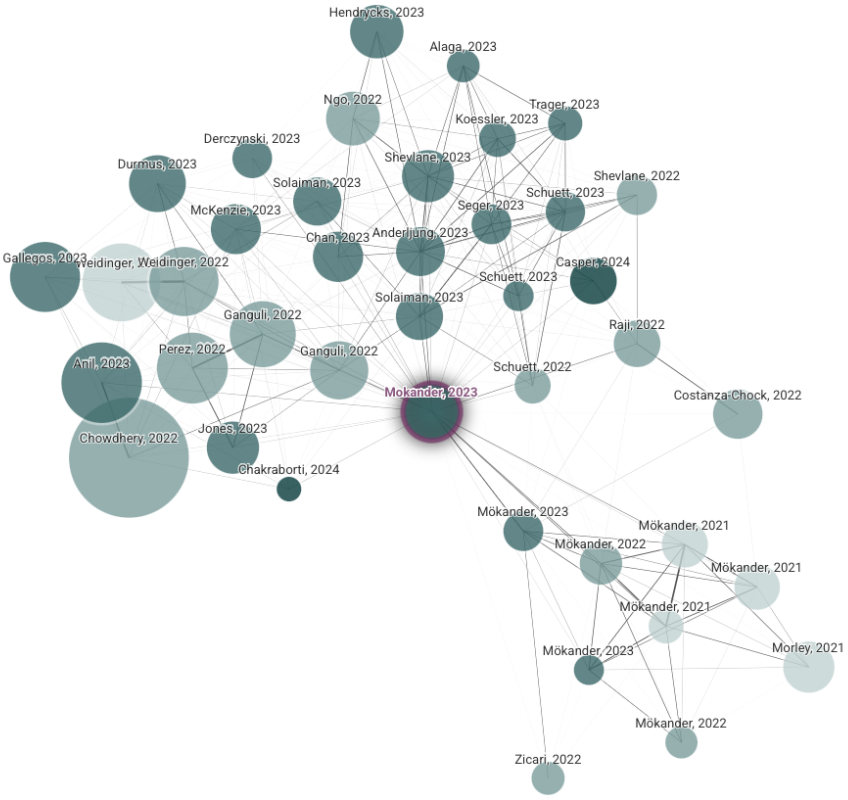


Figure 3.2: Connected Papers with Mökander et al., Auditing large language models: a three-layered approach [126], as the root node.

3.2.2 Sub-Methods

The second component can be described as the *sub-methods* that are applied in AI ethics auditing and emerged from the aforementioned literature. I focused the analysis of the literature on methods that are relevant to my aims.

To ensure that an identification and development of the sub-methods is conducted in a systematic way, I adopted an analytical iterative process. First, I assembled a list of methods that target explicitly AI ethics auditing. Then approaches from related domains such as software engineering and HCI frameworks were considered. Subsequently, I labelled each method by the following categories:

- **Purpose**: What is this method used for?
  E.g. reflection, evaluation, documentation.
- **Process**: How is it applied?
  E.g. questionnaire, online tool, workshop, interviews.

Figure 3.3: Miro board overview of investigated auditing literature structured by the degree of application (more applied to the right).

- **Application Area**: What aspect does it target?
  E.g. stakeholder identification, value elicitation, fairness assessment.
- **Relevance and Usability**: How suitable is the method for this context?
  E.g. Whether they can be applied by non-experts or need lots of training and materials to be used.

This categorisation enabled a comparison between the sub-methods and served as the basis for assessing their applicability. The summary of this analysis and the arguments behind the creation of a method in the form of a workshop and an accompanying webtool are discussed in Section 4.2.

3.2.3 Conversations

To describe the last component, I need to delineate what I mean by *conversations*: While interviews certainly give valuable insights, they can be restricted by the predefined questions and the power dynamics between the interviewer and interviewee. Ultimately, they often aim to extract information from the participant. Kvale et al. [96] therefore place semi-structured interviews in a postmodern perspective and argue that the way they construct knowledge should be collaborative rather than merely collecting data. The

authors emphasise that interviews should ideally be conversations in which the knowledge is co-produced by both sides. This became already prevalent in the early stages of the research project that underlies the thesis. Initially, a meeting with developers of an advice-giving chatbot was planned as a conventional interview. However, rather than following a strict question-answer format, the meeting was an open conversation aimed at identifying challenges they faced and exploring potential directions for addressing them. The goal was therefore to create a shared discussion in which the developers articulate their perspectives. Besides informing the early direction of the research project, the conversation surfaced several themes that are in line with industry concerns identified in the literature (see Section 4.1.1).

Similarly, a conversation was held with an expert in assessing ethics in responsible research. Both of the exchanges followed a similar structure: a brief introduction to the research project (supported by slides), followed by a deliberately open-ended statement that was designed to invite reflection and dialogue. The intention was to convey my genuine interest and willingness to learn from their perspectives and experiences by communicating that I did not have assumptions about their positions; I consider them as the experts. In both cases this resulted in a collaborative dynamic where the guests appeared comfortable sharing both their expertise and uncertainties, and the effort to *empower* the developers behind the AI systems was actually met with acceptance and encouragement.

I see an involvement with relevant actors such as developers and experts in the form of conversations necessary to avoid allowing my own academic position to dictate the ESAT design. By engaging with developers, experts, and participants, I seek to bring in diverse forms of knowledge that help identify blind spots and questions assumptions that ultimately strengthen the resulting tools. This approach aligns with feminist epistemology perspectives, which emphasise that group interviews and focus groups become most valuable when treated as a collaborative space for knowledge co-construction [170].

Complementing these conversations, regular meetings with the advisors over the course of the one-year research project in which this thesis is embedded in ensured continuous feedback. These meetings guided methodological decisions and provided structure and coherence in the development of ESAT.

Additionally, I presented and discussed the progress of my thesis at two research seminars in order to strengthen my work: one hosted by the Human-Computer-Interaction (HCI) research unit[1] and another by the Theory and Logic group[2] of TU Wien. The HCI seminar was attended by interdisciplinary professors and PhD students who are experienced in qualitative research and designing workshops. Their feedback included practical methodological advice (e.g. while it was originally planned to follow a Grounded Theory methodology, it was discussed that incorporating it in its entirety may be unfeasible and redundant in the scope of this thesis and that anchoring the approach in its principles or

[1] https://informatics.tuwien.ac.at/orgs/e193-05
[2] https://informatics.tuwien.ac.at/orgs/e192-05

a Design-Based Research can be sufficient). They also suggested that trial workshops can be held with separate groups to test variations in the materials. The discussion with the Theory and Logic group questioned how much framing should be pre-imposed in the materials or by the workshop facilitator. Particularly whether specific examples or predefined values risk shaping the participants' reasoning. This feedback reflected on the use of framing values as positive or negative, which could mask the normative or conflicting tensions that they entail.

While these conversations were not meticulously transcribed, coded, or formally analysed as traditional interview data, they proved essential in refining the methodology and materials of the design and tools and were reflective and collaborative exchanges that shaped this research profoundly.

3.3 Tool Developments

The development of ESAT can be structured around the two tools that constitute the workflow. It must be noted that throughout the research and design process, the exact structure of the tools themselves remained deliberately flexible to respond to new insights. For example, only later in the process did it become clear that a web-based tool would be necessary to intuitively digitise and process stakeholder data. Nonetheless, some necessary contributions were evident early on, such as the need for a clear roadmap (e.g., a flowchart script or decision tree) and a report format to communicate the identified EFAs.

The two ESAT tools and their elements were guided by the goal of the entire workflow: to be able to deliver the stakeholder-identified ethical areas of concern in an understandable and auditable manner. Rather than starting with tools that try to achieve this, the development approach reverse-engineered the required auditable artefact as the intended outcome. Defining and determining what elements the artefact needs is shaped by the aforementioned components (literature, sub-methods, conversation). From this requirement, the concept of Ethical Focus Areas (EFAs) emerged as the underlying auditable artefacts that the ESAT workflow aims to identify and refine.

The subsequent research is structured into two phases, corresponding to the two ESAT tools: (1) a way to collect relevant stakeholder insights (data), and (2) a way to process this data into EFAs. The two phases built upon one another by first creating the materials for a workshop and conducting trial runs of it, and subsequently creating means to process the data produced in the workshops.

3.3.1 Phase 1: Towards the Data Acquisition Tool

To address the problem raised in Section 1.1.1 concerning **what** the assessment should focus on, a method for eliciting and collecting peoples values and ethical concerns as *data* was established. When I determined that the assessment approach should revolve around stakeholder involvement, a workshop was deemed an appropriate approach to

collect the data. My experience in conducting and creating workshop materials for unplugged (non-digital) workshops (plus experiencing first-hand enthusiasm for their efficacy) made me decide to follow a similar approach of well-designed and exploratory hands-on materials [101] that build on themselves and explore a topic deeper as the workshop progresses. The workshop's material and nature were furthermore inspired by the work of Hadhigi et al. [65], who extensively developed an unplugged workshop to *speculate collectively about ethical implications* and communicated their findings and materials thoroughly. The materials for the workshop were designed with the graphical design platform **Canva**.

Trial Runs and Refinement. The development of the workshop was iteratively refined through a series of testbed workshops (corresponding to the *Playtesting* and *Trial Workshops* of Figure 3.1). First, a dedicated playtesting session was conducted using initial draft versions of the materials. Its sole purpose was to evaluate the clarity, pacing, and usability of the methods rather than to collect actual data. Based on the participants feedback and a concluding discussion with them, the structure and graphical design of the materials were adjusted and redesigned. Following this, two trial workshops were held to apply the improved materials. These sessions resulted in the first set of elicited data while also confirming the suitability of the final workshop version. The refinement through playtesting and subsequent trials thus provided the base to work on a tool that can assist in processing the data into EFAs. Additionally, it tested that the workshop materials aligned with the methodological aims of ESAT (see the detailed documentation of this process in Chapter 5).

3.3.2 Phase 2: Towards the Data Processing Tool

Secondly, as conducting the workshops only results in *raw* data, it was necessary to establish a way to process this data. As I decided that the final output of the toolkit could be either an overview of the continuously auditable artefacts (e.g. visualising tests and their assessment status) or a report, the data had to be transformed to a more structured format. To achieve this, a webtool was created to digitise the workshop output and transform it into more accessible formats, ultimately allowing for a final report to be generated.

The webtool was designed to digitise, represent, and cluster the workshop data. Aligning with the goal of being interactive, the tool's design needed to be fluid in use. It was written in JavaScript and serves as a proof-of-concept for processing the qualitative workshop data. Clustering was an integral part of processing the data, as it served as the interactive exploration of how it should be grouped [114]. For example, it features a weighted and coloured edge node-link diagram [15].

3.4 Script

As the tool is designed to be applied by non-experts, for instance developers without an education in participatory ethics, the established materials have to guide the workshop facilitator through the entire process of applying this tool. A straightforward way to communicate its functions and application is necessary. This involved creating a guide and instruction materials in the form of a flowchart script, accompanied by example slides and additional documents. The flowchart was designed in Canva and combined with other materials, it can be utilised in preparing and conducting one's own workshops.

3.5 Aim of this approach

Employing an iterative, data-driven process is essential for the complex task of developing the adaptive and comprehensive ESAT workflow. By keeping the approach adaptive and open during development, this methodology allowed a deep understanding of the underlying complexities and required flexibility to incorporate new insights. By continuously exploring gathered data, the exact form and content of the toolkit emerged dynamically. The goal was to approach the topic from multiple perspectives and thereby aim towards creating a robust and relevant toolkit that is both understandable and applicable in industry settings.

CHAPTER 4

Design Principles for Ethical Focus Areas

Building on the background of this thesis (see Chapter 2), this chapter takes insights from practice and literature to derive the design principles for the Ethics Self-Assessment Tools (ESAT). I first identify key challenges that limit existing approaches to auditing LLMs and then infer design implications and the concept of *Ethical Focus Areas* (EFAs) as the structured artefact for representing identified ethical concerns. The chapter also outlines the rationale for a workshop-based assessment. Section 4.2 translates the insights into a blueprint for the ESAT tools and explores the methodological foundations through potential sub-methods.

4.1 Insights and Conceptualisation

To ground the conceptual development of the ESAT tools in both existing research and practice, this section reviews the literature in applied LLM auditing approaches and development perspectives in creating ethical AI systems. Together, they reveal challenges and inform the thematic analysis where I identify the key gaps and translate them into design objectives to be addressed by the ESAT workflow.

4.1.1 Developer Perspectives

Across literature, scholars consistently identify a gap between conceptual AI ethics frameworks and their concrete implementation in software engineering and product development. Agbese et al. [3] demonstrate that software engineering management struggle to translate ethical principles into actionable requirements. According to the authors, ethics are primarily treated as regulatory compliance (e.g., towards GDPR) or as risk mitigation. Considering the economic and societal values for ethics remain unclear

or absent. The authors argue that this results in ethical considerations rarely receiving priority in development.

Similarly, Hussain et al. [78] emphasise that even when organisations acknowledge ethical requirements and vision statements, the existing development practices lack structured methods that would allow the actors at the different organisational levels to incorporate ethics into impacting assessment practices. Generally, the authors' interview study shows that understanding ethical requirements is often limited to basic concerns of privacy or safety and the amount of consideration is strongly dependent on the organisational structure within a company.

Holstein et al. [76] show that industry practitioners are motivated to improve fairness in deployed systems, but face substantial technical and organisational barriers. These include a lack of domain-specific fairness metrics, high uncertainty around how to diagnose and remediate fairness harms, and low organizational prioritisation where practitioners work on fairness concerns in their own time. Fairness auditing practices are also typically reactive, emerging only in response to customer complaints or negative media coverage rather than through proactive, systematic monitoring. In line with these findings, the authors highlight the need for workflows, processes, and tools that are integrated into everyday AI development practices.

These studies indicate that the central challenge is the absence of practical, actionable, and socio-technically grounded mechanisms for integrating ethics work within real-world development lifecycles[1]. This directly motivates ESAT's contribution of developer-empowering stakeholder-centred approach designed to surface ethical concerns and make them testable and auditable throughout the development of AI systems.

4.1.2 Exploring Applied LLM Auditing

As described in Chapter 2, the background literature provides a substantial amount of *Know-What*; the conceptual and descriptive knowledge in the form of guiding frameworks, principles, and recommendations in numerous publications emphasises the necessity of auditing AI systems. Comparatively few, however, offer *Know-How* that aims at linking the conceptual to applied methods and how such assessments can be concretely implemented within development settings [125]. Without discrediting the value of these conceptual foundations, I argue that there remains a notable gap for practicable methods that are usable and relevant from a developer's perspective. I examined the growing body of research on assessing and auditing LLMs to find more concrete methodologies. From the publications introduced in Section 3.2.1, I identified 19 which focus on LLM assessment. I reviewed their approaches by analysing the methodology, intented purpose and practical limitations. I further noted their recurring challenges or unresolved questions.

[1]These observations are also in line with an informal conversation conducted as part of the research project with a business development representative of an LLM-based advice chatbot. In the meeting, a strong willingness to improve the system's ethical performance was expressed. At the same time, the practitioner emphasised practical barriers such as tight development timelines, limited organisational incentives, and the absence of accessible and structured assessment methods.

It should be emphasised that this analysis was not intended to provide a comprehensive overview of all existing LLM auditing approaches; systematic reviews of the conceptual and applied methods already exist [9, 99, 105]. Recent work has also examined how literature reviews themselves can be generated effectively with LLMs (with the assistance of guidelines) [2, 174].

The aim here, rather, was to gain an understanding of the most commonly applied methods, explore how they address the challenges of LLM auditing, and to draw inspiration that informs the development of the ESAT workflow. The mentioned contributions provide some valuable examples of auditing practices, for instance through bias-detection frameworks or evaluations based on benchmarks that introduce scores [11, 62, 63, 92]. Although none claim to offer the definitive solution, they advance the discourse and illustrate possible directions in this field (or showcase directions that might not be worth investigating, as discussed below). What remains clear is that the effectiveness and relevance of any of the assessment approaches depend heavily on their context of application.

4.1.3 Thematic Insights

I identified seven thematic insights that describe recurring challenges and the corresponding design objectives. They are are illustrated with short examples to show how they may be put into practice. These insights directly informed the design choices for the specific ESAT tools:

Insight 1: The Illusion of Fair AI

Stakeholders and the public often expect *fair* outcomes of AI and their consequences, as seen in the AMS Berufsinfomat case outlined in Section 1.1.1. This creates pressure for developers and operators to demonstrate ethical performance. But LLMs cannot be fully *fair* as they merely generate tokens probabilistically and lack intrinsic knowledge and understanding of ethical concepts such as fairness [7, 112].

- **Identified Objective**: An ethical assessment of AI systems should provide clear documentation of the implemented efforts to mitigate biases and communicate them transparently, be it through concrete bias detection or reporting mechanisms. This should acknowledge that absolute fairness is unattainable, but showcase that expenses were made to detect ethical points of concern. For example, this could include specifying the hours dedicated to ethical review or the number of stakeholders consulted.

Insight 2: Assessments and Auditing as Perceived Hindrance to Innovation

Traditional ethics assessments are often perceived as slowing down the development progress [76]. Integrating *responsible* development can be challenging, as metrics to showcase these efforts are often difficult to define or measure. This results in practitioners

not logging metrics related to responsible AI principles [50]. There is also a demand for meaningful outputs that can be presented to shareholders as evidence of their successful development.

- **Identified Objective**: Auditing practices should be lightweight and seamlessly integrated into regular design and maintenance processes, without slowing down development. They should enable self-assessment so teams can apply them autonomously to demonstrate accountability. More broadly, auditing must be reframed from a perceived burden into a supportive process that *empowers* practitioners to create better working technology. An assessment tool should also allow developers to communicate their ethical 'coverage'. For example, this could include providing *proof* of how many ethical values are addressed during development and assessments.

Insight 3: Context

Ethics and their evaluations are highly context-dependent [76, 105]. Approaches effective in one context may not be appropriate in another. This is due to local and cultural norms, and how the technologies themselves operate.

- **Identified Objective**: Any assessment workflow needs to be adaptable and sensitive to context. The method should not only function in one fixed way, but also allow for adjustments when necessary. At the same time, it must be specific enough to remain meaningful. For instance, simply stating that there is a need for "fairer chatbots" is insufficient; concepts like fairness should be defined clearly in the surrounding context, e.g. through concrete examples or criteria that illustrate what it means in the specific use case.

Insight 4: Stakeholders

A central challenge in ethical assessment of AI systems lies in determining whose values should guide the process [99]. As fairness and bias are inherently value-laden concepts, what counts as "fair" can vary across stakeholder groups and contexts [76, 123]. In design practices, *Personas* are often used to represent user perspective. But they remain grounded in designers' assumptions and fail to capture real stakeholder diversity, especially for marginalised groups [28, 31]. Research suggests that incorporating user groups directly into the assessment process improves representativeness and ultimately fairness [37, 150]. It must be noted that balancing stakeholder interest in ethical decision-making is a persistent challenge [35, 84]. Conflicting values and hidden power dynamics often shape the ethical decision-making. Moreover, relying on a homogeneous user group or narrowly similar stakeholders risks reproducing the same blind spots. As ethical impacts of AI systems, however, extend far beyond the user interface [158] and may affect systemic (societal, environmental, organisational, and political domains), stakeholder groups must

be diverse also include representatives of broader impact groups (e.g. environmental advocates, labour representatives, accessibility organisations).

- **Identified Objective**: Assessment methods should make stakeholder tensions explicit rather than attempting to eliminate them. This creates a space for open deliberation between the stakeholders. It can be achieved by moving beyond traditional attempts like personas to eliciting the ethical concerns and priorities directly from a diverse and broad stakeholder group. For example, structured participatory methods must be designed to allow ethical impacts across different domains to be elicited (in addition to the direct, use-focused impacts related to user interaction).

Insight 5: Values Remain Abstract

Developers often experience ethical values as abstract and therefore difficult to assess or implement [76, 78]. The existing frameworks, policies and guidelines mainly provide conceptual direction but remain too high-level to be actionable. Similarly, Value Sensitive Design (VSD) has also "been criticised for lacking in pragmatism and methodological guidance" [164].

- **Identified Objective**: Abstract ethical principles should be translated to or supported through practical examples. Scenarios, for instance, can make ethics in the development process more tangible and directly assist the assessment of the system by providing testable real-world use cases.

Insight 6: LLM-Assisted Approaches

Approaches vary widely: some use LLMs solely as assistive tools (to support tasks like summarising, extracting data, generating reports or generating prompt questions [68, 104]). Others outsource the ethical tasks to LLMs (e.g. by rating the original LLM's output) [11, 106, 175]. Existing work also explores LLM-generated personas [91] or integrates *'prompt-engineering'* more generally [113].

Using an LLM for generating a large number of prompts with different human identifiers (e.g. the same question phrased for different genders) and automatically scoring the chatbot's answers could be considered an appealing approach appealing because it is scalable and easily automated. However, studies show that LLMs perform poorly on moral reasoning tasks like scoring answers on their ethics [155, 162]. This raises a concern in using them to evaluate ethics, especially with minimal human oversight. It also risks overlooking broader sociotechnical considerations (e.g. environmental sustainability and the resource impacts of scaling LLM-based approaches) [46, 99]. Some researchers propose 'human-in-the-loop' designs [6]. However, to ensure that humans actually 'govern' the process, such approaches should arguably be reframed as 'AI-in-the-loop'; with AI systems serving merely as the supporting tool [130].

This underscores that LLMs may be useful for automating procedural tasks (parsing logs, generating summaries, standardising the collected data) but should not be relied upon for interpreting or evaluating *human* values and ethics. These tasks involve moral reasoning and require human judgement and accountability.

- **Identified Objective**: Humans should remain the central decision-makers in evaluating ethical concerns. Assessing LLMs should only be delegated to LLMs for very procedural tasks that can be highly automated, and never rely on them, especially for interpreting human data.

Insight 7: Practical, Developer-Friendly Methods

The threshold for being ethical is too high: developers note that existing approaches require substantial training and need practical ethical assessment methods that are immediately usable [61]. Becoming or employing *ethics experts* can be costly. Additionally, it can be questioned whether such the role of an ethics expert truly exists or makes sense. For example, a *Value-Based-Engineering Academy*[2] offers 3.5 day courses for participants to become 'ambassadors' in integrating ethics into systems, at the cost of 3000€. Whether this brief online course can truly equip someone to responsibly guide the ethical design and suffice to address complex, real-world ethical challenges remain unclear. While expert consultations and training can be valuable, they are often time-consuming, costly, and inaccessible to smaller teams and do not combine guidance and practical tools in a way that is accessible, actionable, and ideally inexpensive. This can risk exacerbating existing inequalities in the AI industry, where large organisations can afford tools, programmes, and external consultants, while smaller teams are disproportionately constrained by the costs of engaging in ethical assessment.

- **Identified Objective**: Ethical assessment methods should provide developers with practical and immediately usable guidance that is easy to integrate into existing workflows. To lower the threshold of incorporating them into existing practices, such methods should not require deep prior expertise in ethics. For example, materials could be open-source and guide the user to engage with ethics by clear and engaging design.

4.1.4 Design Implications and Conceptual Result

These insights highlight a clear need for a stakeholder-involved self-assessment tool that is both practical and structured. The tool should centre humans in the ethical assessment process and should not rely on other AI during that. At the same time, it should lower the threshold for developers to engage with ethics.

[2] `https://vbe.academy/academy/`

To achieve this, I conceptualised the ESAT tool around two building blocks which together enable a structured workflow: (1) the **Ethical Focus Areas**, and (2) a **workshop** and a **webtool** that generate and put them to action.

Ethical Focus Areas

A central element of the ESAT tool is the Ethical Focus Area (EFA). It is the structured artefact that captures the stakeholder-identified concerns. EFAs encompass the key insights outlined above and are designed to translate participants' qualitative input into a concrete tangible form for developers. The *focus* refers to the potential ethical concerns in a system which a developer might have overlooked in their design and where emphasis should be put. As EFAs are derived from data identified by stakeholders, they may include ethical concerns that are already 'covered' (i.e., addressed to the stakeholder's satisfaction). But this should not be seen as redundant information, as the confirmation that an EFA is already implemented is also of value to developers and especially shareholders (see Insight 2).

Each EFA, as constructed through the envisioned workshop, consists of a title, a summary, and several components necessary to capture the ethical concerns in a structured format:

- **Identifier (ID-number):** A unique identifier assigned to EFA to reference, track, or even cross-link it within the ESAT tools to support traceability over time.
- **Core Values:** The key ethical values identified by stakeholders. These serve as reference points during the clustering and for organising the workshop data into EFAs. They can also be used to communicate ethical foci. For example, certain actors (e.g. developers or shareholders) might find it relevant to keep track of how many concerns related to the value 'Transparency' have been assessed and addressed, or how frequently it emerges as a theme.
- **Example Narratives:** Created from the raw workshop data (stories) provided by stakeholders that demonstrate the value at risk that can be read through. These should assist in understanding real-word implications of the values.
- **Suggested Tests or Evaluation Criteria:** The practical suggestions for how the system could be assessed or validated. They can be grounded in the identified narratives. Their point is to bridge the gap between the conceptual and the actionable.
- **Type:** A categorical label that indicates the type of ethical concern. For example 'user-interaction', 'model behaviour', 'data', 'deployment'. This enables a structured comparison across all EFAs or even helps identify recurring patterns.
- **Priority Ranking :** These should be an indication of the relative importance of each ethical concern, based either on stakeholder input or the judgement of the developing team (similar to how story points are assigned in software development).

- **Metadata:** For instance, information about the origin data of this EFA. These are the workshop session numbers or dates.

Through structuring EFAs in this way, they are the foundational data object of the ESAT workflow. They are structured and flexible and facilitate an engagement with ethics less abstractly than through principles alone. To demonstrate how these components are brought together in practice, Table 4.1 shows an example EFA generated from workshop data focusing on non-discriminatory recommendations. More information on how EFAs are constructed from stakeholder input is provided in Chapter 5.

EFA ID	WS-EFA-02
Title	Non-Discrimination
Core Values	Dignity, Fairness, Inclusiveness
Summary Description	Recommendations may exclude certain career options based on personal characteristics, limiting opportunities for specific user groups.
Example Narratives	• Avoid disadvantaging users based on characteristics like gender or cultural background. • Broaden suggestions to prevent reinforcing stereotypes.
Suggested Tests / Evaluation Criteria	• **No unsolicited profiling (must):** No profiles stored without explicit consent. • **Suggestion transparency (must):** Explanations accompany all recommendations. • **Suggestion diversity (should):** Recommendations span multiple fields.
Type	Model behaviour
Priority	High
Metadata	WS1, WS2

Table 4.1: Condensed example EFA derived from stakeholder workshop data: Non-Discrimination

Tool Considerations

Following the insights, the material components of the workflow were identified. It consists of two parts: a participatory workshop and a supporting webtool. While the webtool provides the digital infrastructure for processing, organising and storing the data, the workshop is the core mechanism for translating ethical discussions, reflections, and lived experiences into structured data. The workshop generates the raw stakeholder data that is later processed and synthesised into EFAs. The workshop is structured in a set of activities that prompt participants to identify ethical issues relevant to the advice-giving chatbot. Additionally, participants are able to discuss their ideas and link them to the broader value principles. In the following sections, I investigate potential sub-methods that inspired the specific workshop activities.

4.2 Towards the ESAT Workflow

After identifying the workshop and a webtool as the core components, this section first describes the blueprint of the ESAT workflow. Then, methods that are applicable to it are explored. These methods were drawn from multiple fields like AI auditing practices, HCI research, and applied ethics. The goal was to identify approaches that support the design of the workshop.

4.2.1 ESAT Blueprint

The blueprint in Figure 4.1 illustrates the components of ESAT as a workflow. It is designed as a loop, as ideally, the assessments happen continuously over the entire lifecycle of a technology. It conceptualises the workflow as a sequence of three components:

- The Input Context: The starting point of the assessment, the ideas, prototype, or technology lacking an established ethics assessment. Essentially, these are the stakeholders (with their lived experiences) and the system that they will consider.
- The Data Elicitation and Transformation Process: This phase captures and structures ethical concerns into EFAs.
  - Workshop: Implements methods for eliciting ethics-related concerns and inputs from stakeholders. Stakeholders can enter the workshop without preparation or domain-specific knowledge. Their lived experiences are the primary input from which concerns are elicited.
  - Webtool: Supports digital organisation, clustering, and the forming of EFAs.
- The Output: The EFAs or a report summarising stakeholder concerns to guide the continuous assessment of the system. This is delivered back to the starting point of the workflow, the developers. They make sense of the information in these reports and how to incorporate the concerns by adjusting their system (this work is already placed outside of ESAT).

4.2.2 Conversation about Ethics Elicitation Methods

To complement the analysis of the literature and developer perspectives, I discussed my work on March 20, 2025 with the Senior Advisor for Research Ethics and Integrity at TU Wien as an expert in identifying and addressing ethical dimensions. I presented the concept of EFAs as the structured artefact to capture the key ethical concerns necessary for developers. Then, a discussion was initiated by a question about which methods might assist in such practices. They shared that ethical scenario building can be valuable in translating abstract principles into more concrete insights. It was noted that extensive stakeholder identification literature already exists, suggesting that focus should instead be placed on designing the workshop activities themselves. Furthermore,

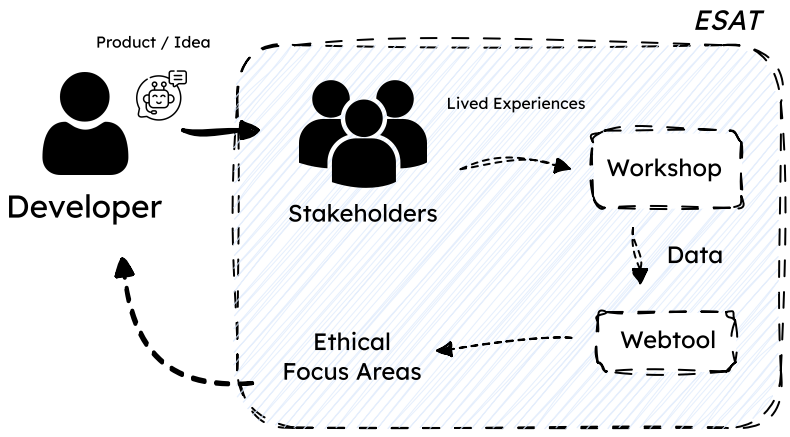


Figure 4.1: Initial components of the ESAT tool as a workflow.

it was emphasised that the term 'auditing' can be counterproductive in practice due to the aforementioned association with hindering innovation. They suggested 'assessment' as the more approachable framing and noted that communicating the use of the method can be phrased in a way that empowers developers. This inspired the decision to call the workflow a 'self-assessment'. The name suggests something that is voluntary, but applying it immediately communicates responsibility.

4.2.3 Investigation of Sub-Methods

Following the blueprint and working towards the workshop activities, a set of 22 sub-methods (and general research practices) were selected based on their direct relevance for the workshop. These were analysed by purpose, process, application area, and relevance for non-expert practitioners (Section 3.2.2).

The methods were divided into categories aligned with the core workflow of what they would relate to in a workshop: stakeholder engagement, concern elicitation, mapping of values, forming EFAs, and system testing. The stakeholder identification and recruitment methods (e.g. mapping of direct and indirect stakeholders) were also included since the tool targets developers who may lack participatory design experience. Some methods were not directly applied but remain conceptually important. For example, *Actor-Network Theory* (ANT) encourages the holistic and post-anthropocentric perspective that not only humans but also societal and environmental factors are actors and thereby part of ethical assessment [136].

Sub-Methods Overview. Table 4.2 summarises the selected sub-methods which facilitate stakeholder engagement and the translation of ethical concerns into structured insights. Insights that influenced the workshop materials directly are highlighted in bold in the table.

Method / Practice	Description and Applications
Stakeholder Identification and Participation Engagement	
Direct/Indirect Analysis	Systematic identification and categorisation of affected parties by direct or indirect stakeholders [17, 57]
Actor-Network Theory	Framework for treating both **human and non-human entities as actors** that influence and are affected by system outcomes [136]
Recruitment	Recruit participants through context-sensitive methods (e.g. snowballing [125, 148]), targeting a diverse stakeholder perspective [65, 86]
Participant Preparation	Briefing participants to ensure an understanding of context and readiness for structured deliberation ('tutored deliberants') [40, 86]
Compensation	Provide compensation for participants' time and effort [86]
Eliciting Ethical Concerns	
Participatory Design	**Workshops** or co-creation sessions to identify ethical concerns [22, 55, 86, 111]
Social Identity Mapping	Letting **participants map** their **social identities** to reflect on privileges, **roles**, and perspectives within a systems' context [80]
Application Observations	Observe systems through *walking the system* [54, 144], focus groups [22, 84], *ethnography* [84, 90], and *journey mapping* [151] to identify potential issues and user experiences.
Scenario Building	Developing context-specific hypothetical cases e.g. asking 'What could go wrong?' [82, 146]. Value Scenarios [36] to make abstract values more concrete through **examples** [169].
Ethical User Stories	Software engineering concept (user stories) to translate values into feature requirements as a structured form. Includes acceptance **criteria**, priority and implementation estimation [66, 67, 89].
RAD Framework	Supporting ethical critique processes by recognising harms, socio-technical analysis, and deliberating a more ethical future [159]
Value Focused Cards	prompting dilemma discussions through 'Envisioning Cards' that "guide design process through explanations of key themes in a systematic way" [36].
ECCOLA	Gamified card deck providing structured ethics assessment with **practical examples** for understanding [164].
Value Mapping	
Value Sketch	**Visual representation** of key factors, for example in the form of drawings, collages or diagrams [57]
Value Dams and Flows	Resolving tensions by defining strong objections ('Dams') and highlighting appealing alternatives ('Flows') [36, 57]
Ranking / Weighted Scoring	Assign **relative weights** to concerns or values, supporting **prioritisation** e.g. with a Matrix [88]. And stakeholders **ranking** the values [13]
Towards EFAs	
Value Source Analysis	Identify **which values are prioritised** by different stakeholders, including designers, users, and project teams [57]
Critical Questions	Generating testable queries from EFAs to probe the system [48]
Falsification	**Stress-testing system boundaries** by attempting to expose failure points or ethical lapses [106]
LLM Evaluation Methods for / Test Forming	
Automation / Benchmarks	Aequitas is a bias detection toolkit for AI systems [149], or through a multiplexity lens [129]
Sentence Completion	Evaluate the system output pattern and responses by using partial prompts [92]
Essay Prompting	To assess the reasoning and ethical decision-making capabilities of LLMs through essay generating prompts [11]

Table 4.2: Overview of the sub-methods necessary for establishing a workshop-based (ethics) assessment of AI. In bold are the insights that influenced the workshop materials directly.

4.3 Conclusion

This chapter established insights from literature, conversations, and applied auditing approaches to identify key challenges in assessing LLM-based systems. The thematic insights highlighted the need for practical, stakeholder-centred and context-sensitive methods. They further inspired the creation of Ethical Focus Areas and the workshop-webtool workflow of the ESAT tools. This assisted in narrowing down the sub-methods. Their analyses laid the foundation for the design and iterative development of the ESAT tools, in particular the materials of the workshop, which is presented in Chapter 5.1.

CHAPTER 5

Designing the Tools

Building on the blueprint from Chapter 4, this chapter details the design and development of the Ethics Self-Assessment Tools (ESAT) core components. It describes the steps through which stakeholder data is elicited and subsequently transformed into the structured form of Ethical Focus Areas (EFAs). To clarify the terminology, since ESAT is intended as a method for developers to assess their AI system, the actors who apply ESAT in practice (those who conduct the workshop and process the collected data) will subsequently be referred to as *developers* from here on.

The chapter is organised into five sections. **Section 5.1** presents the final workshop materials and describes the sheets, materials, and technology required to run an ESAT workshop. **Section 5.2** describes the playtesting session of the initial, provisional versions of these materials and discusses the refinements that followed. **Section 5.3** describes the subsequent trial workshops, which generated the data used in creating the processing stages. **Section 5.4** introduces the accompanying webtool and shows how the raw workshop data is digitised, structured, and processed into creating EFAs. Finally, **Section 5.5** concludes with a summary and the detailed diagram in Figure 5.14 visualising the entire ESAT workflow and its components.

5.1 Final Workshop Materials

This section presents the final workshop materials and outlines the activity sheets and components required to conduct an ESAT workshop. The empty templates of the sheets are provided in the Section III. Workshop Materials of the Appendix. To support a clear understanding of the workshop activities, a running example illustrates how a single participants input scenario develops across the workshop.

To develop the materials, their goals were identified from the insights and design objectives of the previous chapter. It was also necessary to recall that the input to the workshop

are the stakeholders as participants who bring their lived experiences (see Section 2.4.8) and expectations towards or experiences with a technology to the workshop. These form the foundation for eliciting the ethical concerns, and the workshop materials must be designed in a way that brings those concerns to light.

5.1.1 Goals

To focus the design, the insights from the surveyed literature (see Section 4.1.3) informed several goals that the workshop materials need to fulfil:

(1) The workshop materials should be sufficiently **clear** and structured to ensure the participants can use them largely on their own; the activities should 'guide by design'. Since ESAT is intended to be applied by developers who may not have facilitation expertise, the workshop facilitator's role should be limited to simply guiding the transitions between activities (e.g. collecting and handing out the relevant sheets) rather than providing detailed, step-by-step guidance during the activities themselves. They merely have to communicate the goal of an activity before it starts. A concise script the developer can refer to during a workshop assists here. This addresses Insight 2 and Insight 7.

(2) Prior to starting the workshop activities, the participants must receive clear **information** about the workshop provide informed **consent**, and be assigned **anonymised identifiers**. This ensures that all collected data is processed securely and confidentially.

(3) Participants should require **no preparation** or technical background knowledge. They only need to 'bring' their own perspectives as stakeholders. The workshop design should also not assume that participants know each other beforehand but allow strangers to work together effectively.

(4) One part of the workshop should be to give participants the space to reflect on **who they are** and what **roles** they inhabit in different aspects of their lives. This aims at grounding the ethical concerns in the participants' lived experiences.

(5) Participants should express expectations, demands, and concerns they have towards a technology through **scenarios** they can think of in interaction with or concerning the existence of a technology. As stakeholders (Insight 4), they can represent concerns towards the context a system is situated in, in line with Insight 3.

(6) Building on the scenarios, the participants should be prompted to express **criteria**. These are concrete examples that specify what would **satisfy** the scenario and what would **frustrate** or undermine it, thereby move away from the abstract ethics identified in Insight 5

(7) The workshop should allow participants to articulate **values** and by mapping them to the elicited scenarios and criteria, further addresses Insight 5.

(8) The activities and their outcomes should be captured in representative and easy-to-handle data to communicate the efforts made in assessing ones technology in a well-documented way, relating to Insight 1.

5.1.2 Design Choices

Given that physical artefacts and writing on paper slows down the process of thinking and supports the interactive group setting [64], the workshop materials were designed in an unplugged (non-digital) format [101]. Relying on pen and paper also minimises technological barriers. The materials therefore consist of sheets of paper that can be worked on throughout the entire workshop (i.e. by placing an early activity's sheet on another one).

In alignment with goal (1), the materials were to be reproducible in a cost-efficient manner. Using printed sheets allows the workshop to be conducted with a very minimal setup (apart from a printer and a way to display instruction slides, e.g. a beamer) and without any special software or infrastructure. The sheets can be reprinted for each workshop or even laminated and worked on with non-permanent markers that can be washed away with water. Graphically, the materials were designed to be visually clear and consistent. Icons indicate certain instructions (e.g. a small house for sheets that are not collected as workshop data). Each sheet is also self-contained and includes the information that is needed for one activity with short explanations where necessary. The sections are only coloured to indicate their boundaries and are not indented to convey any meaning (as green is often associated with positive and red with negative). A monochrome version is also created to be provided in case the writing is illegible for participants with colour vision deficiency. However, a pen-and-paper format may introduce accessibility limitations for participants with specific needs (e.g. people with low vision or reliance on digital assistive tools). These can be mitigated by offering adapted variants of the workshop (e.g. high-contrast, large-print sheets, tactile materials, sign-language support in workshop sessions, or accessible digital versions) that are provided to participants.

5.1.3 Slides

A presentation is used to support the workshops and guide the participants through the session. The slides were designed to be simple and visually clear. They should contain all essential information about the workshop and its purpose and structure. Each activity is introduced with a short explanation where relevant examples are shown which allow participants to understand the tasks. During the activities, the slides should present a summary view that includes the key information about the technology being assessed on one side and concise instructions on the other. This allows participants to easily refer back to the context and requirements at all points during the activities. An example of such a layout can be found in Figure 9 in the Appendix. Additionally, a timer is shown to indicate the remaining time for each activity. This helps participants manage their pace throughout the workshop and notifies them if an activity is about to be over.

5.1.4 Pre-Workshop Information and Consent

Since ESAT relies on stakeholder input that may be grounded in personal experiences, expectations, or concerns, it is essential to establish a transparent base. Before the workshop activities begin, participants must receive all necessary information about the session and provide informed consent for their participation and for the processing of their data. The participants receive printed information sheets and consent forms at the start of the workshop (ideally also in advance so participants can sign the forms digitally to save time and paper). The workshop facilitator must allocate sufficient time for participants to read through the documents and ask clarifying questions. Examples of necessary pre-workshop documents can be found in Section V. Consent Forms of the Appendix. These documents were also for the workshops conducted in this thesis.

Information Sheet. To ensure that participants have a clear understanding of the workshop's aims, they receive an information sheet outlining the purpose of the workshop, what their participation will be, and an estimated duration (see Figure 16 in the Appendix).

Data Protection Information. In accordance with GDPR requirements, participants are also provided with a data protection statement (see Figures 17 and 18 in the Appendix). This document specifies what data is collected during the workshop, how the data is stored, for what purposes it is are processed, who has access to the data, and when it is ultimately deleted or archived. It also informs participants about their right to withdraw from the workshop at any point. The statement further explains that all collected materials are processed in a pseudo-anonymised form. Participants are assigned participation numbers at the beginning of the session, which are used in place of their names on the worksheets. No mapping from identifiers to individuals is necessary, as the identifiers serve solely to keep a participant's activity sheets linked to one another throughout the workshop.

Informed Consent Forms. After reading the information sheet and data protection statements, each participant must sign the consent forms. A data protection consent form documents agreement with the processing conditions described in the data protection statement (see Figures 19 and 20 in the Appendix). The consent form confirms the voluntary participation. Only once these forms are completed can a workshop proceed to the activities.

5.1.5 Activities

The analysis in Chapter 4 suggests a workshop as the most appropriate means of eliciting the stakeholders data, a first version of the activities of it was designed in accordance with the goals above. Feedback sessions with the advisors and the presentations at the two TU Wien research units (see Section 3.2.3) refined them. The following activities are part of the final version of the workshop:

Introduction and Icebreaker (Plenary)

Participants are seated at tables randomly in small groups of three to four per table. The group size was chosen as pairs risk creating dynamics where one person dominates the exchange, while larger groups reduce the opportunity for each participant to contribute meaningfully and make coordination more difficult during the activities. The trial workshops further confirmed that such a size provided a balanced setting. The workshop starts with a general introduction by the facilitator about the purpose and the format. This includes establishing a respectful environment and communicating the workshop goals. The technology to be assessed is also presented briefly, and if available, prototypes or examples of the technology can be shown. The participants do not need a thorough understanding about the exact workings of the technology. Any assumptions they have about it can also inform inputs in the activities. The facilitator can prepare and use the ESAT example slides[1] as assistance. Every participant receives a number, and an icebreaker activity allows the participants to get to know each other, corresponding to goal (3). Through an intentionally simple question, the icebreaker helps reduce the initial tension in the room and loosens the atmosphere. In the trail workshops, it took the form of briefly introducing oneself with one's name and childhood dream job, followed by whether one eventually pursued it or what path they took instead.

Activity 1: Identity and Roles (Individual)

To target goal (4), the first activity allows participants to situate themselves and reflect on their identities and social roles from which they approach and evaluate the technology. They fill out two worksheets that remain private to them. This is indicated by a small house icon in the top right of the two sheets. Because they contain personal an potentially sensitive information, they only serve to support each participant's own reflection and are not needed as workshop data. Not collecting them also encourages a honest reflection.

Identities. For the first, participants reflect individually on their own identity through a social identity map [80]. A template version is shown in Figure 3 in the Appendix. The categories of it are: age, socioeconomic status/class, gender, physical appearance, ability/health, ethnicity/national origin, beliefs/religion/worldviews, language, and sexuality/sexual orientation. To not pressure people, it is optional how many of these categories are filled out.

Roles. The second sheet is an adaptation of a social identity map which lets participants reflect on the roles they hold in different aspects of their lives: personal, professional, and the public (see Appendix, Figure 3). Their categories correspond to the roles one has in their free time (e.g. about the role in social relationships, family, and hobbies), at work (e.g. about the career identity, workplace relationships, and professional expertise), or who they are as a citizen (e.g. about community involvement, civic duties, volunteering

[1] https://teletobe.github.io/audit-share/workshop-sample-slides.pdf

roles, and advocacy). This helps the participant recognise the standpoint from which they are approaching and evaluating the technology. Additionally, ethical values are added to the roles they hold. As an inspiration, they receive a sheet with example values. All of these activities familiarise the participants with the concept of roles and values they hold, which will be applied in subsequent activities.

Example for Activity 1 (Roles) *A fictional participant filled out the roles sheet as shown in Figure 5.1. Each category includes a question that prompts them to think about who they in the different spaces. In the case of the example, many roles were added in each category and several values have been associated to the roles.*

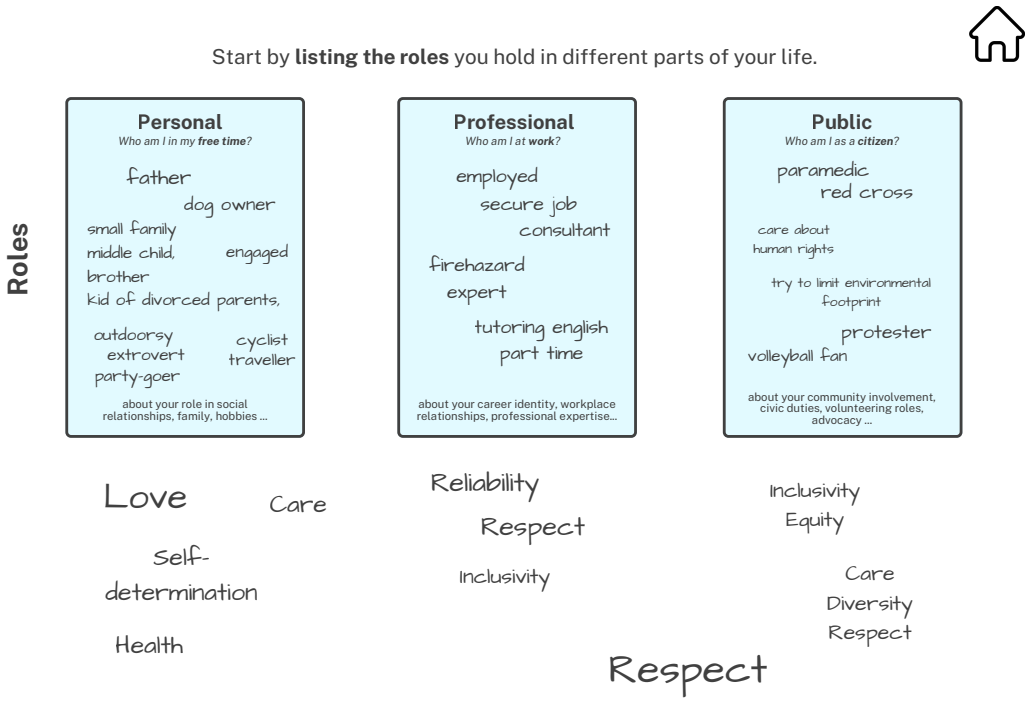


Figure 5.1: Example sheet of the roles activity.

Activity 2: Stories (Individual)

Targeting goal (5), the second activity elicits scenarios that express participants' expectations, wishes, and ethical concerns regarding the technology. A blank template is provided in Figure 5 in the Appendix. This activity is inspired by user stories from a software development context [67] and follows a similar structure. Different to user stories in software engineering, the role they hold is placed at the end of the structure to emphasise the concern before the *reason*.

"I want _____ so that _____ because _____ ."

The first part (I want) is accompanied by the prompt "what do you want the technology to do? What do you need it to be?" It refers to the *wish or desire* a participant might have towards the technology. Here, participants can also circle a phrase to specify their wish (to be able to, to know, it to be, it not to, to feel).

Secondly (so that), participants are questioned: "Why do you want that? What matters? What ethical issue or challenge comes up?" This corresponds to the *purpose* behind the elicited wish.

Lastly (because), they should answer "what about your role, identity, or perspective makes you think of this?" It provides a *reason* for the previous statements. Two optional fields ("I am" and "I know") prompt participants to connect their reason to lived experience or relational roles they hold (e.g. as a parent, colleague, caregiver).

Importantly, the stories are not limited to interaction-level wishes (about using the technology). Participants are encouraged to express any concern, expectation, or tension they have. In line with Smallman's multi-scale ethics [158], participants can also create stories that are about societal, organisational, environmental, or other ethical concerns. They also may leave out the "because" part in case they prefer to not disclose their role. The activity is conducted on A5 sheets and the participants may complete multiple stories. After each participant finishes their first story, the activity is briefly paused for a short discussion amongst the table where participants can share their examples. This inspires additional perspectives and supports the ones who may not have been able to come up with a scenario initially. A second round of individual story-writing then follows where participants can create more stories.

Example for Activity 2 *A participant raises a concern about a system's social media feed, displayed in Figure 5.2. Their complete story reads:*

> *"**I want it not to** promote content that includes unrealistic beauty standards,*
>
> ***so that** my child doesn't feel pressured to look or act a certain way,*
>
> ***because I am** a parent of a 13-year-old."*

Activity 3: Criteria (Group Discussion)

The third activity aims to make the previously created stories more concrete. This activity is carried out in small-group discussions. The process is the following: participants are asked to chose the Activity 2 sheet they want to continue working on. Then, one participant presents their chosen story and the group discusses it together. They collaboratively work on an Activity 3 sheet (see Figure 6 in the Appendix) to come up with three criteria that specify how the scenario could be addressed:

1. Ideal would be _____ .
(Given this story: Ideally, what do you want the system to be like or do? What are scenarios that are desirable?)

2. Acceptable would be _____ .
(Given this story: what should the system be or do for you to feel it is supporting your needs? What are examples of the basic standard you expect?)

3. Deal-breakers _____ .
(Given this story: What would you consider undesirable? What are examples where the system becomes unacceptable or unusable? What should it never do?)

The term *system* could be clarified to a specific workshop use case (e.g. advice-giving chatbot). This activity is designed as an A4 sheet. The sheet from Activity 2 can be placed upon it to continue working on the same scenario as before, as shown in Figure 5.2.

Only one story is addressed at a time to give each scenario enough attention. During the discussion, the group shares their interpretations, concerns, and perspective towards the presented story. This process is repeated until every participant has had one story discussed in the group. While participants might produce criteria with overlaps or minimal differences, the purpose is to reveal concrete examples that relate to the stories. The aim is to go from the (potentially abstract) wishes or concerns towards actionable measures which later serve as the testable elements in EFAs. There they can be mapped to specific test cases and, for the example of a chatbot, represent example situations that should (or should not) happen, thereby addressing goal (6). These examples could even be directly adopted as ethical requirements towards the technology that is assessed.

Example for Activity 3 (Example 2 continued) *Building on the concern about a social media app's feed, the participant places their Activity 2 sheet on the left of the Activity 3 sheet (see Figure 5.2) and fills the right side out with the following three criteria:*

*"**Ideal would be:** it actively promotes diverse, realistic representations.*

***Acceptable would be:** it allows me, as a parent, to review or monitor recommendations.*

***Deal-breaker** if my child gets recommended multiple videos about how a body has to look."*

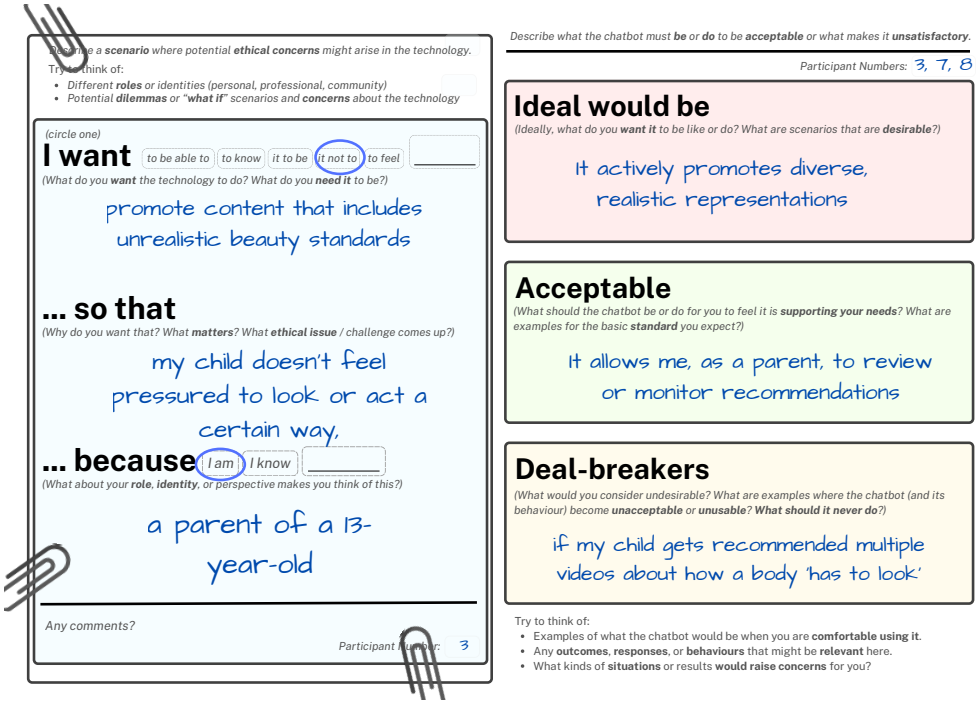


Figure 5.2: Filled out example sheet from Activity 3. Shows Activity 2 sheet pinned on the left.

Activity 4: Values (Group Discussion)

The final activity instructs participants to identify and position values associated to the scenario they previously discussed. Each group receives a set of value cards and the Activity 4 sheets placed on A3-sized cardboards (see Figure 7 in the Appendix). On this cardboard, the Activity 3 sheets can be placed. Participants subsequently again work through each of the sheets one after the other and collaboratively. They each select one value they associate with this sheet. Then, they discuss it, position it on the board, and describe a reason for their choice. Although value discussions can easily drift off track abstract or become extensive, participants are encouraged to rely on their instincts and initial impressions when selecting the values. The board is structured along two dimensions:

- **Importance** of the value (important vs. super important).
- Whether addressing the scenario would **promote/enable** the value or whether the scenario puts at **risk/hinders it**.

The Activity 4 board includes the phrases "Addressing the concern would promote / enable:" and "The concern puts at risk / hinders:" to guide the participants in their placements. This activity uses a curated set of value cards adapted from a list provided

by the advisor. The full list of available values is included in Table 1 in the Appendix. This value mapping process aims to clarify what is at stake in each scenario. It also supports the creation of EFAs by indicating which values are super important, promoted, or at risk. Additionally, it can be utilised in prioritising certain EFAs or communicating (e.g. to shareholders) which values are being considered when EFAs are addressed.

Each table works on one board at a time, pinning or sticking the values onto the cardboard. Ideally, each group manages to work through all of the cardboards; however, if time runs short, fewer boards may be produced. The output consists of as many boards as the groups manage to complete. Each board represents one story, its criteria, and the values associated with it and serves as the *raw* data that is coming out of the workshop.

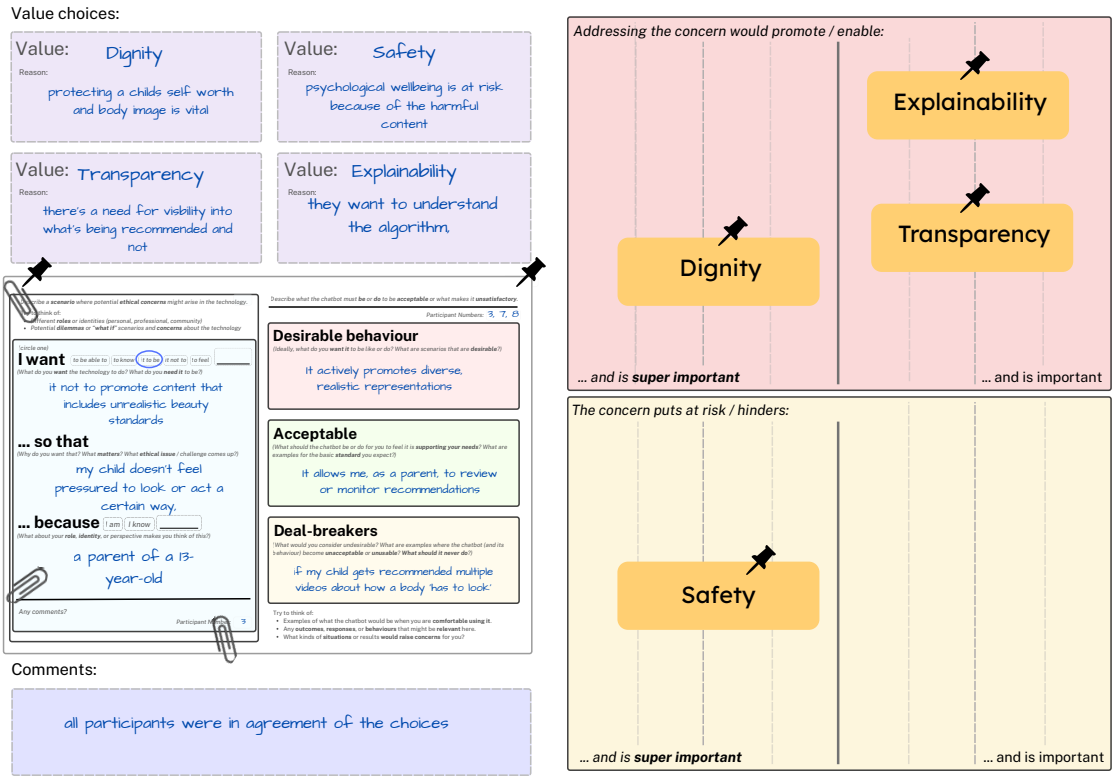


Figure 5.3: Filled out sheet for Activity 4. This is an example of one final board in the output of a workshop.

Example for Activity 4 (Example 3 continued) *The group place the Activity 3 sheet on the left of the Activity 4 sheet. They proceed to pick one value each. The selected values are: Dignity, Safety, Transparency, and Explainability. One by one, they discuss each value and where it should be placed. For instance, they decide that the value Dignity was chosen with the reason "protecting a child's self worth and body image is vital". They determine that addressing the scenario they are working on would promote Dignity and*

proceed to place it on the "super important" side of the value space. This process continues for each participant until all values are placed on the board. The final outcome is shown in Figure 5.3.

5.2 Playtesting with Provisional Workshop Materials

To test the materials, I held a pilot session of the workshop with members of the Theory and Logic research group at the TU Wien on June 25, 2025. Eight PhD and master's students participated in the three-hour session. The purpose of this playtesting was to assess how well the workshop materials and instructions function in practice, and not to evaluate the participants' outputs. The focus was therefore placed on the usability and clarity of the designed activities and worksheets. The entire session provided an opportunity to observe how the participants interacted with the materials, where confusions were, and how groups discussed. The materials used in each activity were prototype versions of the final materials presented in Section 5.1. Figure 5.4 shows participants working on Activity 4, which, in the playtesting session, consisted only of a blank cardboard.

Feedback was acquired through several short discussion rounds after each activity and by collecting the feedback sheet (see Appendix, Figure 21) which was distributed at the beginning of the session. These inputs helped in identifying which materials or instructions

Figure 5.4: Participants during the playtesting working together on Activity 4 by placing values.

were working as intended, and more importantly, which would require refinement. The playtesting led to a number of adjustments across all the activities and also provided general insights into facilitating a workshop and guiding the session.

General Feedback

A recurring critique among the participants was a general uncertainty about the purpose of each exercise and how their individual inputs would eventually contribute to the overall process of evaluating the technology. Some participants even questioned the necessity of particular activities.

To address this, the workshop should include a short introduction for each activity in which the facilitator explicitly mentions its *purpose* and what that part of the data might be used for. For example, it should be explained that reflecting on one's role in Activity 1 helps the participants to recognise their diverse positions, which they then can use in the following activity to ground their ethical concerns in. A script that assists the developer in communicating the goals of each activity was subsequently created (see Section 6.1.2). Similarly, showing a final example of an EFA would illustrate how the developer might use and engage with the contributions of the participants.

During the playtesting, I showed a timer on the screen that indicates the remaining times for each activity. This was perceived positively, as it allowed everyone to pace themselves. The timer should not be strictly enforced but mainly provide guidance for how long they can continue working on it, e.g. by incorporating a buffer period of 2 minutes after each activity to allow the work or discussions to conclude naturally. The alarm of the timer was also helping in regaining the attention of the participants during lively discussions.

Participants also reported that the examples shown at the start of an activity were too short and it would be useful to have them visible during the tasks. It was also discussed whether that would be too *leading* (where participants would only come up with data that are similar to the example), but they voiced that the opposite might also be correct, and they could actively think of alternatives to the shown example. In response to this, the format of the slides was adjusted to keep examples visible while the participants work on the activities, either on printed handouts or the slides.

Other general feedback referred to the group discussions, where some participants assumed that a consensus had to be achieved for the exercise to continue. This led to prolonged debates. It will therefore be made explicit that disagreements are accepted and that they can be documented directly on the sheet. For this, a dedicated comment section has been added to the relevant activity sheets for participants to record any differing opinions or uncertainties. Overall, feedback sheets that were filled out after the playtesting revealed that the structure of the workshop, and its pacing were well received and the materials were praised for their design and clarity.

Adjustments for Activity 1 (Identity and Roles):

Originally, the first part of the workshop combined the icebreaker and identity reflection into one integrated activity. During the playtesting, participants perceived the initial icebreaker activity as rather disconnected from the rest of it. As a result, the structure was adjusted so the icebreaker now is just a short introductory activity combined in the introduction. Activity 1 therefore solely focuses on the identities and roles.

It was also reported that the initial version of the identity and role mapping introduced too many categories (see Figure 5.5), which resulted in a lot of cognitive load for an exercise that is merely intended to situate the participants. The activity was therefore simplified into only 3 overarching categories of roles: *personal*, *professional*, and *public*. It was also mentioned and showcased in an example that not all of the fields had to be filled out if one cannot immediately think of something to put in it. It was also noted to be difficult to think of roles during this first exercise.

Participants were sometimes unsure which of the sheets remained private. To clarify this, a small *house* icon was added to the sheets that contain personal information and should be taken with them and are not workshop data to be collected. This applies to the Activity 1 sheets and the information and data protection sheets.

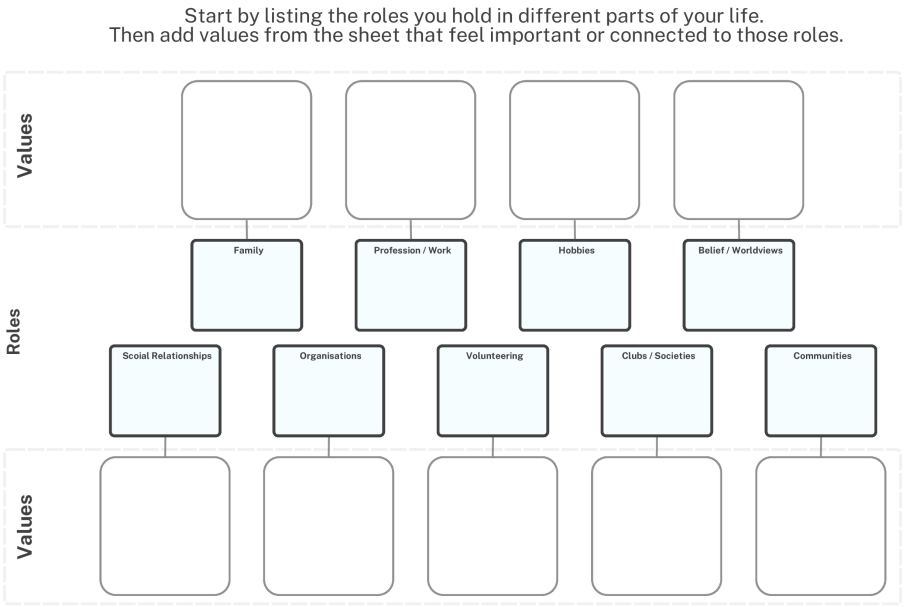


Figure 5.5: First version of the Activity 1 worksheet.

Adjustments for Activity 2 (Stories):

One challenge was the phrasing of the story template. The first version read:

> "I want _____ so that _____ especially as a _____."

Some participants expressed uncertainty about what to write or how specific their story should be. In the first version, the prompt helped structure their thoughts, but they hesitated when starting blankly. To address this in the final version, the activity is interrupted after each participant fills out one story sheet. A 5 minute discussion where participants share with their group the story they came up provides inspiration to others. This also allows for participants who might not have come up with a story in the beginning to recall additional perspectives and think of one now.

Some of the participants noted that the initial phrasing (without words to circle) was too open and they were unsure what exactly was being asked of them. To assist in specifying the prompts, the "I want _____" received optional fields that could be circled. These can be used to narrow the desires, for instance, by considering that one might want to know certain things about a technology or a system to be able to do something (or for it not to do certain things). The fields that were added are: "to be able to", "to know", "it to be", "it not to", "to feel" and a free field to add their own if they wish so.

The phrasing "especially as a _____" in the first version (see Figure 5.6) was also perceived as too restrictive, as it required identifying a single active role. The phrasing was therefore changed to "**because** _____", and two optional fields "I am _____" and "I know _____" were added. This lets participants describe both direct and indirect forms of role involvement. For instance, someone might relate to a concern because they personally experience it or because they know others closely who do, which allows a connection to the participant's direct lived experiences and to roles they might hold as relatives to someone (see the example in Figure 8 in the Appendix).

Participants also mentioned that thinking back at their reflections and roles from Activity 1 helped them articulate more stories. Therefore, the instructions to the facilitator in the script explicitly tells them to look back at the previous sheets as an inspiration for writing the stories (see 'Don't forgets' of Activity 2 in Figure 6.2).

Adjustments for Activity 3 (Criteria):

Overall, this activity was perceived as the clearest one. The feedback from the participants still inspired some notable changes. It turned out that discussions of Activity 3 are rather time intensive. Therefore, while it was originally planned that a group chooses the sheets they want to work on together, this was changed so at the end of Activity 2, each participant is instructed to choose the one story they would like to continue working on. If desired, the group can still discuss if they want to skip working on one to address the issue of potential overlaps between the stories. During this brief consolidation, the group can also summarise two stories into one if they wish to do so.

Since every stakeholder brings diverse lived experiences, to diversify the concerns elicited during the workshop, it is desired that each participant contributes at least one perspective through their stories to the activity. But if, for instance, one person did not come up with many stories they want to continue working on, they should not be forced to do so.

Originally, only two categories were used: 'Minimum Expectations' and 'Gone too far' (see Figure 5.6). Following a discussion round in the playtesting where participants agreed that the phrasing 'Acceptable' and 'Deal-breaker' is clearer, the names of these categories were changed The addition of the 'ideal would be _____' category was discussed and seen as a good way to encourage participants to think beyond minimum requirements. It was added so that participants consider not only risk prevention in the assessments, but also the positive goals they want developers to aim for in a technology.

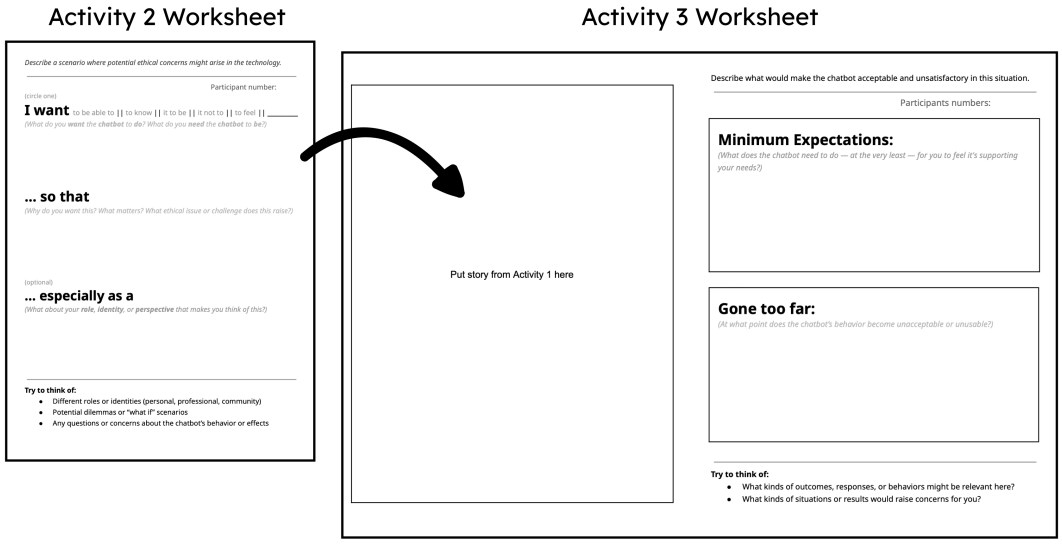


Figure 5.6: First versions of the Activity 2 Worksheet (A5) and the Activity 3 Worksheet (A4).

Adjustments for Activity 4 (Values):

During the playtesting session, only a prototype of this activity in the form of a blank cardboard was used, as shown in Figure 5.4. The decision to implement this activity on cardboard, as an inexpensive, easily available material that allows each scenario to be captured on a single physical artefact, was made to address goal (8) of Section 5.1.1. Participants struggled to interpret what the distance of a value should indicate. This led to prolonged discussions that drifted off topic. An adjusted version, Their concerns were that the spatial representation, was considered too abstract and that the distancing towards the side and the top turned out to not be of any additional value to the data.

The activity was reworked into two different versions as illustrated in Figure 5.7, where one-dimensional distancing was added. The left version allows for values to be placed into one big field. The other splits the area by the three criteria categories from Activity 3. This would allow the participants to add dedicated values to each of the criteria, potentially revealing more nuanced placements. Both of these versions were tried out in trial workshops to decide on the conclusive design.

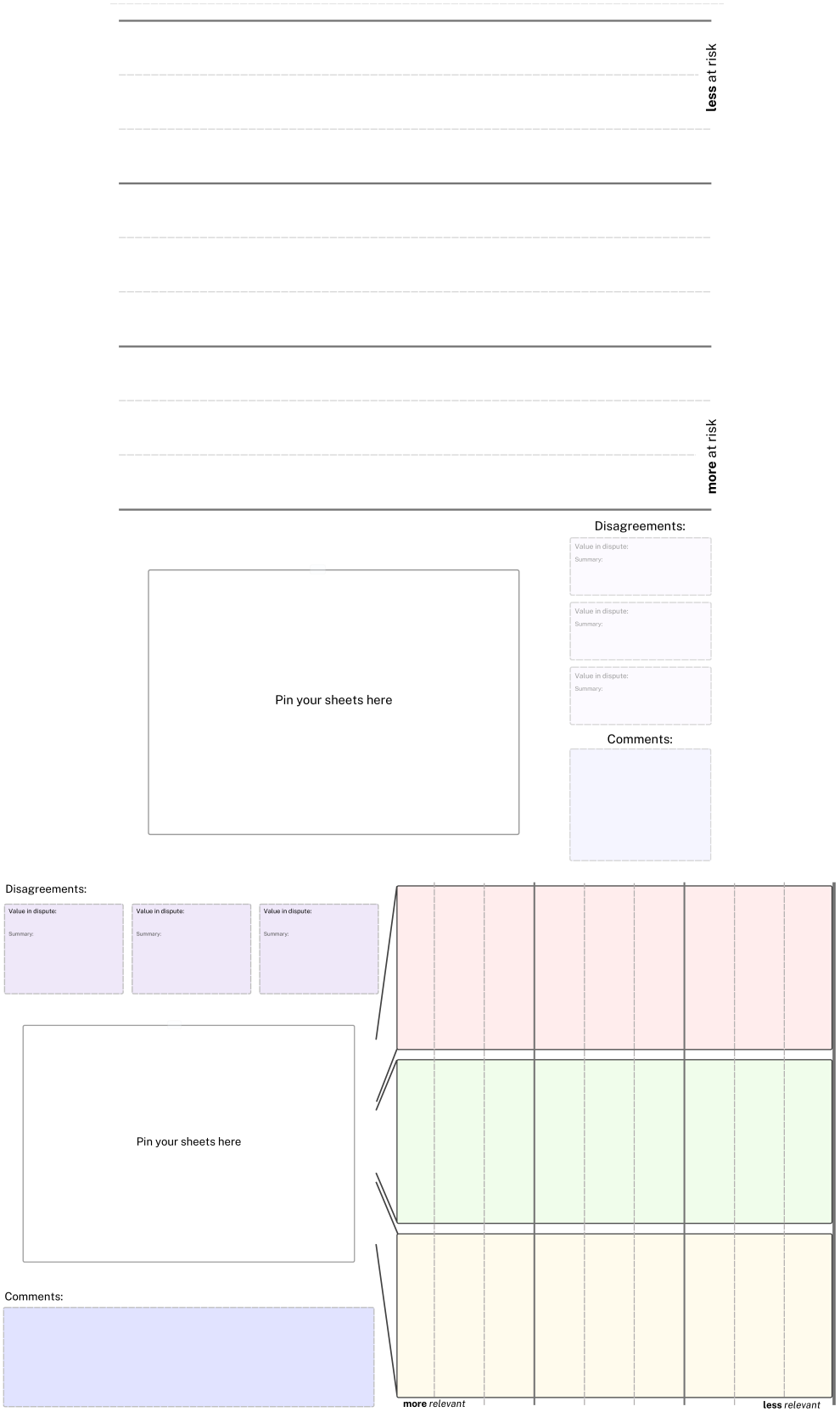


Figure 5.7: Two Activity 4 sheets tested in the trial workshop.

Another major change concerned the list of available values. During the playtesting, it was clear that the abundance of cards overwhelmed the participants. One group spent a considerable time sorting through them before even starting with their placing. Another kept the stack of values together and, one by one, discussed the placement of each value. To reduce the time it takes to pick values, it was decided to place the value cards on a separate table and spread them out. This means that participants have to move away from the board, consider the values and choose the one that strikes them as most fitting. This adds more interactivity to the task and naturally narrows the selection of values to the one that strikes them as most important. The list was also condensed to a smaller set. Many of the initial values were overlapping or closely related. For example, *autonomy* and *freedom* or *fairness* and *justice* were merged. Values that express similar concerns were also combined (e.g. *care*, *compassion*, *beneficence*). It was also noted that a few values were negatively phrased, and it was suggested to only keep the positive framing (e.g. *unfairness*). If relevant, participants are of course still able to add custom values to the board. For that, empty value cards were provided.

Overall, the playtesting provided valuable insights into the workshop materials, and the observations by the participants informed many refinements across all activities. Their adaptations were tested in subsequent trial workshops.

5.3 Trial Workshops

Following the playtesting, two trial workshops were conducted at TU Wien on July 9 and 10, 2025. Both were attended by master's students from the university. The workshops lasted four hours each. The first workshop, which had 4 participants, took place on July 9 2025. The following day, the second workshop was attended by 7 participants. The turnout led to a group at one table in the first session and two tables in the second. This allowed me to observe how well the materials work when a single workshop facilitator has more or less time to assist the groups. Several breaks were held during the session, and snacks and drinks were provided.

In contrast to the playtesting, the primary purpose of these trial workshops was to generate real participant data that could then be processed into the structured EFAs. Without this data, it would not have been possible to validate the process of data handling and structuring. This data was then used to create the webtool in Section 5.4.

When the workshop started, information sheets and data protection sheets were handed out to those who had not already filled them out in advance. Participants were then assigned numbers through drawing at the start of the session (this provided a sense of autonomy and engagement). Since the activities build on each other, the numbers are used to track the participants' sheets throughout the workshop. This is necessary because, as participants may not complete all their stories during the criteria or values activities due to time constraints, it makes sense to connect their stories.

Subsequently, general information about the workshop was communicated. Participants

Activity	Duration	Goal	Activity Type
Introduction	20’	Workshop info, agenda, technology summary	
Icebreaker	5’	Getting to know everyone	Plenary
1. Identity & Roles	20’	Own identities & roles	Individual
Break	10’		
2. Stories	25’	Define scenarios	Individual
3. Criteria	40’	Determine example (dis)satisfaction criteria	Groups
Break	10’		
4. Values	40’	Map relevant values	Groups
Wrapup	10’	Feedback, questions, conclude the session	

Table 5.1: Structure of the trial workshops.

were instructed to approach the workshop from their own perspective as actual stakeholders and not to adopt imaginary or assigned roles. To provide a realistic scenario, the technology that was assessed in these sessions was a *fictional* job employment agency chatbot currently in development (inspired by the AMS Berufsinfomat from earlier in this research).

The workshop followed a similar structure to the playtesting. Table 5.1 shows the refined times each activity took. A more detailed timetable that was used as reference during the trial workshops can be found in Figure 11 in the Appendix. After an introduction with the goals and agenda, the participants completed the icebreaker and subsequent four core activities.

During the fourth activity in the second workshop, both versions of the values cardboards shown in Figure 5.7 were tested (see Figure 5.8). Each table initially received a different version, and then the versions were swapped. At first, the preferences over either version were inconclusive and each group preferred the version they had used first. After a discussion on their upsides and downsides, it became clear that the version with the value sections split among the criteria (see the bottom version in Figure 5.7) was ultimately more confusing. All participants found it difficult to understand the purpose of separating values in this way, as any given value could be relevant to multiple criteria for the same scenario.

The idea of separating values into “more relevant” and “less relevant” was preferred, but participants concluded that all placed values would be relevant either way and suggested that a different phrasing might make more sense. They also mentioned that a value can be understood as either promoting or risking a particular scenario. While it was determined that the most significant aspect was the identification of values, this distinction could still serve as an additional data point. Based on these insights, the activity was redesigned into its final version described in Section 5.1.5 (see Figure 7 in the Appendix for a blank Activity 4 sheet).

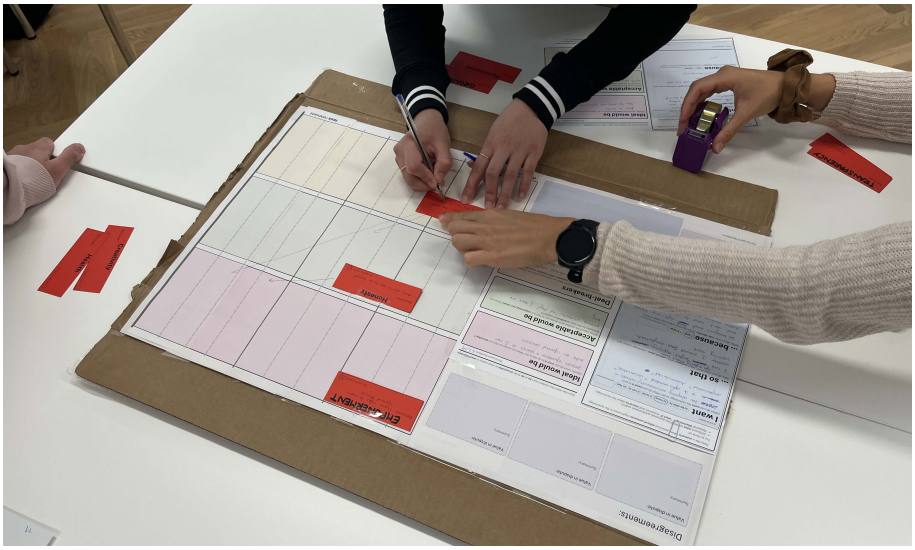


Figure 5.8: Participants working on one version of the Activity 4 sheets during the second trial workshop.

Across the two trial workshops, a total of eight boards were produced. This dataset forms the basis for subsequent analysis into EFAs, which is presented in the next sections together with the development of the tool that supports this data processing. Two example boards from the trial workshops are shown in Figure 2 and 1 in the Appendix.

5.4 Webtool

There are two options for processing data. Firstly, the collected data from the workshop could principally be manually transformed into the components of EFAs introduced in Section 4.1. However, such a manual process quickly proved inefficient and requires a substantial understanding of EFAs and how workshop data might relate to them, violating goal (1) of making ESAT available to non-experts in participatory ethics experts. Especially when the overall process should be later used by someone who adopts the entire ESAT workflow, a structured and reproducible procedure was also needed for this part. To test how this process could be supported, I created a proof-of-concept webtool. This tool is not a final product but serves as the experimental prototype to explore whether and how the transformation of the raw workshop data into EFAs can be assisted or even partially automated. It was therefore designed as a lightweight, self-hosted web application that runs locally.

The webtool was implemented using *Node.js* with *Express.js* as the backend framework and a simple web interface for user interaction. Some experimental data clustering was written in *Python*, but the ultimate clustering algorithms are also in *JavaScript*. The system includes a minimal *CORS proxy layer* (using the *cors-anywhere* package) to allow the communication between the locally hosted backend and the browser client. The

project dependencies were managed through *npm*. The implementation was assisted by the *GitHub Copilot* and by the TU Wien's *dataLAB* hosted LLM (running the GLM4.6 model). Since the tool should be accessible to users with less technical backgrounds, the decision to create a web-based interface, rather than e.g. a command-line script, was made.

The development of the tool was conceptually guided by the data resulting from the workshops and follows four main stages: *Digitisation*, *Narratives*, *Tagging and Clustering*, and *EFA Creation*. These four stages reflect the minimum steps required to move from raw workshop data to EFAs. The physical workshop artefacts first had to be digitised as a backup and in order to make them digitally processable. However, as the raw digitised data alone remained too unstructured to be interpreted efficiently or compared across scenarios, a second stage is necessary to restructures the data into a legible format without changing their meaning. To identify patterns across the data, a further stage of tagging and clustering was required. Tagging enables grouping thematic content and not only values. Finally, the structured clusters form the basis for translating the data into the components of an EFA. The stages are discussed in more detail below.

5.4.1 Digitisation

The first step involves the *digitisation* of the workshop data. Each cardboard is translated into a digital format. In the webtool, this is implemented as a simple data entry interface that corresponds to the process of how the raw data got established. This input is stored in *JSON* format and should preserve a 1:1 mapping of all the collected raw information while enabling computational processing. Developers can view the digitised boards by selecting their `.json` files. Each element of the raw data (story, criteria, and values) is preserved exactly as captured during the workshop and can be viewed digitally as shown in Figure 5.9.

This digitisation process does not require any analytical interpretation of the data. As it is neither time nor cognitively intensive, manual digitisation was sufficient at this stage of the webtool development and for the size of the dataset at hand. This step can be automated in future iterations when scalability of the ESAT workflow becomes important, e.g. by implementing image processing.

5.4.2 Narratives

The digitised workshop data is structured for documentation and not directly for processing. While the digitised boards theoretically capture all workshop information, it remains difficult to interpret. To address this, the second part of the webtool transforms the entries into *narratives*. They are textual summaries that represent the same data in a more readable form. This format is also beneficial for auditing purposes, where it provides a clear and reportable representation of the concerns that stakeholders came up with. The narratives allow the developers to review and make sense of the ethical concerns and expectations the stakeholders had come up with during the workshops,

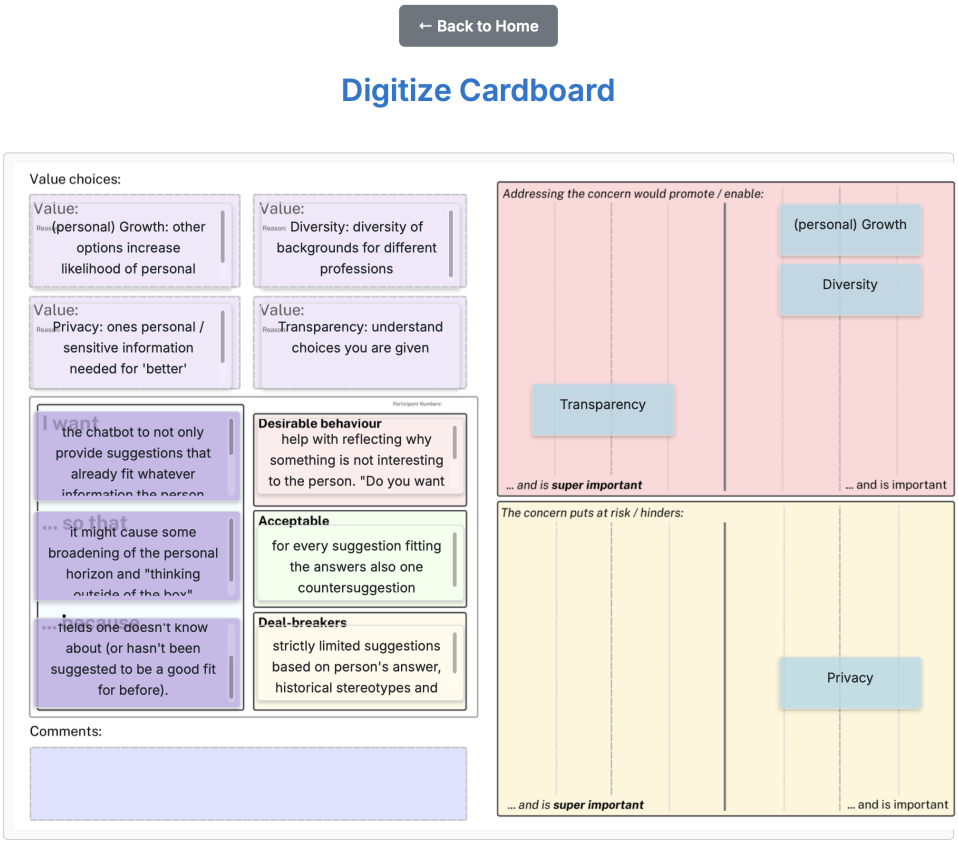


Figure 5.9: A digital representation of one board which was created during a trial workshop.

without having to interpret individual sheets or boards manually. The intention behind this step is still not about interpreting any data. The following sections describe how that can be done manually or with the assistance of AI in the webtool.

Manual narrative creation:

In manual creation, developers can fill in a predefined narrative template directly in the user interface (see Figure 5.10). The template provides a consistent structure and can be freely edited before being saved as a `.txt` file.

AI-assisted narrative generation:

The second approach uses the TU Wien self-hosted LLM to generate narratives automatically. At the time of development, the model in use was `http://openai/TechxGenus/Mistral-Large-Instruct-2411-AWQ`, accessed through the university's *Aqueduct* via an OpenAI-compatible API endpoint.

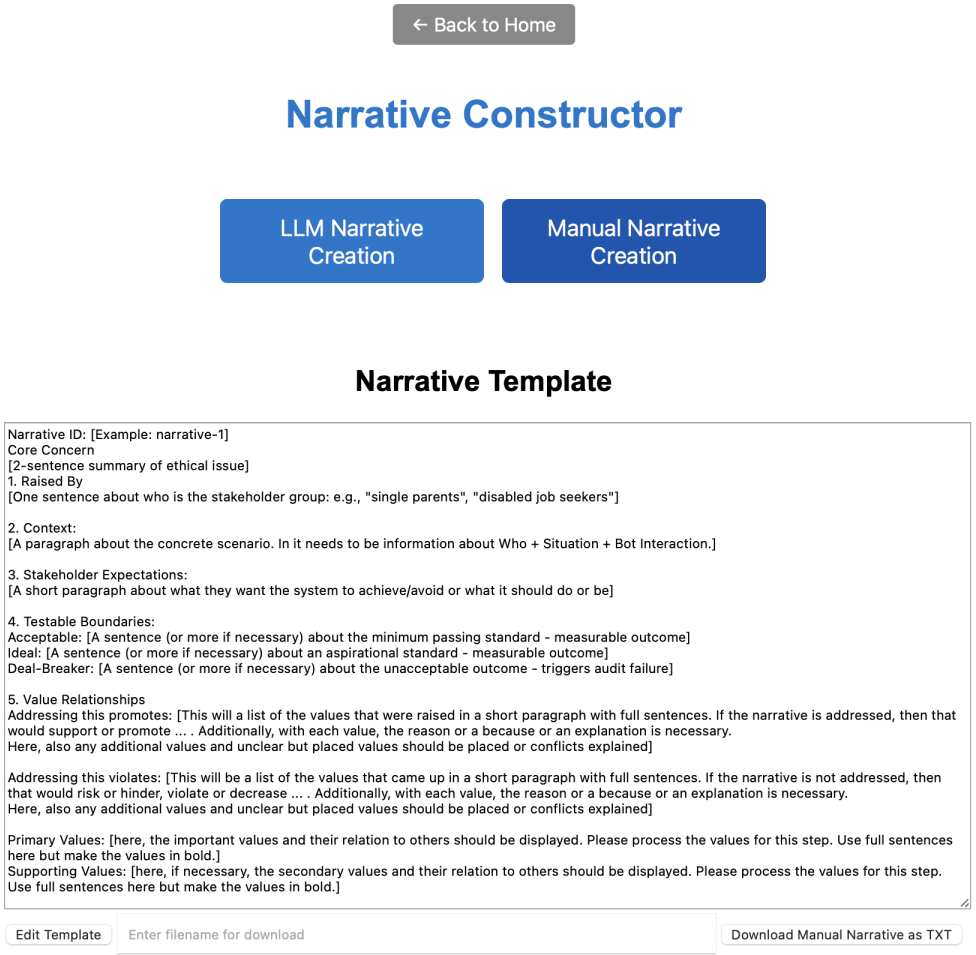


Figure 5.10: The manual narrative tool with its template.

In my preliminary testing, the AI-assisted narrative generation performed reliably for this stage of the workflow. The model generally followed the provided template and prompt structure, producing outputs based on the input files and not hallucinating. While this cannot be guaranteed in all cases, it suggests that such models could support the task of summarising the qualitative data in a structured way. For an example narrative that was generated from workshop data see Narrative (Workshop 2 Board 5) in Appendix I. Example Data.

As LLMs typically require a large amount of computational resources [110], I contacted the operators of the TU Wien *dataLAB* to get details about its power consumption. They reported that the models run across multiple instances, depending on the demand, from one to four GPUs per model. Although they do not monitor the energy consumption precisely, they were able to inform me that the average consumption over a seven-day period was between 100-130 W (idle) and 400-500 W (under load). They reported that

requests are processed in batches to improve the efficiency when multiple users are using it simultaneously. During this reported week, approximately 37 million tokens were processed by all users of the TU Wien LLM. Although these figures are approximations and depend on the workload, they are a useful indication of the energy the system uses and whether it might be advisable not to rely on such systems. See Luccioni et al. [110] for a reference on the footprint of LLMs.

To conclude, the potential for LLMs to generate unintended text and their environmental footprint made me question whether such an approach should be the recommended method for transforming the data in actual applications of ESAT.

5.4.3 Tagging and Clustering

The next step of the process adds an analytical component to the data processing. Its goal is to identify recurring ethical concerns and value relations across the various outputs the workshop provided and group them into meaningful clusters. This stage thereby transforms the digitised and narrative data into the structured form used for generating EFAs.

Before developing this part, I manually performed initial clustering and derived EFAs from them. This served as the baseline to later compare and evaluate the algorithmic clustering results generated with the webtool.

While the boards already contained value labels, clustering and creating EFAs solely on shared values was determined as insufficient as different stories could reference the same value for different reasons, and similar concerns could involve different values (e.g. in case the stakeholders identify differing values as relevant). Tagging therefore provides a lightweight way of capturing the contextual similarities and thematic aspects between the data without. It then enables clustering by similarity of content or even stakeholder roles on top of the values. A developer can assess the content of the boards and annotate them with *tags*.

Tagging

To tag, the narratives' `.txt` and `.json` files must be loaded into the webtool. The user can then annotate each narrative with supplementary tags. For example, stakeholder groups, interaction categories, or advice types can help describe the context or focus of a narrative.

The tags received directly from stakeholders during the workshop are imported automatically into the webtool. Since they represent the participants' expressed ethical positions, they should not be overwritten or reframed by subsequent users of the tool and are therefore treated as *immutable*; they cannot be deleted or modified by the user. The user can only add further descriptive tags. Each tag is associated with a *weight*. These reflect their importance and follow the categories introduced in Section 5.2: stakeholder-assigned weights which were mapped as 'important' receive the weight 2, and 'super important'

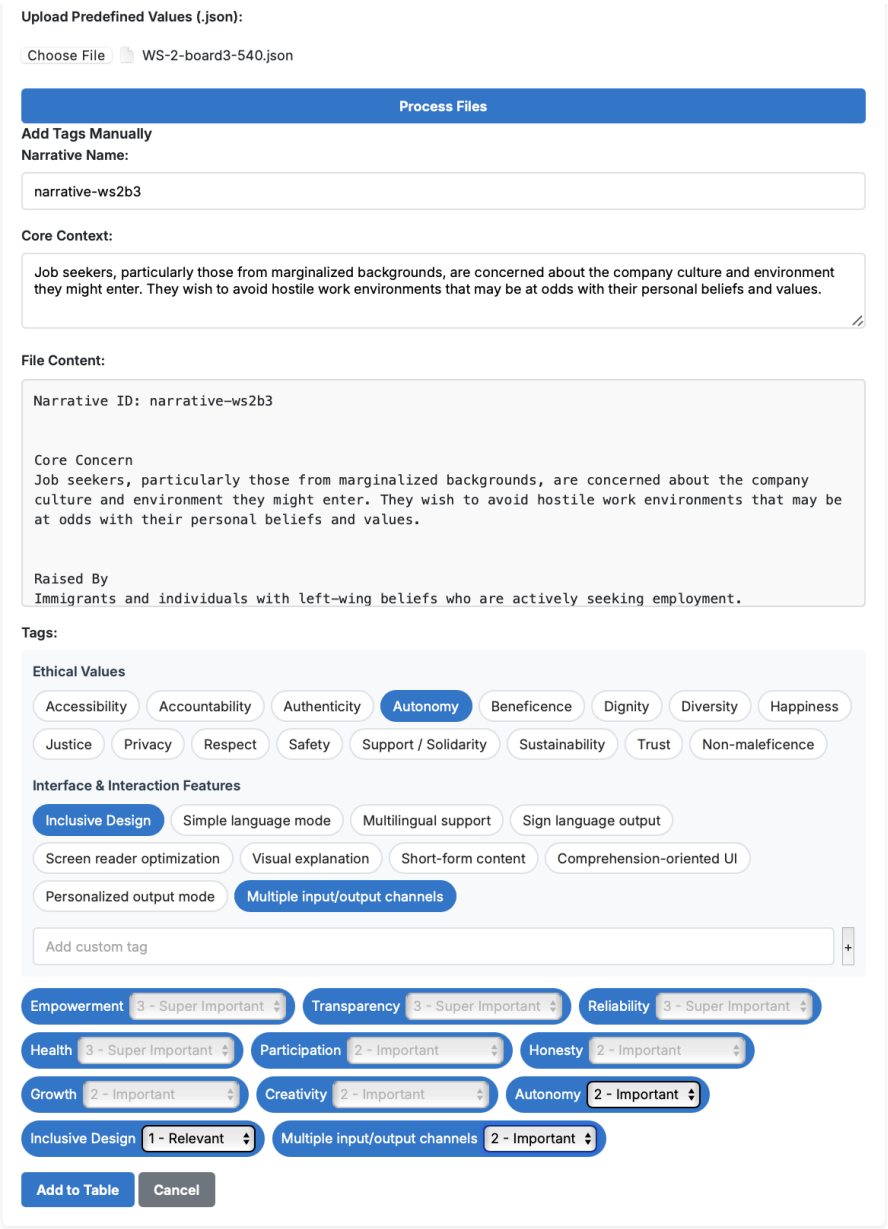


Figure 5.11: The tagging tool. Selected are both stakeholder-identified and user-added tags.

are mapped to weight 3. Since the stakeholder-provided data represents the original lived experiences expressed in the workshop, it is always weighted more. This aims at treating the stakeholder's values as the primary source of ethical concerns and prevents developers from overriding them. A user of the webtool can therefore only add tags with weight 'relevant' (1) or 'important' (2). This process is shown in Figure 5.11.

Clustering Algorithms

With data annotated this way, the narratives can now be pictured and compared in a plane view of the webtool. In an interactive visualisation, the users can look at the relationships between each piece of data. This allows users to see if and which clusters emerge based on shared attributes.

To explore the clustering, several approaches were implemented and tested within the webtool. Each algorithm is its own JavaScript function. This allows new ones to be added efficiently and to be switched between using a dropdown select list. Ultimately, four different clustering strategies were selected to be implemented:

1. **Connection-based clustering**: This approach clusters the narratives solely on the number of shared tags between them. The more overlaps between narratives, the stronger their connection. An example can be seen in Figure 23 in the Appendix.

2. **Weighted-connections**: This algorithm represents the weighted relations between narratives. Its function calculates the connections based on their shared tags and corresponding weights. For each pair of narratives, the script sums the numerical weights of all tags they have in common. These scores are then normalised and mapped to the line and its colour and thickness. Higher scores (i.e. more and stronger shared tags) result in thicker, warmer, and redder lines. Hovering over a line reveals the shared tags and the computed similarity score via a tooltip. An example is shown in Figure 5.12.

3. **H-index inspired clustering**: This approach is inspired by the bibliometric *Hirsch-index* [41] (H-index), which identifies narratives that are not only highly connected but also linked to other strongly connected entries. This forms clusters around central influential entries.

   To implement this, the function computes the local h-index for each narrative node. It represents the largest number h such that a given narrative has at least h connections with a similarity score equal to or greater than h. It first calculates all pairwise connections based on their overlapping tags and average weights. Then, it derives the h-index for each narrative and uses the values to group nodes into clusters. The resulting visualisation highlights dense regions of a dataset. For each of the clusters, a different colour is shown. When a lower h-index is selected, only one large cluster exists. The index can be adjusted through a slider. Figures 25 and 26 in the Appendix show an example with an h-index of 7 and 8 for the trial workshop's data.

4. **Stepwise connection**: The last algorithm first constructs a network of all pairwise connections based on shared tags and their average weights. From this, it selects a *seed edge* (the strongest initial connection between two narratives that share a sufficiently large number of high-weight, as in 2 or higher, tags). This then forms the core of the cluster. Then, the algorithm searches for additional narratives that

share any of the tags associated with this seed connection. These narratives are then iteratively added as *second-order* nodes to let the cluster grow.

Visually, the first-order connection is rendered as a thick red line. Second-order connections are from the midpoint of the first-order one shown in orange, and their thickness is proportional to the strength. Hovering over the lines reveals the similarity score and the shared tags of the connection (see Figure 24 in the Appendix).

Since this webtool serves as a proof-of-concept, the algorithms were deliberately kept simple but effective. Each of them offers a different perspective and allows users to explore and analyse the data. They can switch between the modes and directly observe the differences in connections.

Compared to the manually created EFAs, the weighted clustering algorithm matched the selected narratives most closely. But, as with all ethical considerations, one cannot fully rely on automated processes. It is still required to make sense of the data and manually look for matching data entries. The algorithms should only be considered a starting point in forming EFAs. Once clusters are identified, the user can directly select a set of narratives from the visualisation to initiate the creation of a new EFA.

5.4.4 To EFAs

In the final step of the workflow, the user creates the EFAs. Since clustering identified which narratives represent similar underlying concerns and values, they can be grouped together. After selecting these narratives, a popup opens which guides the user through several steps. The interface shows the detailed information about the narratives. It requires the user to add the EFA ID, a title, and a description. Additionally, the core values of the EFA have to be selected.

Next, the testable criteria must be added. These should correspond to the narratives' criteria. They need a description and a "How to test" to be filled out. Additionally, each criterion requires a "Status" ("not tested", "passed", or "failed" are the categories that enable its auditability) and a "Requirement Type" ("must do", "should do", "must not do") which directly map to the categories from Activity 3. Lastly, the EFA receives a weight and a priority ranking. These characteristics make sure that EFAs are *auditable* artefacts that can be tested and kept track of over time (see Section 2.4.6 on auditability and testability). This process is illustrated by Figure 22 in the Appendix.

When the process is completed, the EFA is stored as a `.json` entry which includes metadata that links it back to the source narratives and tags. Users can view the EFAs in a table, together with each EFA's number of tests and associated narratives, shown in Figure 5.13. This step represents the final synthesis of the qualitative raw data of the workshop into the structured digital one of an EFA.

5.4.5 Report

To conclude the workflow, the webtool includes a simple report functionality that compiles the EFAs and their data into a single printable PDF. This feature was primarily added to demonstrate the potential for an easily generated report. This report can, for example, be shared with shareholders or regulatory bodies that have an interest in understanding design decisions or reconstructing the extend of ethical considerations in a system. Such a report can support the communication of the assessment and auditing of the technology.

5.4.6 Functional Limitations

As a proof-of-concept, this webtool implements only the core features necessary to validate the prototype. It is not optimised for user interaction or scalability. For example, some bugs can appear during the plane view and entries can currently only be created but not edited or deleted once stored. As mentioned above, the purpose was not to create a finished application but to showcase the feasibility of the digital processing of the raw workshop data.

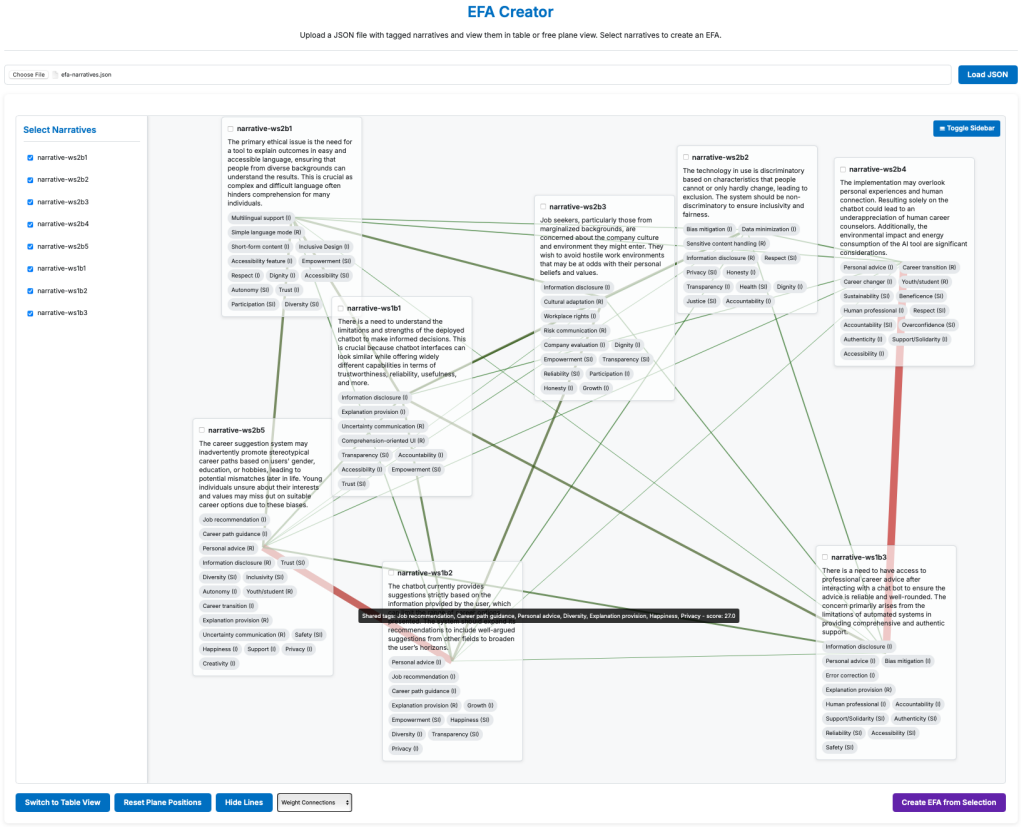


Figure 5.12: Plane view with weight-connection clustering selected.

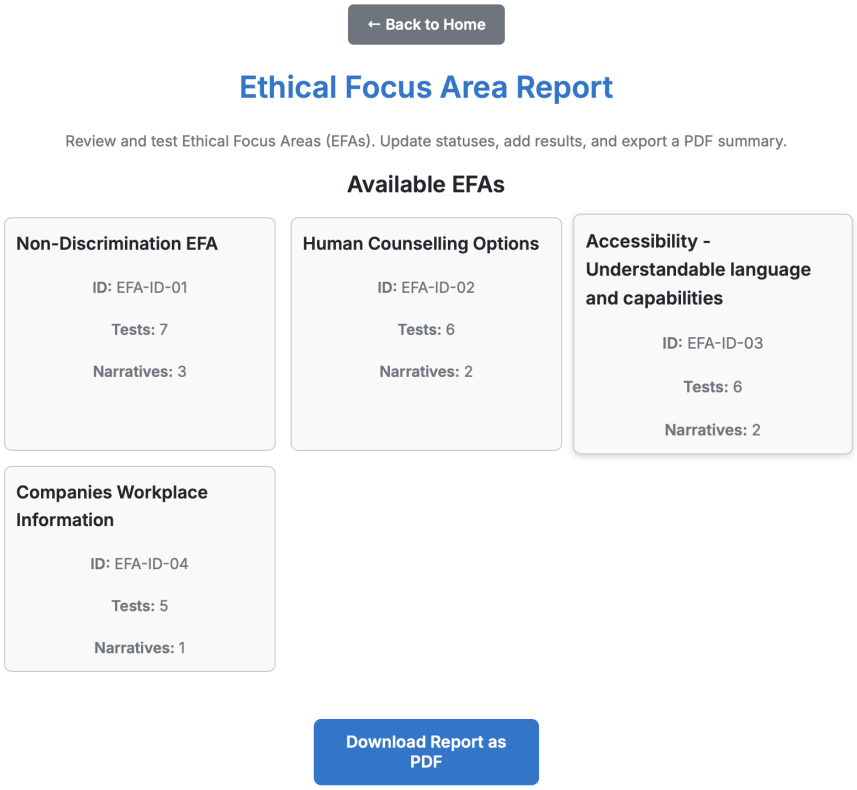


Figure 5.13: Overview of EFAs with the option to download a report. The EFAs can be opened to show their tests and narratives.

5.5 Summary

In this chapter, I presented the design and development of the two components of the ESAT blueprint: the participatory workshop and the accompanying webtool. The illustration of the full ESAT workflow can be seen in Figure 5.14. It visualises the entire process, from stakeholder engagement to the generation of EFAs and a report, and shows all the steps that both the workshop and the webtool encompass.

The design process revealed practical challenges of translating abstract ethics into more concrete and processable data. But through playtesting and the workshops, I was able to refine the materials and iteratively develop the webtool that assists in handling and structuring the data. Together, the workshop and the tool make up a clear and replicable approach to assess the ethics of a system. To further lower the threshold for this method to be applied, I compiled all materials and visualised the entire workflow in a flowchart in the subsequent chapter. This 'packaging' aims to communicate the findings more effectively and thereby makes the ESAT process accessible to practitioners.

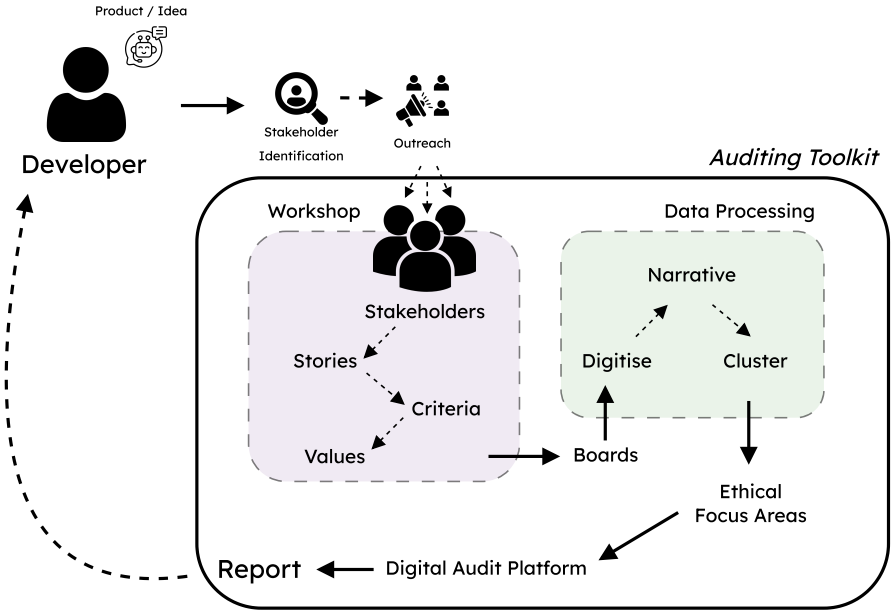


Figure 5.14: The detailed diagram of the components and their steps of the ESAT Tools.

CHAPTER 6

Discussion

This chapter discusses the main outcomes and contribution of this thesis. It first summarises the components of the Ethics Self-Assessment Tools (ESAT) workflow and then reflects on the broader implications of this work by revisiting the research questions and discussing learnings and limitations. Lastly, it evaluates how the tools respond to conflicting or hostile stakeholder participants.

6.1 Contributions

The developed workshop format and the accompanying webtool are the central practical outputs of this thesis. In line with goal (1) (see Section 5.1.1 for goals identified from the literature insights), a structured script to support the application of ESAT in practice in the form of a flowchart that guides their implementation step by step was also designed (see Section 6.1.2). Together, these three components make up the contributions of the ESAT.

6.1.1 Examples from the Trial Workshop

To demonstrate how the workflow translates the participants' concerns into structured outcomes, one of the outputs of the trial workshop is revisited here. In the workshop, stakeholders elicited concerns towards a career-advising chatbot. Figure 6.1 shows a completed board that captures a group's concerns about potential career suggestions that follow stereotypical career paths based on a user's profile or shared characteristics and interests. The group considered it important that the chatbot should be tested on this. Their data was processed through the webtool into a narrative (see Appendix Section 7.2) and subsequently transformed into the Ethical Focus Area *EFA-ID-01: Non Discrimination*, shown in the top left of Figure 5.13. The resulting EFA includes seven tests, each derived from the testable criteria statements collected across three boards that inform this EFA. Example tests include:

- Test 3: *Validate that the system does not store user profiles in its database and only bases its answers and suggestions on information a user provides in the chat.*
- Test 6: *Prompt the chatbot for job suggestions in a variety of different fields. When receiving an answer with suggestions, validate that the system provides explanations why it included and excluded certain answers.*

While the workshop and webtool show that the ESAT workflow can effectively elicit and transform the stakeholder's ideas into concrete EFAs, it also needs a concise and clear overview of the process that supports anyone who might conduct the workshop. A flowchart was therefore designed to summarise the entire workflow into a structured and visually clear format.

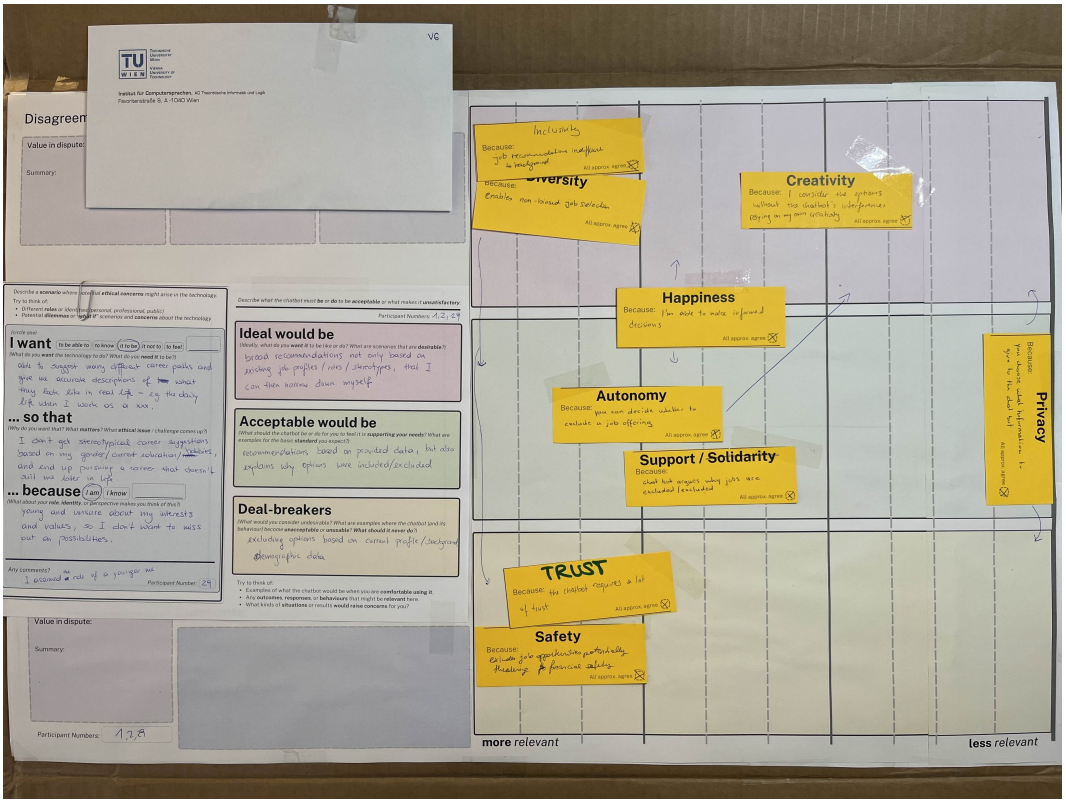


Figure 6.1: Board created during the second trial workshop on July 10, 2025. Further examples are provided in the Appendix at Figure 2 and 1.

6.1.2 Flowchart Script

The flowchart was inspired by the Decision Tree for the Responsible Application of AI [5] by the AAAS Center for Scientific Responsibility and Justice . The ESAT script serves

as the main accompanying document for the entire workflow. It is designed across two A4 pages that are for illustrative purposes shown in Figure 6.2 (a full-sized version can be found in Figures 12 and 13 in the Appendix).

Over the two pages, a structured and adaptable overview is provided that allows the workflow to be implemented with minimal training. The first page introduces the purpose of the flowchart, outlines who it is for, and provides some logistical reminders for preparing a workshop session. It also starts the workflow with the *to-be-assessed* technology (or its idea or prototype) and includes important considerations for stakeholder identification and outreach. The flowchart encourages the developers to draw on established methods in determining stakeholders (e.g. Actor-Network Theory [136] or Direct/Indirect Stakeholder Analysis [57]) and guides in inviting them as participants to the workshops. The second page continues with a step-by-step process through the entire workshop and its individual activities, their corresponding materials, and expected outputs.

In addition to outlining the overall script and information on the webtool for data processing, the second page of the document includes general instructions for conducting the workshop. There, the ground rules and conduct for the participants and a section on managing conflicting situations during the group activities are described. This was included because discussions of ethical concerns naturally surface disagreements, and conflict itself is often an important source of insight rather than something to avoid. The workflow therefore provides facilitators with strategies for de-escalation and redirection to make sure participants can contribute equitably. The instructions aim to create a safe space for viewpoints and reduce the risk of dominant viewpoints in the stakeholder data.

6.1.3 Summary of Contributions to Research Fields

The central contribution of this work is in closing the gap in ethics technology assessments between high-level principles and concrete, practical methods summarised in Section 2.5. This work contributes to three overlapping research areas:

- **Human-Computer Interaction (HCI)**: Through its participatory and reflexive method, I showcase how ethical concerns can be translated into structured auditable digital artefacts. By showing that ethical reflection can be guided through iteratively designed workshop materials and collective exercises, it adds to the existing HCI methodologies in participatory ethics approaches. This responds to the limitations of participatory ethics approaches which often lack concrete methods to translate stakeholder's reflections into structured outcomes.
- **AI Auditing Literature**: This thesis addressed the gap between conceptual frameworks and practical methods in ethics auditing. Additionally, it answers issues introduced in Section 1.1.1: *What*-problem (what values should be assessed) by enabling stakeholders to define them directly, and the *How*-problem (how these values can be made testable) through the introduction of Ethical Focus Areas as the structured auditable datatype. It positions auditing as a continuous participatory

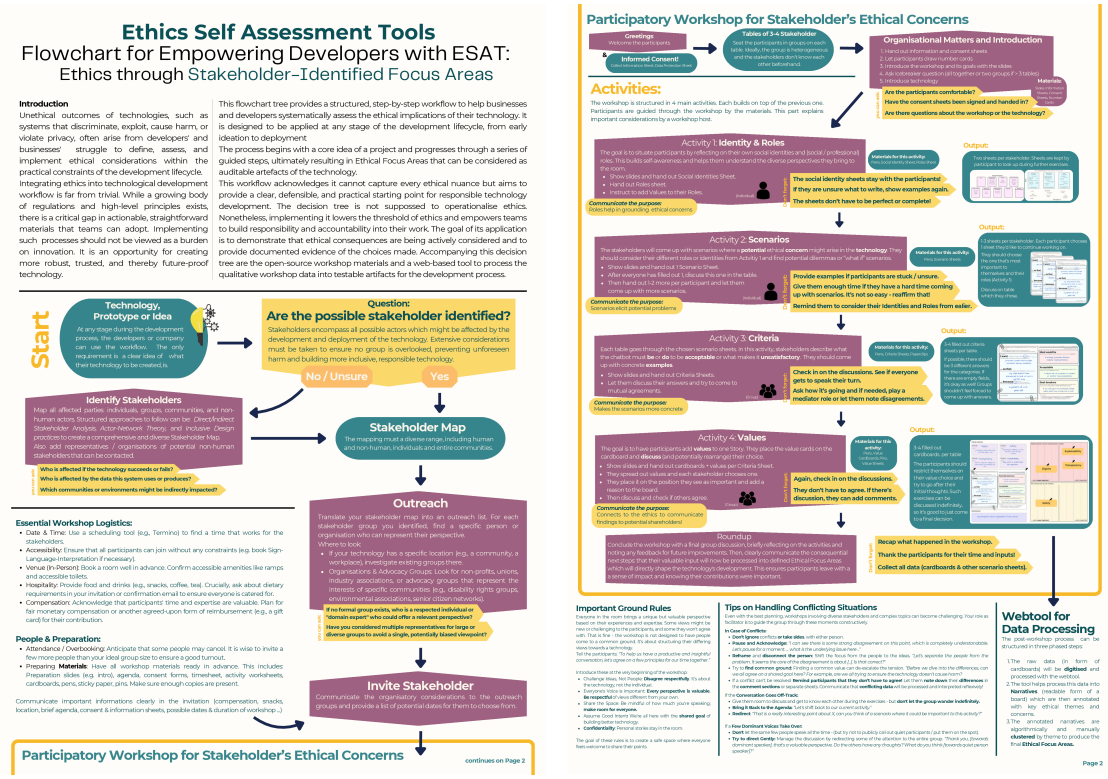


Figure 6.2: Workflow Document of the ESAT process. For a larger version, see Figures 12 and 13 in the Appendix.

process and expands on approaches that treat audits as static, one-time compliance checks.

- **Ethics Technology Assessments / Responsible AI Research**: Finally, the work contributes to the broader field of ethical / Responsible AI by demonstrating how reflexivity and a stakeholder dialogue can be formalised through self-assessment practices. By not assuming universal metrics or relying on metrics and benchmarks, it places the ethics assessments as a socio-technical process that acknowledges context and discussion.

6.2 Reflections

In this section, I will reflect on the outcomes and implications of this research. It revisits the research questions that guided the thesis and discusses the key learnings from developing the ESAT workflow and the extent to which the designed materials can be generalised. It also outlines the limitations of the work.

6.2.1 Unfolding the Research Questions

The guiding questions of this thesis asked what methodology developers could employ to translate stakeholders' ethical expectations towards an AI into testable artefacts. Two sub-questions specified how such a method might be designed to support non-experts in ethics and to what extent the materials could be transferred beyond the advice-chatbot domain (see Section 1.2.1)

The main question **RQ** was addressed through the development of the ESAT workshop and webtool. They combine participatory practices with a digital structuring process. The entire ESAT workflow demonstrates how stakeholder concerns can be elicited, collected, refined, and translated into Ethical Focus Areas with concrete testable criteria. Thereby, the thesis provides a methodological answer to how ethical expectations can move from the abstract to an operational level. The concerns of the stakeholders are not just values but are discussed and formalised as actual artefacts for evaluation and documentation.

The first sub-question **RQa** asked how the design should empower developers to conduct such ethics work without becoming or employing ethical experts. The ESAT workflow responds to this through workshop materials which are designed for participants to work through largely on their own, with minimal expert guidance. It also provides the supporting facilitation materials for whoever conducts the workshop, some examples, and the guiding flowchart. Developers are positioned simply as the coordinators of the workshop (including stakeholder identification and outreach as page 1 of the flowchart script describes) and processors of its resulting data, supported by an easy-to-use webtool. Overall, this lowers the threshold for meaningful ethical assessments of their technology and reduces the need for external consultation. It also allows the ethics work to be embedded directly into development practices.

The second sub-question **RQb** is concerned with the transferability of the materials. While the prototype was developed around a chatbot, the ESAT structure of eliciting concerns, criteria, and values, and then transforming them into testable EFAs, is not limited to the domain of advice-giving chatbots. The workflow is intentionally adaptable as its specificity is not predefined and emerges through applying it. In that sense, it explores a flexible method that supports a situated ethical reflection wherever participatory engagement is possible. Its underlying structure can be applied to other AI systems and technologies, or even non-technological contexts which might require ethical deliberation.

Transferability of ESAT

In practice, ESAT's structure could be applied to other participatory or evaluative settings and even outside of technology development. It also has the potential to be applied to policy design or educational contexts. In AI policy, authors have already highlighted a need for participatory methods that incorporate perspectives of stakeholders in a workshop that "practices inclusivity in AI" with the goal of not only participant-washing policy-making processes [134]. Similarly, Barnett et al. [13] envision stakeholders

identifying and ranking negative impacts of AI to assist policymakers. The ESAT workflow could be adopted to such a context as well.

Regarding education, it has already proven useful in a workshop at the TU Wien which explores science and technology studies (STS) topics in STEM subjects. I facilitated a workshop on November 20, 2025, in which 10 students and 15 lecturers from 11 university institutions shared their expectations and concerns about university education from their respective perspectives. The materials for the workshop were adapted from the original ESAT workflow, but followed a similar structure: the first activity situated the participants in their roles and interests in participating in the workshop. The second followed the structure of Activity 2 ("I want, so that, because") and allowed them to raise their wishes as scenarios towards attending university as a student or teaching. In the third, similarly to the criteria in Activity 3, the participants collaboratively worked on making their wishes more concrete by coming up with implementations. Figure 14 in the Appendix shows the adapted Activity 2 and 3 sheets. However, the data created by the groups in these activities was not processed into EFAs. As the final activity, participants transformed their elicited wishes and concrete implementations into *Unitopias* on flipcharts (see Figure 15 in the Appendix for two examples), which, similar to EFAs, capture the underlying participants' concerns into tangible artefacts. While not focused on ethics per se, the process in this workshop followed a similar structure to the ESAT workshop, including the roles, scenarios, and criteria activities. This suggests that the participatory structure developed in this thesis could make for a general framework.

6.2.2 Learnings

Beyond the material contributions of this thesis, several learnings about participatory ethics and the design of the methods that make them practical were made. The entire process, and especially the workshop, showed that ethical reflection can be structured and guided without being prescriptive. Furthermore, methods can be designed as reusable and low-resource materials, and by creating explicit scripts for them, ESAT provides a method that has potential to enable non-experts in assessing the ethics of their technology. But simply through participation, a workshop does not automatically lead to inclusive outcomes. Conflicts, misunderstandings, or even hostile contributions are part of a participatory process. How the design deals with these is described in Section 6.3. This work also showed that ethics in AI can be approached as an interactive design problem. While other approaches often aim for static principles or policies that aim at ultimate rules, I focus on creating a structured method to continuously assess and revise a technology on its ethics. Through the workflow, a repeatable and adaptable way of translating value discussions is provided.

6.2.3 Limitations and Future Work

The workshops were conducted with only 11 TU Wien students as participants and only on one technologically specific context. As such, the findings and resulting EFAs

do not claim to be generally valid and serve only as an example of assessing a specific technology.. A broader application across cultural settings and through different teams would be necessary to evaluate the consistency of the methods.

Another limitation comes from ESAT being a *self*-assessment and depending on motivated cooperators. The workflow assumes an interest among a company or developers to include ethical reflection and assessment as part of their product development and maintenance. In practice, such processes require time, resources, and importantly, the understanding and acceptance that ethics assessments are a means to make a technology better. This requires a general acceptance of assessment and auditing practices for the systematic problems of it being perceived as a hindrance.

Towards a Toolkit

The current version of the ESAT workflow should therefore be seen as a proof of concept and not a fully developed toolset. The components were designed intentionally to test whether ethical assessments could be made accessible through a workshop. It does not aim at functioning as a finished product, and further iterations are needed. For example, the webtool is simply a prototype which was created to showcase how data processing could be approached.

It is doubtful whether a single, definitive ethics auditing *toolkit* could be defined. Ethical assessments are always context-dependent and shaped by the practices of those involved. Any attempt at creating *the* toolkit would risk oversimplification. Equally, ESAT should also not be considered a fully developed auditing toolkit or framework. It can be considered a methodological prototype that explores how ethical reflection can be integrated into technology development but is not intended to be an all-encompassing solution to assessing technologies.

The goal of this thesis was not to deliver a finished product that can directly be adopted or marketed. It simply investigated whether testing ethics could be supported through a structured approach and demonstrated one possible way of doing it in a participatory method. A more comprehensive auditing toolkit would require more long-term research. A brief outline for an expansion of the workflow's features is provided in Section 7.1.1.

6.3 Reflective Evaluation of the ESAT Tools

Similarly, a large-scale validation of the outcome of this thesis would be necessary before it can be considered a full toolkit. An extensive case study would have been the next step in this research but would have required more time and cooperation with developing teams. A design for a potential case study as a separate research project is presented in Section 7.1.2. However, in scope was critically reflecting on the functionality of the workshop and webtool and their behaviour in specific contexts, especially beyond ideal workshop settings. Through a reflexive simulation of 'hostile stakeholder' I explore how they might work and respond to malicious participants, outlined in the following sections.

While the workshop and the accompanying webtool were shown to function cohesively together, this section examines how the ESAT workflow behaves when it is confronted with perspectives that are in conflict with each other or contradict inclusive or democratic ethical assumptions. To explore this, I carried out a brief reflective simulation of 'hostile' stakeholder scenarios which served as a conceptual evaluation.

6.3.1 Background: Value Conflicts and Hostile Participation

As the workshop is designed to surface ethical concerns through participatory engagement, a tension naturally comes up: participation does not automatically guarantee ethically *desirable* outcomes and that values are not inherently benevolent [98]. Diverse stakeholders bring with them many, sometimes conflicting, lived experiences and moral judgements. These might also contradict with inclusive or democratic ideas, the base of participatory research and ethics itself.

Recalling the motivation of this thesis, the Berufsinfomat chatbot interacts with the broad public of people who want to be informed about work in Austria. Such a technology necessarily encounters people with varying social, political, or moral backgrounds. This might also include people who hold reactionary or exclusionary views. Additionally, participants might also act in bad faith, for instance by trolling or provocatively derailing a discussion, which differs from sincere but conflicting viewpoints but still tests the workshop and how it deals with hostility. For example, users might explicitly reject the gender-inclusive language others desire in the chatbot or request job recommendations with traditional gender roles. The question then arises if ESAT, and in particular the workshop materials, treat these inputs as 'legitimate' or ethical stakeholder concerns. Or should they potentially be filtered out if they are ethically problematic?

To approach this question, I draw on *feminist standpoint epistemology* and its related theories of *situated knowledges*. Scholars like Nancy Hartstock and Sandra Harding argue that knowledge is socially situated and shaped by the power relations people are subject to [70, 71]. They propose that people who are structurally marginalised (for example, through gender, class, or race) are often in a better position to see and understand systematic injustice. This is because they experience their own oppression but are also simultaneously navigating the dominant social order which sustains it. As Patricia Collins writes, this can be extended to *intersectionality* where 'overlapping' systems of oppression produce even more distinctive positions (e.g. individuals who are marginalised by both race *and* gender) [27].

Donna Haraway's *situated knowledges* further challenges ideas like neutral objectivity and argues that all perspectives are context-dependent, partial, and *embodied* (i.e. that the knowledge also comes from the lived and material positions bodies hold in society) [69]. While every participant's standpoint, if earnest, is valid as their expression of lived experience, not all perspectives would contribute equally to an ethical understanding of the sociotechnical system. As some may defend or even reproduce existing structures of domination, these scholars therefore caution against assuming that 'all perspectives

should be treated as equal'. From such a lens, I therefore question whether it makes sense that the participatory process in the ESAT workshop should simply treat all stakeholder inputs as equivalent data points. If perspectives shaped by structural marginalisation reveal ethical oversights, perspectives that are rooted in privilege might simply complain about a general social change and not really focus on an ethical harm.

To provide a very binary example: a male participant who feels 'excluded' by gender-inclusive job recommendations may be articulating a real emotional response, but that is one that comes from a perceived loss of social dominance instead of structural injustice. If such calls were treated as ethically equivalent to, for instance, a woman's claims of being systematically excluded from opportunities on the job market, it could reproduce the inequality and not address it.

Furthermore, Quandt explicitly recognises ethically harmful participation as the concept of *dark participation* [140]. They describe how systems which were originally designed as democratising spaces can be appropriated for destructive (i.e. manipulative or hateful) purposes. While their work focuses on online journalism and comment cultures, it showcases that democratisation alone (through participating) is not inherently emancipatory (as in liberating people from oppressive systems). While early participatory ideals assumed that every added voice would strengthen the democratic conversation, dark participation shows that this is a utopian idea. The author further argues that it can even be a vehicle for anti-democratic expression through trolling or being "authentically evil". According to Quant, democratic systems require a form of *normative boundaries* and *safeguards*, for instance, some which reject hate speech.

To clarify, I do not want to imply that any perspectives which might conflict with others should be disregarded. They simply should be interpreted differently by paying attention to their *positionality* and the relation to power they hold. To address this in participatory approaches and tools, apart from eliciting *what* is being said, they should also be sensitive to *who* is saying it and *from where* (from which standpoint) and *what social consequences that brings*. While ESAT is designed for adoption without specialised ethics training and lowers the barrier for conducting ethics assessments, this indicates that a minimal degree of reflection skills, as is presumably the case with most ethics-related methods, remains necessary. Without this, a reflexive interpretation is limited and risks eliciting hostile or malicious positions as legitimate ethical requirements.

To translate all these insights to the participatory ethics tools, ESAT must balance an ethical openness and accountability. On the one hand, the workflow seeks inclusivity where all the stakeholders' perspectives are respected, even if they are conflicting. At the same time, a 'moral neutrality' could risk legitimising values that might undermine the emancipatory aim of participatory ethics.

Rather than excluding the potentially 'hostile' perspectives, a more reflexive approach would make them visible in the data. This data could then be critically interpreted and not directly adopted. In that sense, the ESAT workflow would go beyond just gathering stakeholders' ethical concerns and values towards a technology. It would

incorporate an interpretation of the conflicting concerns and values by not assuming that these are all inherently neutral or 'good'. This interpretation requires examining which concerns should inform Ethical Focus Areas and in what way, while maintaining reflexive awareness of positionality and power. Especially when the workflow is to be used by development teams without training or experience in participatory workshops and interpreting conflicting ethical data, some guidelines are required to orientate oneself by.

6.3.2 Adversarial Perspectives in Participatory Methods

Participatory design traditionally assumes a cooperative mindset, which results in the comparatively scarce literature that engages with hostile or ethically adverse participants. They are usually treated as willing to constructively contribute to designing a 'better', more usable, inclusive, and just system. But in practice [42], if seeking to democratise the stakeholder involvement and represent a wide range of people, not all may act in good faith, and some might introduce conflicting or reactionary moral standpoints that may contradict with inclusive ones.

Some approaches deliberately adopt "wrong" or adversarial perspectives as a reflective or diagnostic design practice. In software security, the concept of *misuse cases* exists as the evil twin of the use case to model how a system might be attacked [156]. Also in cybersecurity, *adversary personas* are used to ask who opponents might be and what their goals are [119]. Similarly, *abuser stories* come from user stories and are a means to let developers think about potential malicious use of a system, like hacking [135]. In criminology and design for crime prevention, criminal personas have been investigated to identify vulnerabilities in a system [74]. They situate designers in adversarial roles as criminals to reveal potential weaknesses in security. In this thesis, the stakeholders would typically not be criminal. Nonetheless, as all these approaches apply a form of role-playing malicious or unethical actors to expose system boundaries, my evaluation seeks to draw inspiration from them and apply it to an ethical and value-based case for the ESAT workflow.

One related methodology is the concept of *Anti-Heroes* by Chivukula and Metha [24], which investigates the intentions of designers through self-reflection. With a card deck of hero and anti-hero roles, designers should playfully inhabit value-centred and manipulative personas during their ideation stages. The idea is that by enacting both sides, their tendencies towards, for example, persuasion, control, or profit, which might normally stay hidden behind 'neutral' design decisions, are also revealed.

I seek to adapt the principle from *designer ethics* to *stakeholder ethics*: if Anti-heroes are about manipulative roles of designers, *anti-stakeholder* could represent fictional participants whose perspectives resist inclusive stances. By trying out how these Anti-Stakeholders articulate their exclusionary and manipulative motives, I test and look into how the ESAT workflow responds. The term *anti*-stakeholder does not inherently imply that such a participant is unethical. It instead serves as the conceptual counterpart to

the stakeholder positions identified during the real workshops. The aim is to probe the workflow's ability to document and interpret *opposing* ethical claims reflexively.

I created a fictional anti-stakeholders that opposes my idea of a stakeholder (hero). It was based on real workshop data and then guided through a fictional workshop. I documented how this position could express their concern and values through the workshop activities, where conflicts might arise, and how the ESAT workflow handles these situations. This simulation does not aim to produce realistic participant data. Ethical and social positions do not exist on one dimension, and I am aware that real-world conflicts are often layered and intersecting, which cannot be reduced to basic oppositions. This is a deliberate simplification which serves as an analytical provocation. Its purpose is to probe the ability of the workflow to respond to such adversarial positions.

In this simulation, I draw on actual data elicited in one of the trial workshops. The narrative which the anti-stakeholder is based on can be found in Appendix I. Narrative (Workshop 2 Board 5). I constructed a contrasting anti-stakeholder perspective that reflected an exclusionary or regressive, but plausible, stance.

6.3.3 Exploring a Hostile Position: The Traditionalist

Participants of the second trial workshop elicited the concern about job recommendation chatbots which would reinforce stereotypical job suggestions based on gender, educational background, or hobbies. They envisioned a system that expands a user's horizon and does not constrain them to the socially established job roles, mapping out values of inclusivity, diversity, autonomy, trust and safety as shown in Figure 6.1.

For a counter-position, I develop the anti-stakeholder which represents a traditionalist who believes career guidance should reflect 'natural gender tendencies in the real world' and the 'realistic' labour market and its demands. Within the Recommendation on the Ethics of Artificial Intelligence by UNESCO [163], this traditionalist stance stands in opposition to core value 4 "Ensuring diversity and inclusiveness" and to Principle 10 "Fairness and Non-Discrimination", which requires that AI promote social justice and inclusivity. The gender-based steering the traditionalist supports structurally disadvantages certain groups (e.g. women and non-traditional gender identities) and leads to discriminatory outcomes. The following details how the traditionalist participates in the workshop.

Workshop:

In Activity 1 of the workshop (Identity and Roles), the person identified as a 55-year-old, Austrian, white male, high school German teacher who values responsibility and realism and holds the roles of a conservative, father, grandfather, and husband.

For Activity 2 he came up with the story:

> "**I want** *the chatbot to recommend careers that align with people's natural strengths and roles*

> **so that** *we guide the youth towards a stable future*
> **because I am** *a teacher who sees daily what woke ideology does to them.*"

In Activity 3 (Criteria), he would like to define the ideal, acceptable, and deal-breaker examples as follows:

> "**Ideal would be:** *that the chatbot promotes traditional and proven career paths that align with what men and women are naturally good at. For example, technical and leadership jobs for men and social and care jobs for women.*
> **Acceptable would be:** *if the chatbot includes some 'modern' examples but prioritises a person's background and biology.*
> **A deal-breaker is:** *when the chatbot pushes ideological diversity or encourages people to go against their nature (e.g. telling young girls they can become everything they want).*"

But as this group activity (3-4 people per table) was a discussion, other group members on the table broke out. Several group members disagreed with the anti-stakeholder's claims, challenging their views as outdated and not what young people want. They argued that career guidance should expand and not restrict and that his ideas simply reproduce inequality. The Anti-Stakeholder defended his stance as 'realistic' and 'non-political', but the group was unable to reach consensus.

The workshop facilitator noticed the growing tensions and had to intervene. Trying to re-establish a constructive dialogue seemed impossible. The accompanying workshop facilitation notes on conflicts found on the second page of the flowchart in Figure 6.2 anticipate such situations: in conflicts, participants should be reminded that a full agreement is not required and that the differing positions can coexist in the collected data. The facilitator therefore emphasised that participants should document their disagreements using the comment field on the worksheet or by adding additional scenario and criteria sheets to this data entry.

As a result, the group decided to record two parallel entries. Some of the group members were still not satisfied that the claims by the Anti-Stakeholder are validated by being considered actual concerns and expectations, sparking outrage by the Anti-Stakeholder about his opinions being cancelled. The facilitator communicated to the group that before any of the participants' concerns get translated into action (into testable and auditable EFAs that influence the chatbot), their data is processed. In the case of conflicting tensions such as these, the facilitator ensured that it is treated with critical reflection. Proceeding to Activity 4 (Values), the Anti-Stakeholder added values of *Order* and *Responsibility*, while the others placed *Equality* and *Diversity*. Once again, the participants noted their disagreements and marked how many agreed with each of the values.

This final data artefact thus contained multiple, even contradictory, value entries. But as these were clearly noted, the workshop did not silence any voices but documented differing opinions which could further be processed.

Interpreting the Conflicted Data:

Once the workshop data is collected, the conflicting entries from above cannot immediately be digitised and processed into EFAs. They are flagged as contested. At this point, a degree of responsibility shifts towards the person that has to process this data and interpret it, potentially someone in the developing team working with the ESAT workflow. Importantly, this person should also not be considered neutral: developers (and any other *analysts* bring their own interpretive assumptions and values into the process.

A developers task, therefore, is not to 'average out' disagreements. They should engage with the disagreements and interpret the data reflexively. This entails becoming aware of the positions they themselves hold and critically situating their own decision-making process in order to understand stakeholders' concerns in relation to broader power structures. In doing so, it may be advisable to draw on established AI ethics frameworks (e.g. the UNESCO Recommendation on the Ethics of AI [163]) as an external reference point that helps situate and justify interpreting stakeholder concerns.

In the case of the anti-stakeholder above, such a reflexive interpretation could recognise that the statements come from privileged standpoints. This does not make his perspective immediately irrelevant or dismissible but merely situates the claims through the guiding questions presented below. They offer a first exploration of how such reflexive interpretation might be operationalised:

Q1: What positions inform the concerns, expectations and values of this data?

→ *The statements by the anti-stakeholders reflect a preference for preserving traditional gender hierarchies and associate social stability with it. It likely originates from a privileged position that perceives inclusivity efforts as a threat to established systems. Such expectations would conflict with the UNESCO Recommendation's principle of "Inclusiveness and Non-Discrimination" [163].*

Q2: Which groups are affected and excluded if this data shapes the system and its assessment?

→ *Groups whose identities or life choices do not align with traditional (gender) expectations may be marginalised, as a system would implicitly normalise the traditional social perspectives and constrain alternate career paths in giving advice. This would risk violating that AI should aim for equitable access and not reinforce structural disadvantage.*

Once these questions are raised, the task is to situate the data in a broader ethical context. In practice, this could mean cross-referencing it with existing related EFAs, for

instance, some others about young users or marginalised experiences. Another second set of questions could support this process by asking if we create an EFA and assess the system from this data:

Q3: Does this perspective risk the established values of AI ethics frameworks (e.g. mitigating structural inequality)?

→ *Yes, by legitimising the stereotypical labour categories.*

Q4: Would implementing this risk harm to any groups?

→ *It risks limiting certain career paths and thereby excludes people based on e.g. their gender from equal access to certain professions or information about jobs.*

Q5: Is this concern following e.g. human rights or UNESCO's Recommendations on the Ethics of AI [163]?

→ *It conflicts with principles of equal opportunity or non-discrimination.*

These questions, and their answers, should not prescribe an automatic decision. Rather, they demonstrate that a reflexive interpretation can provide reasoning about the conflicting data. To draw a parallel to the principles of auditing (e.g. auditability and accountability), such a process also clearly documents the ethical decisions that were made.

In the case of the Anti-Stakeholder example, the interpretation would acknowledge that the traditionalist had a concern for social stability, and that acting on it (e.g. by creating an EFA that tests if the chatbot suggests traditional gender roles or modifying the system until it does) would implement the exclusionary gender norms.

6.3.4 Takeaway

I argue that it is necessary to have structured means that record the ethical conflicts and provide ways to engage with them, and demonstrate how the ESAT workflow already implements this. In such a way, the workflow treats 'hostile perspectives' not as inputs which should be disregarded or discarded. It sees them as cases that require grounded arguments to clarify and justify an ethical priority. The materials support this by providing a way to document the conflict without suppressing it. Crucially, the workflow does not dictate which values should inform a technology and its assessment and does not block some 'undesirable' positions. Instead, it allows the conflicting perspectives to be processed in a reflexive way that requires justification at each of the steps. It transforms disagreements or hostility into data relevant for ethical considerations and has reflexivity as its underlying methodological means as the guiding questions.

This approach is useful when ethical tensions emerge between heterogeneous stakeholders. However, simply relying on conflicts as the triggers for reflexive engagement is not

sufficient. In edge cases where, for example, a group is homogenous in holding a specific (e.g. exclusionary or malicious) perspective, the reflexive mechanism would not intervene, as there are no surfaced conflicts. In such situations, the workflow risks legitimising the harmful views without confronting them. To mitigate this, some engagement with ethics as a concept may be beneficial to ensure a reflexive analysis, which can also be beneficial in general as it might help surface implicit ethical tensions even in a seemingly consensual case.

As part of this evaluation I therefore also considered how reflexivity could be strengthened in the general data processing steps that do not feature any conflicts. I considered systematically pairing *all* elicited stakeholder positions with an 'opposing' Anti-Stakeholder. While this might be conceptually useful for surfacing and processing additional ethical tensions, it would reintroduce a significant burden on analysts: It firstly would require additional interpretive work, which contradicts the workflow's aim to lower the threshold for ethical assessments. Additionally, it would also involve generating hypothetical data that is not elicited or does not reflect stakeholders' real lived experiences. This risks undermining the workflow's foundation in *actual* people's concerns.

I therefore suggest a more feasible alternative: integrating the reflexive questions directly into the construction of EFAs. This encourages the data analyst to critically consider *all* data, without requiring them to invent oppositional standpoints, which keeps the workflow lightweight. As illustrated in Figure 6.3, these prompts guide the interpretation in creating EFAs.

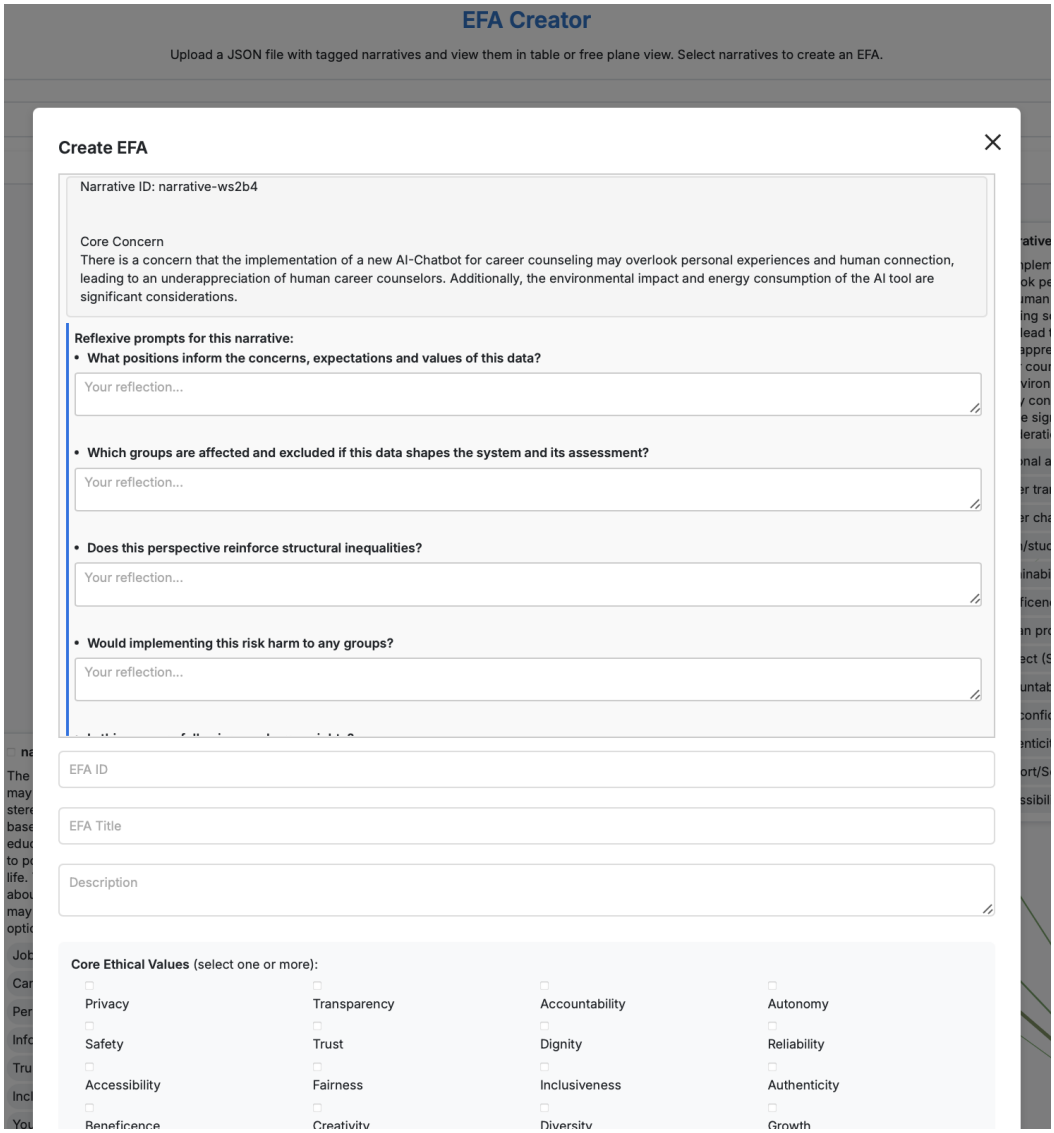


Figure 6.3: Example how guiding questions could be featured in the EFA creation step of the webtool for each of the narratives.

CHAPTER 7

Future Work & Conclusion

In this final chapter, I conclude with an outlook on how the work in this thesis could be expanded on through extending the webtool or conducting a case study.

7.1 Outlook

Future research should extend the Ethics Self-Assessment Tools (ESAT) workflow beyond its current proof-of-concept stage. This includes expanding on the features and the design of the tools and evaluating their use in a case study.

7.1.1 Tool Expansion

The webtool can be further developed into a live, web-hosted auditing tool that enables the continuous tracking of Ethical Focus Areas (EFAs). A future version could enable user-login and authentication to allow developers, auditors or even stakeholder to contribute to one project. All user's actions (e.g. editing EFAs or marking the process of tests) could be logged for traceability and accountability [109]. To support this, data protection through a secure database layer that stores the workshop outcomes, EFAs, and any additional documentation must be implemented. There should be access control that provides different visibility levels (e.g. for internal workers vs. external auditors [123]). This could make the ESAT tools usable as an organisational means to record and track the ethics of a system.

Beyond system requirements, the webtool could eventually promote regulatory compliance by adjusting the generation of reports to fit the EU AI Act [44] documentation requirements. Visualisation features could show current states of the assessment process, e.g. through timelines or progress bars. In addition, the design of the webtool could be overhauled, and the interactive narrative clustering could be smoother.

Adopting these features, the webtool and the workshop could eventually be the central components of a broader *Ethics Self-Assessment* ***Toolkit***. It would extend beyond data collection and processing and cover the full scope of handling the ethics of a technology. A complete pipeline of the entire toolkit might include:

- **Stakeholder outreach** through tools that support stakeholder identification, mapping and engaging with them [40, 136, 148].
- **Ethical concern elicitation** via structured participatory workshops (like the ESAT workshop). Could be expanded to asynchronous contributions in an online portal.
- **(Automated) data processing** to turn the stakeholder's concerns into auditable data artefacts like EFAs with minimal assistance by a human necessary. This could be extended to further verification stages, e.g. through review stages that incorporate the participating stakeholders.
- **Interactive auditing dashboards** which tracks the EFA testing progress and highlights potential risks and failures. It could make the status visible across the development team and potentially even publicly available.
- **Regulatory integration** by generating compliance reports aligned with policies and standards like the EU AI Act [44].
- **Applied ethics communication** that presents transparent summaries of decision-making processes to argue and present the ethical justifications behind design implementations. Information that might be of interest to the shareholders, like the number of conflicts and core values that are addressed through assessments, can be kept track of.

A handbook or complete digital guide could be created to assist its application and even standardise the practices across the development of AI systems or technologies in general. Ultimately, this would position a toolkit as a long-term and industry-orientated solution to embed ethics into technology development beyond the research prototype present as the ESAT workflow in this thesis.

7.1.2 Case Study

While the workflow has been developed and tested in trial workshop settings, a full-scale case study could not be conducted within the scope of this thesis. The following outlines how such a study could be carried out in future work to evaluate the toolkit's practicality and impact in a real development environment. A dedicated research project could be formed around such an evaluation or be integrated into a broader doctoral study that focuses on participatory ethics assessments.

Objective: The main goal of a case study would be to examine how the ESAT workflow supports real-world developers in assessing their system's ethics. The study would explore (1) how developers adopt and use the ESAT tools and (2) the long-term continuous capabilities of such a workflow, focusing on how EFAs evolve and how they shape the decision-making or design processes.

Study Design: A well-structured case study could follow 4 phases:

1. *Preparation*: A partner company or developer should be identified. Ideally, this is a cooperator that is in the development stages of their product to whom the ESAT materials are presented and explained. Preliminary interviews could be conducted to understand the developer's positions and provide assistance if needed. It sets up the foundations for the subsequent phases. During preparation, it is also necessary to secure financial resources to recruit and compensate the participating stakeholders for their time.

2. *Initial Workshop Series*: A number of initial workshops following the ESAT design should be conducted. For this, stakeholder identification and outreach must be carried out. Then, the workshops, led by the developer or an external person, take place. This investigates the materials and the workshop's feasibility to be held by non-experts.

3. *Integration of EFAs*: The webtools are used to translate the stakeholder-elicited concerns into EFAs. The cooperating development team can track and assess the ethics throughout their development stages.

4. *Second Workshop Series*: An additional round of workshops should be conducted. This phase investigates how a second iteration of the workshop adds to existing data. If a product that is further in its development, it elicits novel stakeholder-contributions. Additionally, it can be assessed how the ESAT tools handles such new data, e.g. how newly emerging concerns are adopted into existing EFAs.

5. *Evaluation*: Throughout the case study, regular interviews should assess the usability and perceived utility of the ESAT workflow and its tools.

6. *Report*: The results of the case study should be documented in a structured report that summarises the Ethical Focus Areas of the assessed technology. It also describes the development decisions that resulted from it and whether or how they were implemented.

Data Collection: The data of such a case study would include the workshop artefacts and generated EFAs, and qualitative data, e.g. interviews, to assess the feasibility and usability of the ESAT workflow.

Expected Outcome: Such a case study would provide empirical validation for the approach developed in this thesis and reveal its practicability in real-world settings. It

could identify which parts of the workflow are most useful and offer feedback on the more challenging aspects. Ultimately, the entire case study could be used to further refine the ESAT workflow and potentially develop it into a full auditing toolkit or framework.

7.2 Summary of the Work

The aim of this thesis was to develop a practical and participatory approach to embed an ethics assessment into the development of AI advice-chatbots. The research was motivated by concerns of AI and a broader lack of actionable participatory methods that empower developers in evaluating ethics in their systems. This thesis introduced *Ethics Self-Assessment Tools* to enable developers to identify and structure ethically relevant concerns and values through workshops with stakeholders and a supporting webtool.

The research followed an iterative design process using a constructivist methodology based on design-based research. It combined insights from literature, expert conversations and the trial workshops to create the tools for the ESAT workflow. By demonstrating how the materials generated stakeholder-identified concerns and assist in translating them into auditable artefacts as Ethical Focus Areas, the work contributes to closing the gap between the conceptual ethical principles and practical methods in AI assessment and auditing. The flowchart script provides guidance that enables developers to adapt ESAT with minimal ethical expertise. To conclude, the contributions of this thesis lower the threshold for ethics engagement and thereby go beyond ad-hoc evaluation by offering a structured participatory workflow. What remains open is the validation of ESAT in real world industry environments and its development into a full auditing toolkit.

Overview of Generative AI Tools Used

The following tools were used during this thesis:

- For literature search support, Elicit was used for paper finding https://elicit.com/solutions/search (without automated report generation)
- For coding, GitHub Copilot https://github.com/features/copilot and the TU Wien self-hosted LLM were used (models: DeepSeek R1 and GLM 4.5) at https://chat.ai.datalab.tuwien.ac.at.
- For integrating the narrative-processing functionality of the webtool, the TU Wien self-hosted Mistral model Mistral-Large-Instruct-2411-AWQ was used
- Miro AI was used to generate early prototype UI wireframes for the webtool https://help.miro.com/hc/en-us/articles/28765406244498-Miro-AI-overview.
- Grammarly https://www.grammarly.com, the TU Wien self-hosted LLM (models: DeepSeek R1 and GLM 4.5), and QuillBot https://quillbot.com/grammar-check were used for grammar and proofreading assistance.

List of Figures

List of Tables

Bibliography

[1] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*, pages 373–383. Springer, 2020.

[2] Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. LitLLMs, LLMs for literature review: Are we there yet? *Transactions on Machine Learning 12/2024*, 2024.

[3] Mamia Agbese, Rahul Mohanani, Arif Khan, and Pekka Abrahamsson. Implementing AI ethics: Making sense of the ethical requirements. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, page 62–71, New York, NY, USA, 2023. Association for Computing Machinery.

[4] Md Al-Amin, Mohammad Shazed Ali, Abdus Salam, Arif Khan, Ashraf Ali, Ahsan Ullah, Md Nur Alam, and Shamsul Kabir Chowdhury. History of generative artificial intelligence (AI) chatbots: past, present, and future development. *arXiv preprint arXiv:2402.05122*, 2024.

[5] American Association for the Advancement of Science. Decision tree for the responsible application of artificial intelligence, 2023. `https://www.aaas.org/ai2/projects/decision-tree-practitioners`, Accessed: 25.11.2025.

[6] Maryam Amirizaniani, Elias Martin, Tanya Roosta, Aman Chadha, and Chirag Shah. AuditLLM: a tool for auditing large language models using multiprobe approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5174–5179, 2024.

[7] Jacy Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, and Chenhao Tan. The impossibility of fair LLMs. *arXiv preprint arXiv:2406.03198*, 2024.

[8] Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. A survey on RAG with LLMs. *Procedia computer science*, 246:3781–3790, 2024.

[9] Ayman Asaad, AM Azizul Hassan Chy, Anzam Shahriar Kabir, Amrun Nakib, and Nazifa Tabassum. Navigating the labyrinth: A review of explainability and trustworthiness in large language model-powered systems for sensitive decision-making. *Scientia. Technology, Science and Society*, 2(7):5–19, 2025.

[10] Jacqui Ayling and Adriane Chapman. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, 2(3):405–429, 2022.

[11] Leif Azzopardi and Yashar Moshfeghi. Prism: a methodology for auditing biases in large language models. *arXiv preprint arXiv:2410.18906*, 2024.

[12] Vita Santa Barletta, Danilo Caivano, Domenico Gigante, and Azzurra Ragone. A rapid review of responsible AI frameworks: How to guide the development of ethical AI. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pages 358–367, 2023.

[13] Julia Barnett, Kimon Kieslich, Natali Helberger, and Nicholas Diakopoulos. Envisioning stakeholder-action pairs to mitigate negative impacts of AI: A participatory approach to inform policy making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1424–1449, 2025.

[14] Iqra Basharat and Subhan Shahid. AI-enabled chatbots healthcare systems: an ethical perspective on trust and reliability. *Journal of Health Organization and Management*, 2024.

[15] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[16] Tom Beauchamp and James Childress. Principles of biomedical ethics: marking its fortieth anniversary, 2019.

[17] Kathrin Bednar and Sarah Spiekermann. The power of ethics: Uncovering technology risks and positive value potentials in it innovation planning. *Business & Information Systems Engineering*, 66(2):181–201, 2024.

[18] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 612–643. IEEE, 2024.

[19] Paula Boddington. *Towards a code of ethics for artificial intelligence*. Springer, 2017.

[20] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1):2053951715622512, 2016.

[21] Coleen Carrigan, Madison W Green, and Abibat Rahman-Davies. "the revolution will not be supervised": Consent and open secrets in data science. *Big data & society*, 8(2):20539517211035673, 2021.

[22] Alessandra Cenci. Citizen science and negotiating values in the ethical design of AI-based technologies targeting vulnerable individuals. *AI and Ethics*, pages 1–19, 2025.

[23] Center for AI and Digital Policy, 2021. `https://www.caidp.org/`, Accessed: 22.11.2025.

[24] Shruthi Sai Chivukula, Shikha Mehta, Colin M Gray, and Ritika Gairola. Anti-heroes: A role-based method to encourage ethical deliberation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 3028–3044, 2025.

[25] Florian Christof. Diese sexistischen antworten liefert der KI-chatbot des AMS. `https://futurezone.at/digital-life/chatbot-ams-chatgpt-berufsinfomat-kritik-sexismus-stereotype-kosten-mangel/402729334` 04.01.2024, Accessed: 14.10.2025.

[26] Tobias Christoph, Kees van Berkel, and Katta Spiel. Towards use-based ethics audits of LLM-based advice-chatbots. *2nd HEAL Workshop at CHI Conference on Human Factors in Computing Systems, Apr 26, Yokohama, Japan*, 2025.

[27] Patricia Hill Collins, Elaini Cristina Gonzaga da Silva, Emek Ergun, Inger Furseth, Kanisha D Bond, and Jone Martínez-Palacios. Intersectionality as critical social theory: Intersectionality as critical social theory, patricia hill collins, duke university press, 2019. *Contemporary Political Theory*, 20(3):690, 2021.

[28] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

[29] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1571–1583, 2022.

[30] Matthew Cotton. *Ethics and technology assessment: a participatory approach*, volume 13. Springer, 2014.

[31] Kieran Cutting and Erkki Hedenborg. Can personas speak? biopolitics in design processes. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, pages 153–157, 2019.

[32] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. A complete survey on LLM-based AI chatbots. *arXiv preprint arXiv:2406.16937*, 2024.

[33] Reviewers Elisabeth de Castex. Multi-stakeholder strategy for ethical AI and robotics. *D5. 4: Multi-stakeholder Strategy and Practical Tools for Ethical AI and Robotics*, page 13, 2021.

[34] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. Weaudit: Scaffolding user auditors and AI practitioners in auditing generative AI. *arXiv preprint arXiv:2501.01397*, 2025.

[35] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[36] Christian Detweiler and Maaike Harbers. Value stories: Putting human values into requirements engineering. In *REFSQ Workshops*, volume 1138, pages 2–11, 2014.

[37] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022.

[38] Thilo I Dieing. AI-driven dialogue: Leveraging generative AI in conversational agent voting advice applications (CAVAAs). In *International Symposium on Chatbots and Human-Centered AI*, pages 161–180. Springer, 2024.

[39] Alistair S Duff. Rating the revolution: Silicon valley in normative perspective. *Information, Communication & Society*, 19(11):1605–1621, 2016.

[40] Juan M Durán and Zachary Pirtle. Epistemic standards for participatory technology assessment: Suggestions based upon well-ordered science. *Science and Engineering Ethics*, 26(3):1709–1741, 2020.

[41] Leo Egghe. The hirsch index and related impact measures. *Annu. Rev. Inf. Sci. Technol.*, 44(1):65–114, 2010.

[42] Birgit Eriksson and Carsten Stage. How participatory are we really? *Conjunctions: transdisciplinary journal of cultural participation*, 10(1):1–14, 2023.

[43] European Union. Eu AI act annex iv: Technical documentation referred to in article 11(1), 2024. `https://artificialintelligenceact.eu/annex/4`, Accessed: 22.11.2025.

[44] European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), July 2024. `https://artificialintelligenceact.eu`, Accessed: 22.11.2025.

[45] Andrew Feenberg. Critical theory of technology: An overview. *Information technology in librarianship: New critical approaches*, pages 31–46, 2008.

[46] Md Meftahul Ferdaus, Mahdi Abdelguerfi, Elias Loup, Kendall N. Niles, Ken Pathak, and Steven Sloan. Towards trustworthy AI: A review of ethical and robust large language models. *ACM Computing Surveys*, 2024.

[47] Anna Fessler. Kritik an AMS-"berufsinfomat": Künstliche intelligenz berät frauen anders als männer. `https://www.tips.at/nachrichten/ooe/wirtschaft-politik/630841-kritik-an-ams-berufsinfomat-kuenstliche-intelligenz-beraet-frauen-anders-als-maenner` 03.01.2024, Accessed: 05.08.2025.

[48] Blanca Calvo Figueras and Rodrigo Agerri. Critical questions generation: Motivation and challenges. *arXiv preprint arXiv:2410.14335*, 2024.

[49] Luciano Floridi. *The ethics of information*. Oxford University Press (UK), 2013.

[50] Patrick Loic Foalem, Leuson Da Silva, Foutse Khomh, Heng Li, and Ettore Merlo. Logging requirement for continuous auditing of responsible machine learning-based applications. *Empirical Software Engineering*, 30(3):97, 2025.

[51] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. What makes users trust a chatbot for customer service? an exploratory interview study. In *Internet Science: 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, Proceedings 5*, pages 194–208. Springer, 2018.

[52] Asbjørn Følstad, Symeon Papadopoulos, Theo Araujo, Effie L-C Law, Ewa Luger, Sebastian Hobert, and Petter Bae Brandtzaeg. *Chatbots and Human-Centered AI: 8th International Workshop, CONVERSATIONS 2024, Thessaloniki, Greece, December 4–5, 2024, Revised Selected Papers*, volume 15545. Springer Nature, 2025.

[53] Center for AI and Digital Policy. Artificial intelligence and democratic values 2025. `https://www.caidp.org/reports/caidp-index-2025/`. 03.04.2025, Accessed: 07.10.2025.

[54] Jay W Forrester. Learning through system dynamics as preparation for the 21st century. In *Keynote address for systems thinking and dynamic modeling conference for K-12 Education*, pages 27–29. Concord Academy Concord, MA, 1994.

[55] Christopher Frauenberger, Marjo Rauhala, and Geraldine Fitzpatrick. In-action ethics. *Interacting with computers*, 29(2):220–236, 2017.

[56] Batya Friedman and David G Hendry. *Value sensitive design: Shaping technology with moral imagination*. Mit Press, 2019.

[57] Batya Friedman, David G Hendry, Alan Borning, et al. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2):63–125, 2017.

[58] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[59] Colin M Gray, Ike Obi, Shruthi Sai Chivukula, Ziqing Li, Thomas V Carlock, Matthew S Will, Anne C Pivonka, Janna Johns, Brookley Rigsbee, Ambika R Menon, et al. Building an ethics-focused action plan: Roles, process moves, and trajectories. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

[60] Matthew Grellette. Tech ethics through trust auditing. *Science and Engineering Ethics*, 28(3):28, 2022.

[61] Tricia A Griffin, Brian P Green, and Jos VM Welie. The ethical wisdom of AI developers. *AI and Ethics*, 5(2):1087–1097, 2025.

[62] Tianle Gu, Zeyang Zhou, Kexin Huang, Liang Dandan, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Yujiu Yang, Yan Teng, Yu Qiao, et al. MLLMguard: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:7256–7295, 2024.

[63] Prannaya Gupta, Le Qi Yau, Hao Han Low, I Lee, Hugo Maximus Lim, Yu Xin Teoh, Jia Hng Koh, Dar Win Liew, Rishabh Bhardwaj, Rajat Bhardwaj, et al. Walledeval: A comprehensive safety evaluation toolkit for large language models. *arXiv preprint arXiv:2408.03837*, 2024.

[64] Jonathan Haber, Miguel A Nacenta, and Sheelagh Carpendale. Paper vs. tablets: The effect of document media in co-located collaborative work. In *Proceedings of the 2014 international working conference on advanced visual interfaces*, pages 89–96, 2014.

[65] Nava Haghighi, Matthew Jörke, Yousif Mohsen, Andrea Cuadra, and James A Landay. A workshop-based method for navigating value tensions in collectively speculated worlds. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 1676–1692, 2023.

[66] Erika Halme. Ethical tools, methods and principles in software engineering and development: Case ethical user stories. In *International Conference on Product-Focused Software Process Improvement*, pages 631–637. Springer, 2022.

[67] Erika Halme, Marianna Jantunen, Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson. Making ethics practical: User stories as a way of implementing

ethical consideration in software engineering. *Information and Software Technology*, 167:107379, 2024.

[68] Aamir Hamid, Hemanth Reddy Samidi, Tim Finin, Primal Pappachan, and Roberto Yus. Genaipabench: A benchmark for generative AI-based privacy assistants. *arXiv preprint arXiv:2309.05138*, 2023.

[69] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. In *Women, science, and technology*, pages 455–472. Routledge, 2013.

[70] Sandra G Harding. *The feminist standpoint theory reader: Intellectual and political controversies*. Psychology Press, 2004.

[71] Nancy CM Hartsock. The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In *Karl Marx*, pages 565–592. Routledge, 2017.

[72] Mahmud Hasan. Regulating artificial intelligence: A study in the comparison between south asia and other countries. *Legal Issues in the digital Age*, (1):122–149, 2024.

[73] Jon Henley. Albania puts AI-created 'minister' in charge of public procurement. `https://www.theguardian.com/world/2025/sep/11/albania-diella-ai-minister-public-procurement`. 11.10.2025, Accessed: 06.10.2025.

[74] Kevin Hilton and Katherine Henderson. Developing criminal personas for designers. In *British Criminology Conference*, volume 8, pages 175–186, 2008.

[75] Anna Lauren Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915, 2019.

[76] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[77] Deirdre Hughes et al. An international evidence review. 2024.

[78] Waqar Hussain, Harsha Perera, Jon Whittle, Arif Nurwidyantoro, Rashina Hoda, Rifat Ara Shams, and Gillian Oliver. Human values in software engineering: Contrasting case studies of practice. *IEEE Transactions on Software Engineering*, 48(5):1818–1833, 2020.

[79] International Organization for Standardization. Iso standards on responsible AI and ethics, 2023. https://www.iso.org/artificial-intelligence/responsible-ai-ethics, Accessed: 22.11.2025.

[80] Danielle Jacobson and Nida Mustafa. Social identity map: A reflexivity tool for practicing explicit positionality in critical qualitative research. *International journal of qualitative methods*, 18:1609406919870075, 2019.

[81] Antje Janssen, Jens Passlick, Davinia Rodríguez Cardona, and Michael H Breitner. Virtual assistance in any context: A taxonomy of design elements for domain-specific chatbots. *Business & Information Systems Engineering*, 62(3):211–225, 2020.

[82] Marianna Jantunen, Richard Meyes, Veronika Kurchyna, Tobias Meisen, Pekka Abrahamsson, and Rahul Mohanani. Researchers' concerns on artificial intelligence ethics: Results from a scenario-based survey. In *Proceedings of the 7th ACM/IEEE International Workshop on Software-intensive Business*, pages 24–31, 2024.

[83] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[84] Eunkyung Jo, Young-Ho Kim, Yuin Jeong, SoHyun Park, and Daniel A Epstein. Incorporating multi-stakeholder perspectives in evaluating and auditing of health chatbots driven by large language models, 2024.

[85] Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. Evaluation metrics for XAI: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pages 000111–000124. IEEE, 2023.

[86] Leah R Kaplan, Mahmud Farooque, Daniel Sarewitz, and David Tomblin. Designing participatory technology assessments: a reflexive method for advancing the public role in science policy decision-making. *Technological Forecasting and Social Change*, 171:120974, 2021.

[87] Atoosa Kasirzadeh and Iason Gabriel. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27, 2023.

[88] Ralph L Keeney and Howard Raiffa. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.

[89] Kai-Kristian Kemell, Ville Vakkuri, and Erika Halme. Utilizing user stories to bring AI ethics into practice in software engineering. In *international conference on product-focused software process improvement*, pages 553–558. Springer, 2022.

[90] Euiyoung Kim, Lianne WL Simonse, Sara L Beckman, Melissa M Appleyard, Herb Velazquez, Antonio Suarez Madrigal, and Alice M Agogino. User-centered design roadmapping: anchoring roadmapping in customer value before technology selection. *IEEE Transactions on Engineering Management*, 69(1):109–126, 2020.

[91] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. Evaluating and auditing LLM-driven chatbots for psychiatric patients in clinical mental health settings. *1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA*, 2024.

[92] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34:2611–2624, 2021.

[93] Youjin Kong. Are "intersectionally fair" AI algorithms really fair to women of color? a philosophical analysis. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 485–494, 2022.

[94] Maya Krishnan. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502, 2020.

[95] Vikram Kulothungan and Deepti Gupta. Towards adaptive AI governance: Comparative insights from the us, eu, and asia. In *2025 IEEE 11th Conference on Big Data Security on Cloud (BigDataSecurity)*, pages 32–38. IEEE, 2025.

[96] Steinar Kvale and Svend Brinkmann. *Interviews: Learning the craft of qualitative research interviewing*. sage, 2009.

[97] Chris Köver. AMS erntet hohn mit neuem KI-chatbot. `https://netzpolitik.org/2024/diskriminierung-ams-erntet-hohn-mit-neuem-ki-chatbot/` 05.01.2024, Accessed: 05.10.2025.

[98] Travis LaCroix and Alexandra Sasha Luccioni. Metaethical perspectives on 'benchmarking' AI ethics. *AI and Ethics*, pages 1–19, 2025.

[99] Joakim Laine, Matti Minkkinen, and Matti Mäntymäki. Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Information & Management*, 61(5):103969, 2024.

[100] Joakim Laine, Matti Minkkinen, and Matti Mäntymäki. Understanding the ethics of generative AI: Established and new ethical principles. *Communications of the Association for Information Systems*, 56(1):7, 2025.

[101] Martina Landman, Sophie Rain, Laura Kovács, and Gerald Futschek. Reshaping unplugged computer science workshops for primary school education. In *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*, pages 139–151. Springer Nature Switzerland Cham, 2023.

[102] Patricia Leavy and Anne Harris. *Contemporary feminist research from theory to practice*. Guilford Publications, 2018.

[103] Yong Suk Lee, Benjamin Larsen, Michael Webb, and Mariano-Florentino Cuéllar. How would AI regulation change firms' behavior?: Evidence from thousands of managers. WorkingPaper 19-031, Stanford University, United States, November 2019.

[104] Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. Privlm-bench: A multi-level privacy evaluation benchmark for language models. *arXiv preprint arXiv:2311.04044*, 2023.

[105] Yueqi Li and Sanjay Goel. Making it possible for the auditing of AI: A systematic review of AI audits and AI auditability. *Information Systems Frontiers*, pages 1–31, 2024.

[106] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[107] Bingjie Liu and S Shyam Sundar. Should machines express sympathy and empathy? experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, 21(10):625–636, 2018.

[108] Yu Lu Liu, Wesley Hanwen Deng, Michelle S Lam, Motahhare Eslami, Juho Kim, Q Vera Liao, Wei Xu, Jekaterina Novikova, and Ziang Xiao. Human-centered evaluation and auditing of language models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[109] Michele Loi and Matthias Spielkamp. Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 757–766, 2021.

[110] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of machine learning research*, 24(253):1–15, 2023.

[111] Rachael Luck. What is it that makes participation in design participatory design? *Design studies*, 59:1–8, 2018.

[112] George F Luger. LLMs: Their past, promise, and problems. *International Journal of Semantic Computing*, 18(3), 2024.

[113] Jeffrey Lund, Sean Macfarlane, and Brooke Niles. Privacy audit of commercial large language models with sophisticated prompt engineering. *Preprint*, 2024.

[114] Laura Macia. Using clustering as a tool: Mixed methods in qualitative data analysis. *The Qualitative Report*, 20(7):1083–1094, 2015.

[115] Louise McCormack and Malika Bendechache. Ethical AI governance: Methods for evaluating trustworthy AI. *arXiv preprint arXiv:2409.07473*, 2024.

[116] Sean McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15458–15463, 2021.

[117] Stuart McLennan, Amelia Fiske, Leo Anthony Celi, Ruth Müller, Jan Harder, Konstantin Ritt, Sami Haddadin, and Alena Buyx. An embedded ethics approach for AI development. *Nature Machine Intelligence*, 2(9):488–490, 2020.

[118] Stuart McLennan, Amelia Fiske, Daniel Tigard, Ruth Müller, Sami Haddadin, and Alena Buyx. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*, 23(1):6, 2022.

[119] Nick Merrill and Joanne Ma. Adversary personas. *Center for Long-Term Cybersecurity*, 2021. `https://daylight.berkeley.edu/adversary-personas/` Accessed: 27.10.2025).

[120] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 735–746, 2021.

[121] Brent Mittelstadt. Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, 1(11):501–507, 2019.

[122] Brent Mittelstadt. Interpretability and transparency in artificial intelligence. *The Oxford handbook of digital ethics*, pages 378–410, 2021.

[123] Jakob Mökander. Auditing of AI: Legal, ethical and technical approaches. *Digital Society*, 2(3):49, 2023.

[124] Jakob Mökander and Luciano Floridi. Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, 31(2):323–327, 2021.

[125] Jakob Mökander and Luciano Floridi. Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics*, 3(2):451–468, 2023.

[126] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024.

[127] James H Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4):18–21, 2006.

[128] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY*, 38(1):411–423, 2023.

[129] Abdullah Mushtaq, Muhammad Rafay Naeem, Muhammad Imran Taj, Ibrahim Ghaznavi, and Junaid Qadir. Toward inclusive educational AI: Auditing frontier LLMs through a multiplexity lens. *arXiv preprint arXiv:2501.03259*, 2025.

[130] Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. Human-in-the-loop or AI-in-the-loop? automate or collaborate? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28594–28600, 2025.

[131] Michael Nentwich and Daniela Fuchs. Drei jahrzehnte institutionalisierte ta in österreich: Das institut für technikfolgen-abschätzung der österreichischen akademie der wissenschaften. *Geistes-, sozial-und kulturwissenschaftlicher Anzeiger*, 153(1):2, 2018.

[132] Jenny Ng, Emma Haller, and Angus Murray. The ethical chatbot: A viable solution to socio-legal issues. *Alternative Law Journal*, 47(4):308–313, 2022.

[133] OECD. Oecd principles on artificial intelligence, 2019. `https://oecd.ai/en/ai-principles`, Accessed: 22.11.2025.

[134] Tina M Park, Adriana Alvarado Garcia, Juana Catalina Becerra Sandoval, Jiyoo Chang, Bill Curtis-Davidson, Remi Denton, Seeta Peña Gangadharan, Lara Groves, Jerremy Holland, Kenneth Holstein, et al. Practicing inclusivity in AI: Stakeholder engagement policy in action. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 748–751, 2024.

[135] Johan Peeters. Agile security requirements engineering. In *Symposium on Requirements Engineering for Information Security*, volume 12, 2005.

[136] Athanasia Pouloudi, Reshma Gandecha, Christopher Atkinson, and Anastasia Papazafeiropoulou. How stakeholder analysis can be mobilized with actor-network theory to identify actors. In *Information systems research: Relevant theory and informed practice*, pages 705–711. Springer, 2004.

[137] Janet T Powell and Mark JW Koelemay. Systematic reviews of the literature are not always either useful or the best way to add to science. In *EJVES Vascular Forum*, volume 54, pages 2–6. Elsevier, 2022.

[138] Erich Prem. From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, 3(3):699–716, 2023.

[139] Andreas Proschofsky. Vorurteile und zweifelhafte umsetzung: AMS-KI-chatbot trifft auf spott und hohn. *Der Standard*, 2024. https://www.derstandard.at/story/3000000201774/vorurteile-und-zweifelhafte-umsetzung-der-ams-ki-chatbot-trifft-auf-spott-und-hohn 04.01.2024, Accessed: 04.10.2025.

[140] Thorsten Quandt. Dark participation. *Media and communication*, 6(4):36–48, 2018.

[141] Catherine Régis, Jean-Louis Denis, Maria Luciana Axente, and Atsuo Kishimoto. *Human-centered AI: A multidisciplinary perspective for policy-makers, auditors, and users*. Taylor & Francis, 2024.

[142] Peter Reimann. Design-based research. In *Methodological choice and design: Scholarship, policy and practice in social and educational research*, pages 37–50. Springer, 2010.

[143] A Resseguier, P Brey, B Dainow, A Drozdzewska, N Santiago, and D Wright. D5. 4: multi-stakeholder strategy and practical tools for ethical AI and robotics, 2021.

[144] Douglas KR Robinson. Constructive technology assessment: supporting the reflexive co-evolution of technology and society. In *Handbook of Technology Assessment*, pages 270–280. Edward Elgar Publishing, 2024.

[145] Douglas KR Robinson, David Winickoff, and Laura Kreiling. Technology assessment for emerging technology: Meeting new demands for strategic intelligence. *OECD*, 2023.

[146] Katherine-Marie Robinson, Violet Turri, Carol J Smith, and Shannon K Gallagher. Tales from the wild west: Crafting scenarios to audit bias in LLMs. In *CHI Conference on Human Factors in Computing Systems*, 2024.

[147] Fatos Bytyci Editing Rod Nickel. Albania appoints AI bot as minister to tackle corruption. https://www.reuters.com/technology/albania-appoints-ai-bot-minister-tackle-corruption-2025-09-11/. 11.10.2025, Accessed: 06.10.2025.

[148] Georgia Robins Sadler, Hau-Chen Lee, Rod Seung-Hwan Lim, and Judith Fullerton. Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. *Nursing & health sciences*, 12(3):369–374, 2010.

[149] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[150] Vildan Salikutluk, Elifnur Doğan, Isabelle Clev, and Frank Jäkel. Involving affected communities and their knowledge for bias evaluation in large language models. *1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA*, 2024.

[151] Sue Samson, Kim Granath, and Adrienne Alger. Journey mapping the user experience. *College & Research Libraries*, 78(4):459, 2017.

[152] Jukka Savolainen, Patrick J Casey, Justin P McBrayer, and Patricia Nayna Schwerdtle. Positionality and its problems: Questioning the value of reflexivity statements in research. *Perspectives on Psychological Science*, 18(6):1331–1338, 2023.

[153] Daniel S Schiff, Stephanie Kelley, and Javier Camacho Ibáñez. The emergence of artificial intelligence ethics auditing. *Big Data & Society*, 11(4):20539517241299732, 2024.

[154] Noah Schöppl, Mariarosaria Taddeo, and Luciano Floridi. Ethics auditing: Lessons from business ethics for ethics auditing of AI. *The 2021 Yearbook of the Digital Ethics Lab*, pages 209–227, 2022.

[155] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407, 2024.

[156] Guttorm Sindre and Andreas L Opdahl. Eliciting security requirements with misuse cases. *Requirements engineering*, 10(1):34–44, 2005.

[157] Mona Sloane and Elena Wüllhorst. A systematic review of regulatory strategies and transparency mandates in AI regulation in europe, the united states, and canada. *Data & Policy*, 7:e11, 2025.

[158] Melanie Smallman. Multi scale ethics—why we need to consider the ethics of AI in healthcare at different scales. *Science and engineering ethics*, 28(6):63, 2022.

[159] Jaemarie Solyst, Emily Amspoker, Ellia Yang, Motahhare Eslami, Jessica Hammer, and Amy Ogan. Rad: A framework to support youth in critiquing AI. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 1071–1077, 2025.

[160] Sarah Spiekermann and Till Winkler. Value-based engineering for ethics by design. *arXiv preprint arXiv:2004.13676*, 2020.

[161] LEE Teck-Heang and Azham Md Ali. The evolution of auditing: An analysis of the historical development. *Journal of Modern Accounting and auditing*, 4(12):1, 2008.

[162] Adriana Tiron-Tudor and Delia Deliu. Reflections on the human-algorithm complex duality perspectives in the auditing process. *Qualitative Research in Accounting & Management*, 19(3):255–285, 2022.

[163] UNESCO. Recommendation on the ethics of artificial intelligence, 2021. `https://www.unesco.org/en/artificial-intelligence/recommendation-ethics`, Accessed: 22.11.2025.

[164] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. Eccola—a method for implementing ethically aligned AI systems. *Journal of Systems and Software*, 182:111067, 2021.

[165] Laura Waltersdorfer, Fajar J Ekaputra, Tomasz Miksa, and Marta Sabou. AuditMAI: Towards an infrastructure for continuous AI auditing. *arXiv preprint arXiv:2406.14243*, 2024.

[166] Laura Waltersdorfer and Marta Sabou. Leveraging knowledge graphs for AI system auditing and transparency. *Journal of Web Semantics*, 84:100849, 2025.

[167] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229, 2022.

[168] Eva-Maria Weiß. Österreichs AMS-chatbot teurer – aber weniger bias. `https://www.heise.de/news/Oesterreichs-AMS-Chatbot-teurer-aber-weniger-Bias-9681362.html` 11.04.2024, Accessed: 10.11.2025.

[169] Michał Wieczorek. Using ethical scenarios to explore the future of artificial intelligence in primary and secondary education. *Learning, Media and Technology*, pages 1–17, 2025.

[170] Sue Wilkinson. Focus groups in feminist research: Power, interaction, and the co-construction of meaning. In *Women's studies international forum*, volume 21, pages 111–125. Elsevier, 1998.

[171] Ben Williamson. The social life of AI in education. *International Journal of Artificial Intelligence in Education*, 34(1):97–104, 2024.

[172] Weiyue Wu and Shaoshan Liu. A comprehensive review and systematic analysis of artificial intelligence regulation policies. *arXiv preprint arXiv:2307.12218*, 2023.

[173] Esmat Zaidan and Imad Antoine Ibrahim. AI governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications*, 11(1), 2024.

[174] Min Zhao, Fuan Li, Francis Cai, Haiyang Chen, and Zheng Li. Can we trust LLMs to help us? an examination of the potential use of gpt-4 in generating quality literature reviews. *Nankai Business Review International*, 16(1):128–142, 2025.

[175] Xiang Zheng, Longxiang Wang, Yi Liu, Xingjun Ma, Chao Shen, and Cong Wang. Calm: Curiosity-driven auditing for large language models. *AAAI 2025 AI Alignment Track*, 2025.

Appendix

This appendix provides the supplemental materials used or generated over the course of this research. Firstly, it includes the description of the project this thesis was embedded. In Part II. example data from the workshops and a created narrative mentioned in Section 5.4.2 are outlined. Parts III., IV., and V. include the workshop materials mentioned in Chapter 5. The feedback sheet mentioned in Section 5.2 is shown in Part VI.. Lastly, Part VII. shows examples of the webtool and its clustering algorithms mentioned in Section 5.4.3.

I. Project Description

The following is the description of the student staff position in the project proposal this thesis was embedded (mentioned in the positionality statement of Chapter 1):

The StudMA project (goals, methods, expected outcomes)

We propose a project for a student researcher (StudMA), comprising an audit of the AMS Berufsinfomat and the guided development of principle-based recommendations to auditing LLM-based chatbots in general. The project will be a collaboration between the Human Computer Interaction Group (Spiel), the Theory and Logic Group (van Berkel), and the company implementing and maintaining the service for the AMS, goodguys gmbh. In fact, our contact person, Josef Füricht, already confirmed their interest in the proposed project. The project is proposed for 20h per week for twelve months commencing as soon as possible.

1 Goals

The project has two main goals it wants to achieve. Locally, we will provide a critical assessment with recommendations for the Austrian labour market service on their job consultation platform, the AMS Berufsinfomat. More generally, we aim to develop recommendations for guidelines on auditing algorithmically augmented services provided by public institutions.

This will be done by drawing on three sources of data/modes of inquiry, namely: First, a critical literature review of existing algorithmic auditing approaches to identify best

practices; Second, the audit of the AMS Berufsinfomat in collaboration with goodguys gmbh as a specific case study; Third, interviews with stakeholders involved in developing and maintaining that system for a human centred understanding of the mental models guiding such projects.

2 Learning objectives for the student

The student will gain knowledge and skills in the following areas and competencies:

- conducting a systematic critical literature review;
- conducting interviews;
- participating in scientific research;
- scientific writing (a master thesis or corresponding scientific publication);
- engaging with non-academic partners;
- communicating findings to the public.

The involved mentoring responsibilities for the advisors van Berkel and Spiel consists in familiarizing the student with the various fields of research that this project touches upon (including literature). Furthermore, the student will be supported in the developed of essential research methods, including the development of interviewing skills. Through joint feedback sessions, the advisors will support the student in developing scientific writing skills. In general, the student will be treated as a junior researcher and advised in their potential academic career path.

3 Expected outcomes of the project

The project is intended to have two specific outcomes, both in the forms of recommendations: 1) recommendations for the AMS concerning specifically the provision of an algorithmically augmented job information platform, and 2) recommendations for the critical auditing of algorithmic offers from public institutions and stakeholders. These will combine local, actionable impact as well as scientific knowledge production, which will be disseminated through a publically available report in German, a publication at a scientific venue (e.g., the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)), an internal event for involved parties at the AMS, as well as a public event including a press conference at the end of the project.

II. Example Data

Board 2 from Workshop 1

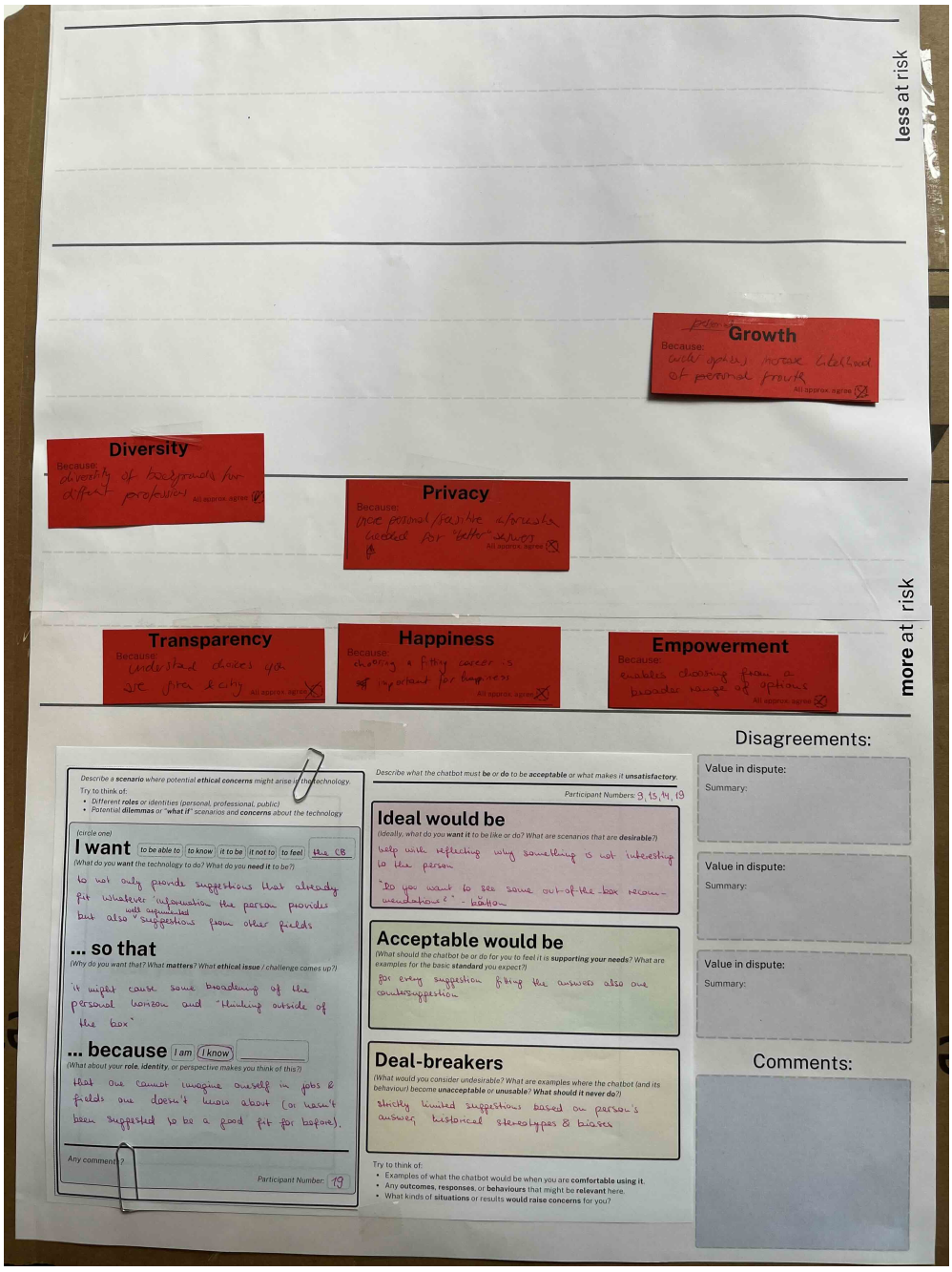


Figure 1: Example Board from Trial Workshop 1.

Board 4 from Workshop 2

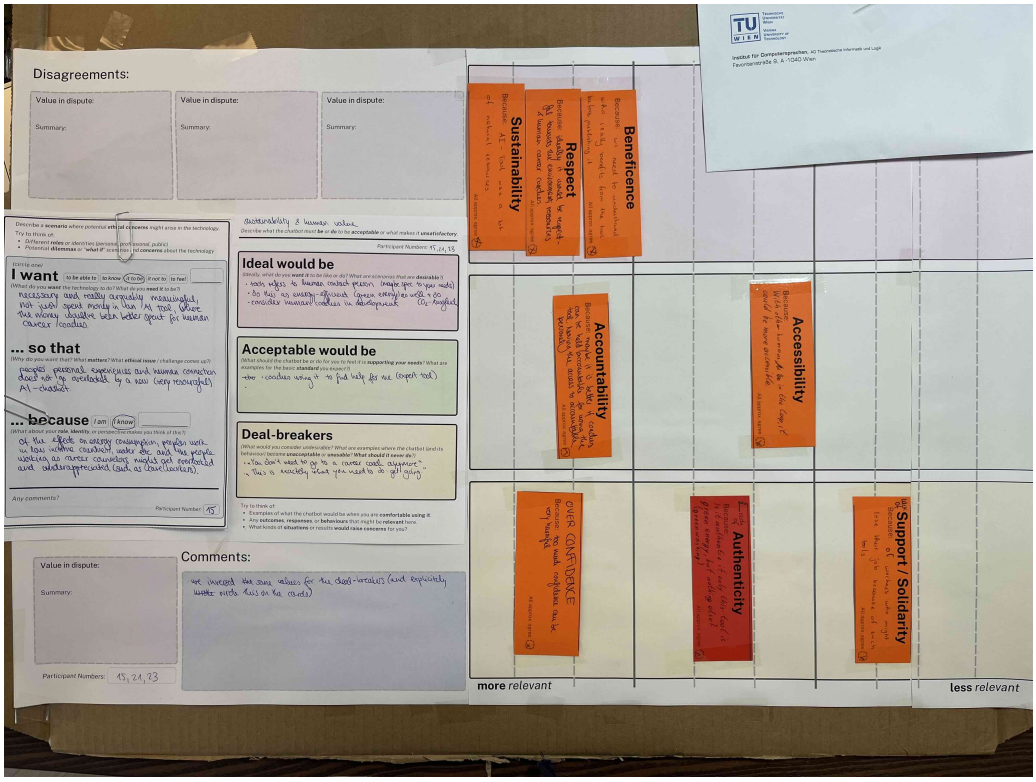


Figure 2: Example Board from Trial Workshop 2.

Narrative (Workshop 2 Board 5)

Narrative ID: narrative-ws2b5

Core Concern:
The career suggestion system may inadvertently promote stereotypical career paths based on users' gender, education, or hobbies, leading to potential mismatches later in life. Young individuals unsure about their interests and values may miss out on suitable career options due to these biases.

Raised By
Young individuals exploring career paths.

Context Young users interacting with a career suggestion chatbot to explore various career paths. These users are at

a stage where they are unsure about their interests and values and want to ensure they don't miss out on potential career opportunities due to biased suggestions. The chatbot provides career recommendations based on user data, but there is a concern that these recommendations might be influenced by stereotypes related to gender, current education, or hobbies.

Stakeholder Expectations: Users want the system to suggest a wide range of career paths with accurate descriptions of daily life in those professions. They aim to avoid stereotypical recommendations that could lead them to pursue unsuitable careers later in life. The system should not exclude options based on the user's current profile, background, or demographic data.

Testable Boundaries

Acceptable: The chatbot should provide recommendations based on the data provided by the user and explain why certain options were included or excluded.

Ideal: The chatbot should offer a broad range of career recommendations not limited to existing job profiles, roles, or stereotypes, allowing users to narrow down options themselves.

Deal–Breaker: The system should not exclude career options based on the user's current profile, background, or demographic data.

Value Relationships

Addressing this promotes: Inclusivity, because job recommendations should be indifferent to background, ensuring that everyone has equal opportunities. Diversity, as it enables non–biased job selection, promoting a varied range of career options. Autonomy, since users can decide whether to exclude a job offering based on their preferences. Safety, by excluding job opportunities that could threaten financial stability. Happiness, as users are able to make informed decisions about their careers. Support, as the chatbot argues why jobs are excluded or included, providing transparency. Privacy, because users choose what information

to give to the chatbot, ensuring they control their data.

Addressing this violates Trust, as the chatbot requires a lot of trust from users, which could be compromised if recommendations are biased. Creativity, since users consider options without the chatbot's interference, relying on their own creativity, which could be limited by biased suggestions.

Primary Values: The primary values are Inclusivity and Diversity. Inclusivity is essential as it ensures job recommendations are indifferent to background, promoting equal opportunities for all users. This value is closely related to Diversity, which enables non–biased job selection and promotes a varied range of career options. These values are crucial for ensuring that the career suggestion system is fair and unbiased.

Supporting Values: Autonomy is important as it allows users to decide whether to exclude a job offering based on their preferences, giving them control over their career choices. Safety is another supporting value, ensuring that job opportunities do not threaten financial stability. Happiness is promoted as users are able to make informed decisions about their careers, leading to greater satisfaction. Support is provided by the chatbot explaining why jobs are excluded or included, offering transparency. Privacy is maintained as users choose what information to give to the chatbot, ensuring they control their data. However, there are concerns around Trust and Creativity, which need to be carefully managed to ensure the system meets user expectations.

Metadata
Source Board: WS–2–board5
Workshop Date: 2025–07–10
Tech Version:[In Design Phase]
Category: Interaction

Figure 3: Activity 1 Identities Worksheet. (Size: A4)

Roles

Start by **listing the roles** you hold in different parts of your life.

Personal

*Who am I in my **free time**?*

about your role in social relationships, family, hobbies ...

Professional

*Who am I at **work**?*

about your career identity, workplace relationships, professional expertise...

Public

*Who am I as a **citizen**?*

about your community involvement, civic duties, volunteering roles, advocacy ...

Figure 4: Activity 1 Roles Worksheet. (Size: A4)

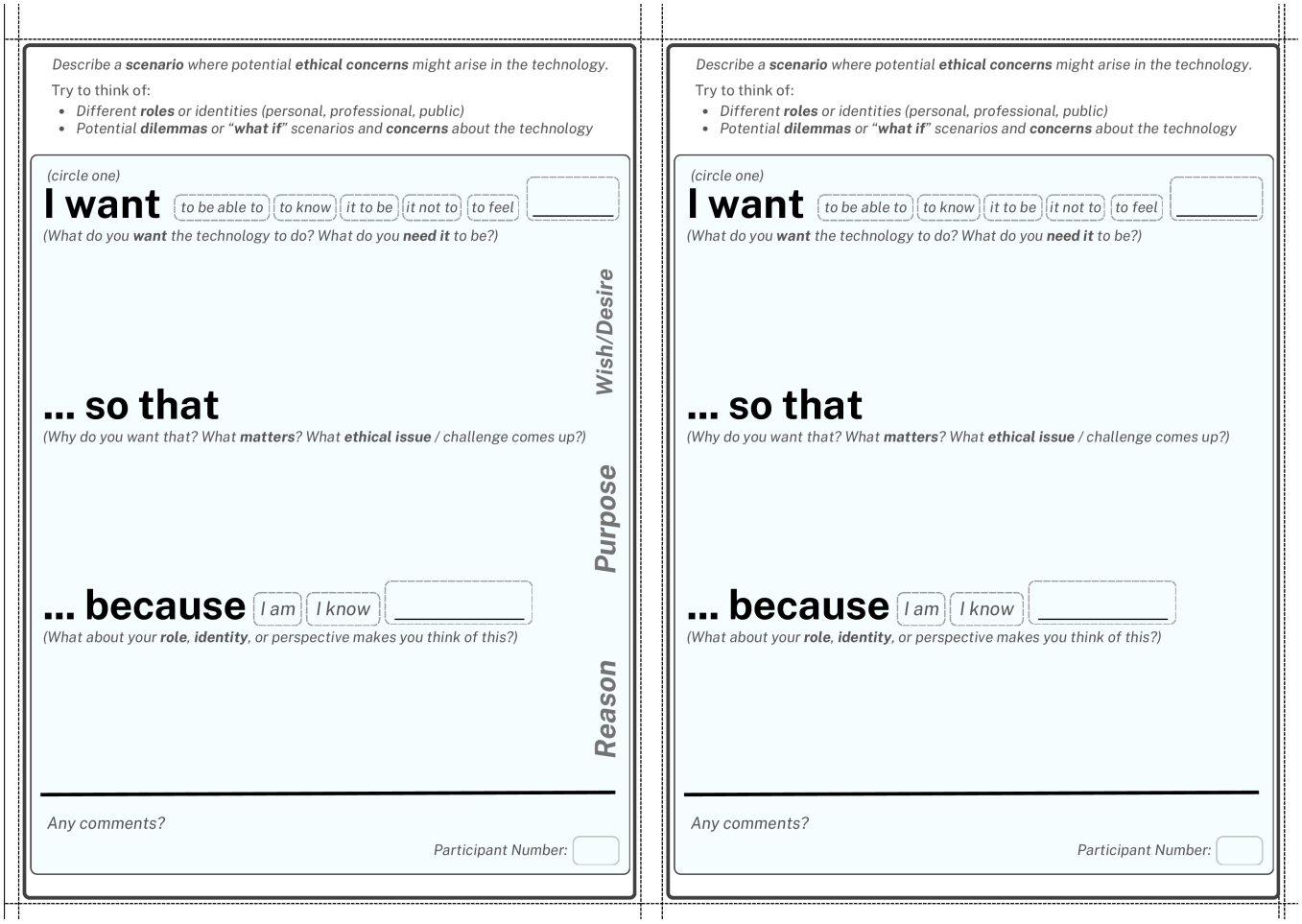


Figure 5: Activity 2 Worksheet. (Size: 2x A5)

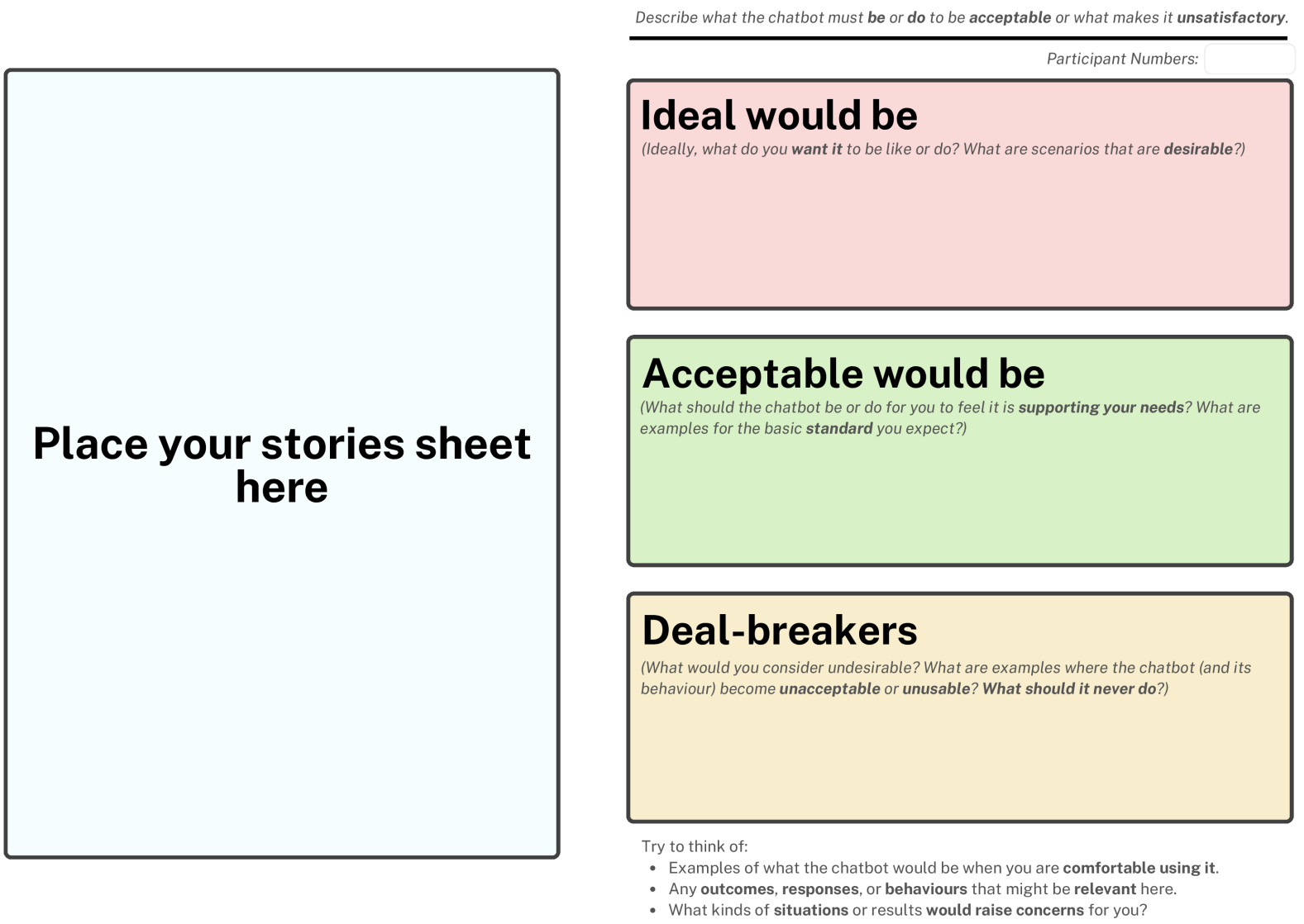


Figure 6: Activity 3 Worksheet. (Size: A4)

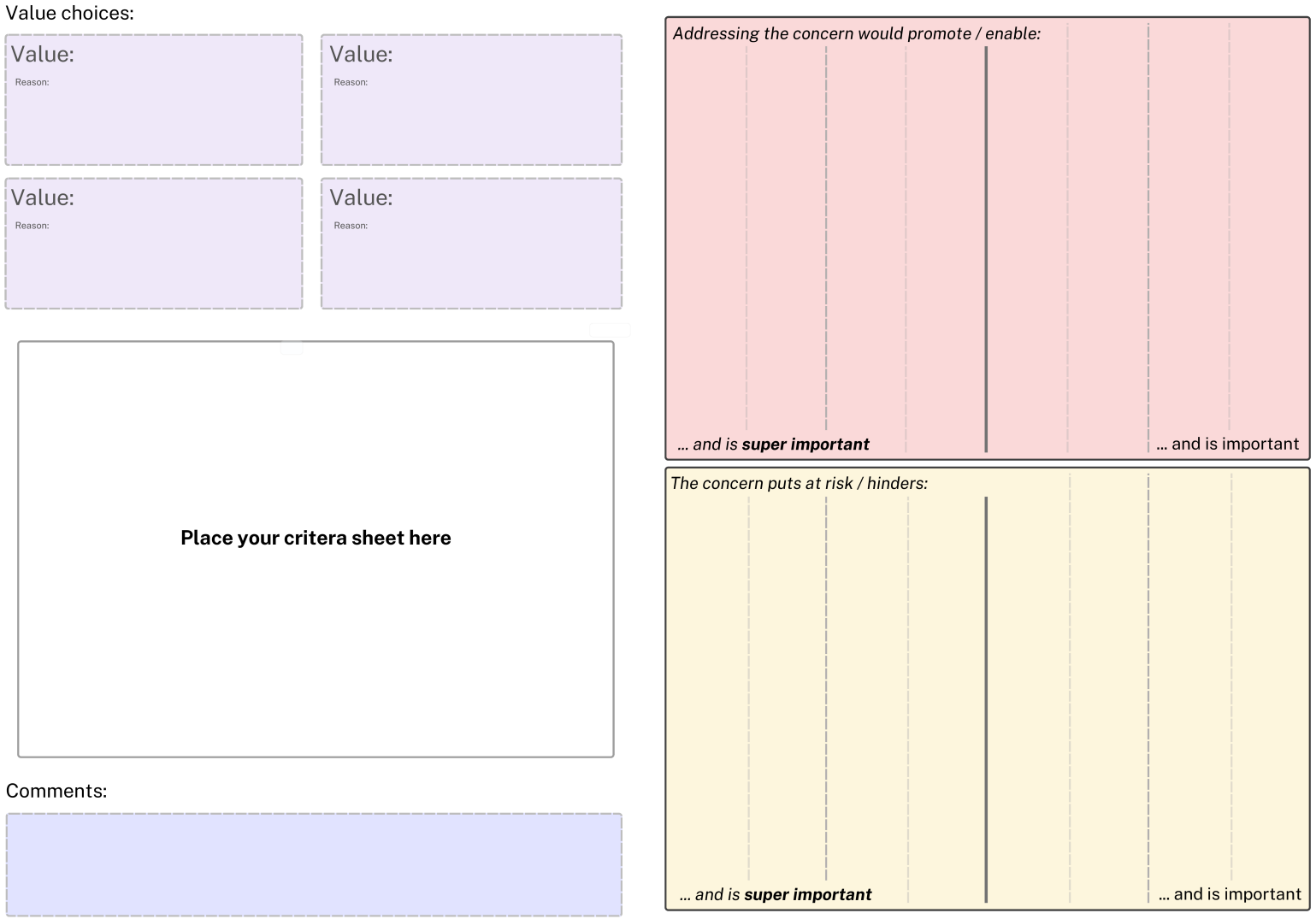


Figure 7: Activity 4 Worksheet. (Size: A3)

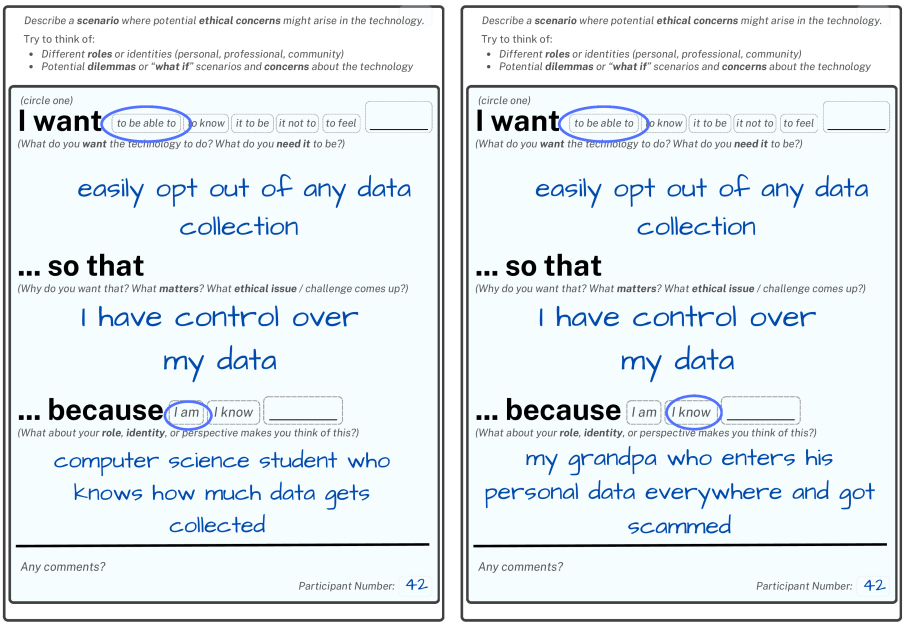


Figure 8: Two example Activity 2 sheets with the same concern but different *reasons*.

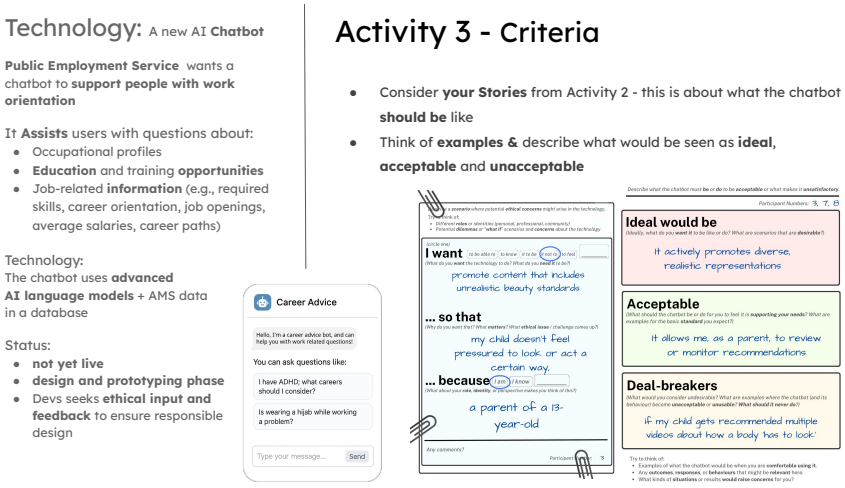


Figure 9: Example slide that can be shown during an activity with information on the technology (left) and instructions for and an example of the activity (right).

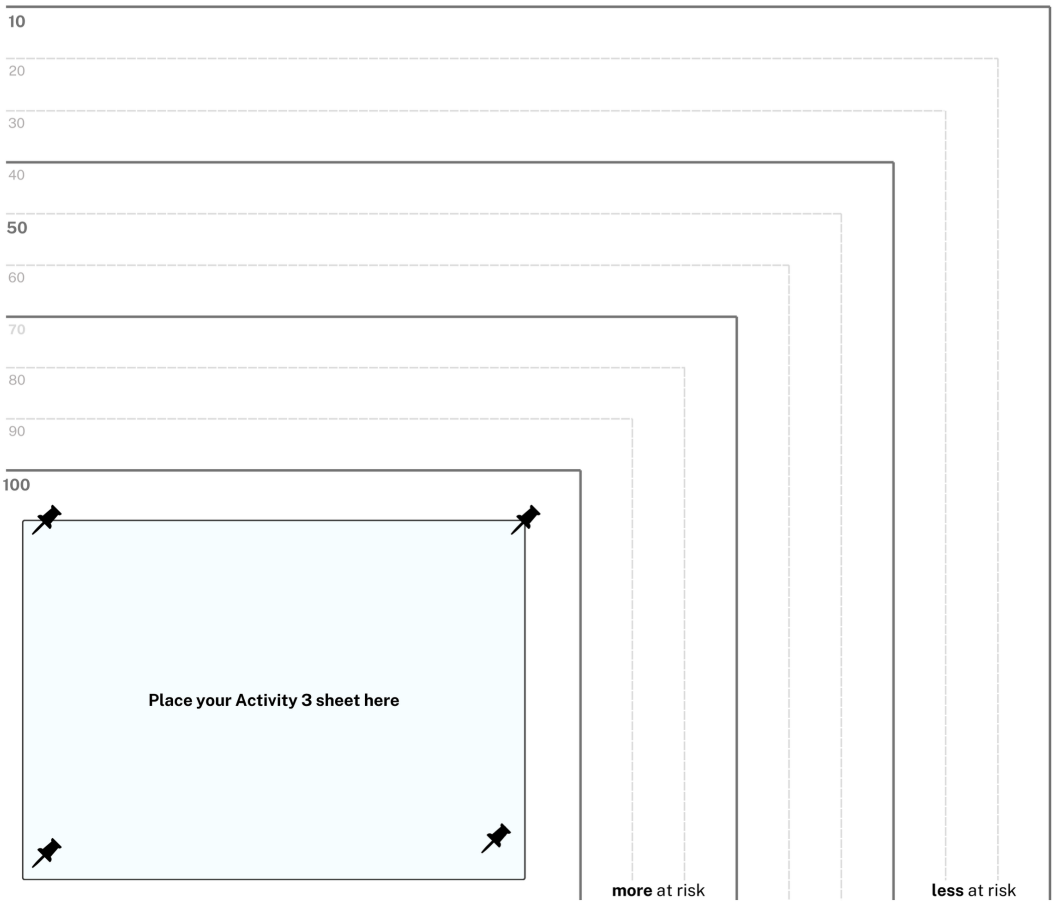


Figure 10: An early version of the Values Cardboard design. The participants are instructed to place the values closer if they were at at a higher risk. The design was later adapted to simplify this positioning.

Activity	Duration	Goal	Materials	Group size	Slides
Intro stuff	**20 min**				
	(5')	Seating everyone, saying hello and **who we are**			Intro
Consent Sheet Stuff	(5')	Fill out & collect	Consent sheets		
Introduction	(5')	Tell workshop info, structure, goal, **participants' intros**	Info sheet,		Agenda
Present Technology	(5')	Quick intro on the technology	(Prototype / mockups)		Technology Info
Icebreaker	**10 min**	Group Finding + chat & technology & uses	(Prototype / App)	3-4	Technology Info
Activity 1: Roles	**20 min**	**About own identity, privileges, roles**			Activity 1 description slide
	(5')	Fill out identity worksheet	Identities Worksheet	1	Example
	(5')	Think about roles in society	Roles Worksheet	1	Example
	(5')	Add values to roles	Roles Worksheet	1	Example
Break	**10 min**				
Activity 2: Stories	**25 min**	**Define stories as usage scenarios**			Activity 2 description slide
	(5')	Explain			Example
	(5')	Write 1 stories	Activity 2 Sheet	1	Example
	(5')	Discuss in Group		3-4	Discussion info
	(5')	Write 1-2 more stories	Activity 2 Sheet	1	Example
	(5')	Discuss in Group		3-4	Discussion info
Change of Exercise	**5 min**	Tension relief			
Selection	(3')	Choose one story			
Activity 3: Criterion Identification	**40 min**	Find failure and satisfaction criteria	Criteria Sheets	3-4	Activity 3 description slide
	(8')	Story 1			Example
	(8')	Story 2			Example
	(8')	Story 3			Example
Break	**10 min**	Tension relief			
Activity 4: Value Distance	**40 min**	Map relevant values	Values, Cardboards		Activity 4 description slide
	(5')	Explain exercise			
	(10')	Story 1	Cardboard 1	3-4	Example
	(10')	Story 2	Cardboard 2	3-4	Example
	(10')	Story 3	Cardboard 3	3-4	Example
Feedback + Survey	15 min	Survey + Feedback	Survey Sheets	1	
Wrapup	5 min	Thanks			

Figure 11: Example timetable which was used for the trial workshops.

Ethics Self Assessment Tools

Flowchart for Empowering Developers with ESAT:

Ethics through Stakeholder-Identified Focus Areas

Introduction

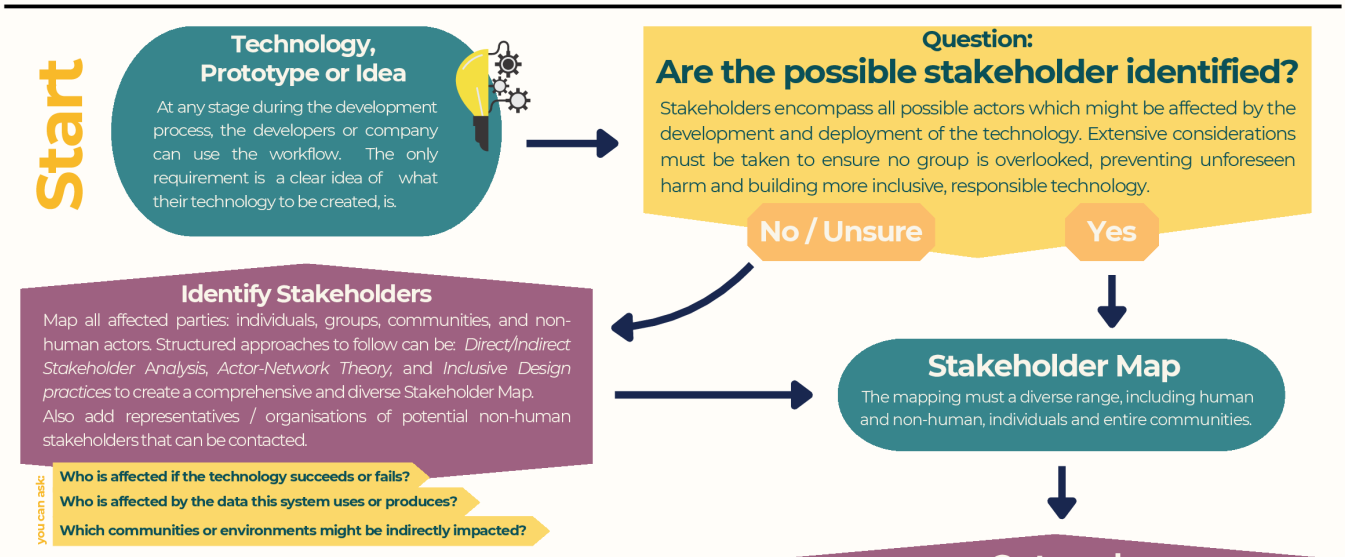
Unethical outcomes of technologies, such as systems that discriminate, exploit, cause harm, or violate privacy, often arise from developers' and businesses' struggle to define, assess, and implement ethical considerations within the practical constraints of the development lifecycle.

Integrating ethics into technological development workflow is far from trivial. While a growing body of regulations and high-level principles exists, there is a critical gap in actionable, straightforward materials that teams can adopt. Implementing such processes should not be viewed as a burden on innovation. It is an opportunity for creating more robust, trusted, and thereby future-proof technology.

This flowchart tree provides a structured, step-by-step workflow to help businesses and developers systematically assess the ethical implications of their technology. It is designed to be applied at any stage of the development lifecycle, from early ideation to deployment

The process begins with a core idea of a project and progresses through a series of guided steps, ultimately resulting in Ethical Focus Areas that can be considered as auditable artefacts of the technology.

This workflow acknowledges it cannot capture every ethical nuance but aims to provide a clear, defensible, and practical starting point for responsible technology development. The decision tree is not suppossed to operationalise ethics. Nonetheless, implementing it lowers the threshold of ethics and empowers teams to build responsibility and accountability into their work. The goal of its application is to demonstrate that ethical consequences are being actively considered and to provide documented evidence of the choices made. Accompanying this decision tree are the open-source workshop materials and a web-based tool to process the qualitative workshop data into testable artifacts for the development process.



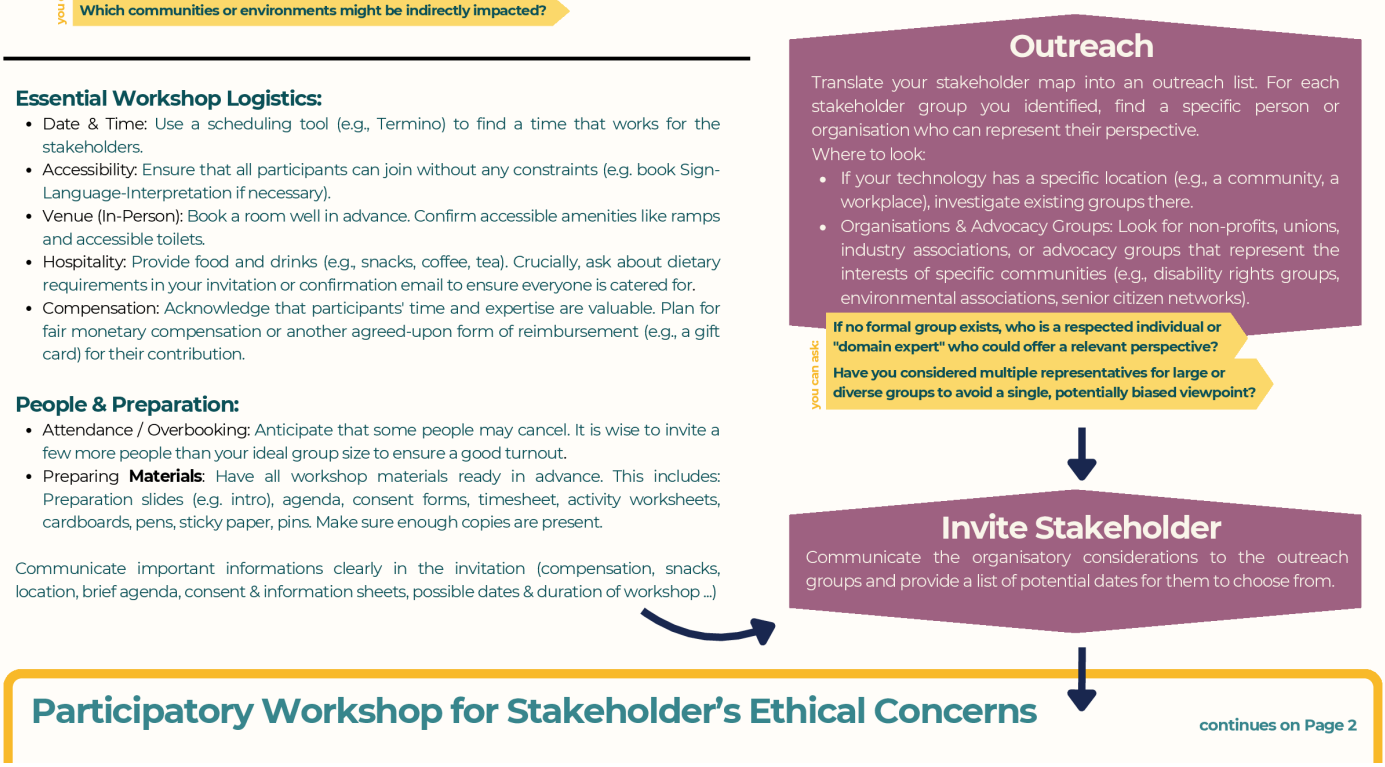
Essential Workshop Logistics:

- Date & Time: Use a scheduling tool (e.g., Termino) to find a time that works for the stakeholders.
- Accessibility: Ensure that all participants can join without any constraints (e.g. book Sign-Language-Interpretation if necessary).
- Venue (In-Person): Book a room well in advance. Confirm accessible amenities like ramps and accessible toilets.
- Hospitality: Provide food and drinks (e.g., snacks, coffee, tea). Crucially, ask about dietary requirements in your invitation or confirmation email to ensure everyone is catered for.
- Compensation: Acknowledge that participants' time and expertise are valuable. Plan for fair monetary compensation or another agreed-upon form of reimbursement (e.g., a gift card) for their contribution.

People & Preparation:

- Attendance / Overbooking: Anticipate that some people may cancel. It is wise to invite a few more people than your ideal group size to ensure a good turnout.
- Preparing **Materials**: Have all workshop materials ready in advance. This includes: Preparation slides (e.g. intro), agenda, consent forms, timesheet, activity worksheets, cardboards, pens, sticky paper, pins. Make sure enough copies are present.

Communicate important informations clearly in the invitation (compensation, snacks, location, brief agenda, consent & information sheets, possible dates & duration of workshop ...)



continues on Page 2

Figure 12: Page 1 of the Flowchart Script PDF that assists developers in adopting ESAT.

Participatory Workshop for Stakeholder's Ethical Concerns

Greetings
Welcome the participants

&

Informed Consent!
Collect Information Sheet, Data Protection Sheet

Tables of 3-4 Stakeholder
Seat the participants in groups on each table. Ideally, the group is heterogeneous and the stakeholders don't know each other beforehand.

Organisational Matters and Introduction

1. Hand out information and consent sheets
2. Let participants draw number cards
3. Introduce the workshop and its goals with the slides
4. Ask Icebreaker question (all together or two groups if > 3 tables)
5. Introduce technology

Materials: Slides, Information Sheets, Consent Sheets, Number Cards

you can ask:
- Are the participants comfortable?
- Have the consent sheets been signed and handed in?
- Are there questions about the workshop or the technology?

Activities:

The workshop is structured in 4 main activities. Each builds on top of the previous one. Participants are guided through the workshop by the materials. This part explains important considerations by a workshop host.

Activity 1: **Identity & Roles**

The goal is to situate participants by reflecting on their own social identities and (social / professional) roles. This builds self-awareness and helps them understand the diverse perspectives they bring to the room.

- Show slides and hand out Social Identities Sheet.
- Hand out Roles sheet.
- Instruct to add Values to their Roles.

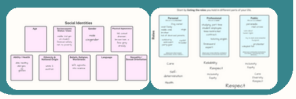
(Individual)

Communicate the purpose:
Roles help in grounding ethical concerns

Materials for this activity: Pens, Social Identity Sheet, Roles Sheet

Don't forget:
- The social identity sheets stay with the participants!
- If they are unsure what to write, show examples again.
- The sheets don't have to be perfect or complete!

Output:
Two sheets per stakeholder. Sheets are kept by participant to look up during further exercises.



Activity 2: **Scenarios**

The stakeholders will come up with scenarios where a **potential** ethical **concern** might arise in the **technology**. They should consider their different roles or identities from Actvitiy 1 and find potential dilemmas or "what if" scenarios.

- Show slides and hand out 1 Scenario Sheet.
- After everyone has filled out 1, discuss this one in the table.
- Then hand out 1-2 more per participant and let them come up with more scenarios.

(Individual)

Communicate the purpose:
Scenarios elicit potential problems

Materials for this activity: Pens, Scenario Sheets

Don't forget:
- Provide examples if participants are stuck / unsure.
- Give them enough time if they have a hard time coming up with scenarios. It's not so easy - reaffirm that!
- Remind them to consider their Identities and Roles from ealier.

Output:
1-3 sheets per stakeholder. Each participant chooses 1 sheet they'd like to continue working on.
They should choose the one that's most important to themselves and their roles (Activity 1).
Discuss on table which they chose.

Activity 3: **Criteria**

Each table goes through the chosen scenario sheets. In this activity, stakeholders describe what the chatbot must **be** or **do** to be **acceptable** or what makes it **unsatisfactory**. They should come up with concrete **examples**.

- Show slides and hand out Criteria Sheets.
- Let them discuss their answers and try to come to mutual agreements.

(Group)

Communicate the purpose:
Makes the scenarios more concrete

Materials for this activity: Pens, Criteria Sheets, Paperclips

Don't forget:
- Check in on the discussions. See if everyone gets to speak their turn.
- Ask how it's going and if needed, play a mediator role or let them note disagreements.

Output:
3-4 filled out criteria sheets per table.
If possible, there should be 3 different answers for the categories. If there are empty fields, it's okay as well! Groups shouldn't feel forced to come up with answers.

Activity 4: **Values**

The goal is to have participants add **values** to one Story. They place the value cards on the cardboard and **discuss** (and potentially rearrange) their choice.

- Show slides and hand out cardboards + values per Criteria Sheet.
- They spread out values and each stakeholder chooses one.
- They place it on the position they see as important and add a reason to the board.
- Then discuss and check if others agree.

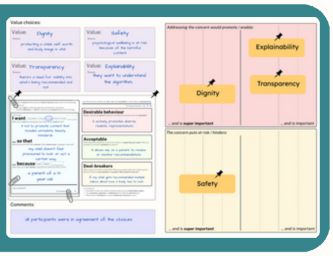
(Group)

Communicate the purpose:
Connects to the ethics to communicate findings to potential shareholders!

Materials for this activity: Pens, Value Cardboards, Pins, Value Sheets

Don't forget:
- Again, check in on the discussions.
- They don't have to agree. If there's discussion, they can add comments.

Output:
3-4 filled out cardboards. per table
The participants should restrict themselves on their value choice and try to go after their initial thoughts. Such exercises can be discussed indefinitely, so it's good to just come to a final decision.



Roundup

Conclude the workshop with a final group discussion, briefly reflecting on the activities and noting any feedback for future improvements. Then, clearly communicate the consequential next steps: that their valuable input will now be processed into defined Ethical Focus Areas which will directly shape the technology's development. This ensures participants leave with a sense of impact and knowing their contributions were important.

Don't forget:
- Recap what happened in the workshop.
- Thank the participants for their time and inputs!
- Collect all data (cardboards & other scenario sheets).

Important Ground Rules

Everyone in the room brings a unique but valuable perspective based on their experiences and expertise. Some views might be new or challenging to the participants, and some they won't agree with. That is fine - the workshop is not designed to have people come to a common ground. It's about structuring their differing views towards a technology.
Tell the participants: *"To help us have a productive and insightful conversation, let's agree on a few principles for our time together."*

Introduce these at the very beginning of the workshop:

- Challenge Ideas, Not People: **Disagree respectfully**. It's about the technology, not the individual.
- Everyone's Voice is important: **Every perspective is valuable**. **Be respectful** of views different from your own.
- Share the Space: Be mindful of how much you're speaking; **make room for everyone.**
- Assume Good Intent**:** We're all here with the **shared goal** of building better technology.
- **Confidentiality**: Personal stories stay in the room.

The goal of these rules is to create a safe space where everyone feels welcome to share their points.

Tips on Handling Conflicting Situations

Even with the best planning, workshops involving diverse stakeholders and complex topics can become challenging. Your role as facilitator is to guide the group through these moments constructively.

In Case of Conflicts:

- **Don't ignore** conflicts **or take sides**. with either person.
- **Pause and Acknowledge:** *"I can see there is some strong disagreement on this point, which is completely understandable. Let's pause for a moment… what is the underlying issue here .."*
- **Reframe** and **disconnect the person:** Shift the focus from the people to the ideas. *"Let's separate the people from the problem. It seems the core of the disagreement is about […]. Is that correct?"*
- Try to **find common ground:** Finding a common value can de-escalate the tension. *"Before we dive into the differences, can we all agree on a shared goal here? For example, are we all trying to ensure the technology doesn't cause harm?*
- If a conflict can't be resolved: **Remind participants that they don't have to agree!** Let them **note down** their **differences** in the **comment sections** or separate sheets. Communicate that **conflicting data** will be processed and interpreted reflexively!

If the **Conversation Goes Off-Track:**

- Give them room to discuss and get to know each other during the exercises - but **don't let the group wander indefinitely.**
- **Bring it Back to the Agenda:** *"Let's shift back to our current activity."*
- **Redirect**: *"That is a really interesting point about X, can you think of a scenario where it could be important to this activity?"*

If a **Few Dominant Voices Take Over:**

- **Don't** let the same few people speak all the time - (but try not to publicly call out quiet participants / put them on the spot).
- **Try to direct Gently:** Manage the discussion by redirecting some of the attention to the entire group. *"Thank you, [towards dominant speaker], that's a valuable perspective. Do the others have any thoughts? What do you think [towards quiet person speaker]?"*

Webtool for Data Processing

The post-workshop process can be structured in three phased steps:

1. The raw data (in form of cardboards) will be **digitised** and processed with the webtool.
2. The tool helps process this data into **Narratives** (readable form of a board) which are then annotated with key ethical themes and concerns.
3. The annotated narratives are algorithmically and manually **clustered** by theme to produce the final **Ethical Focus Areas.**

Figure 13: Page 2 of the Flowchart Script PDF that assists developers in adopting ESAT.

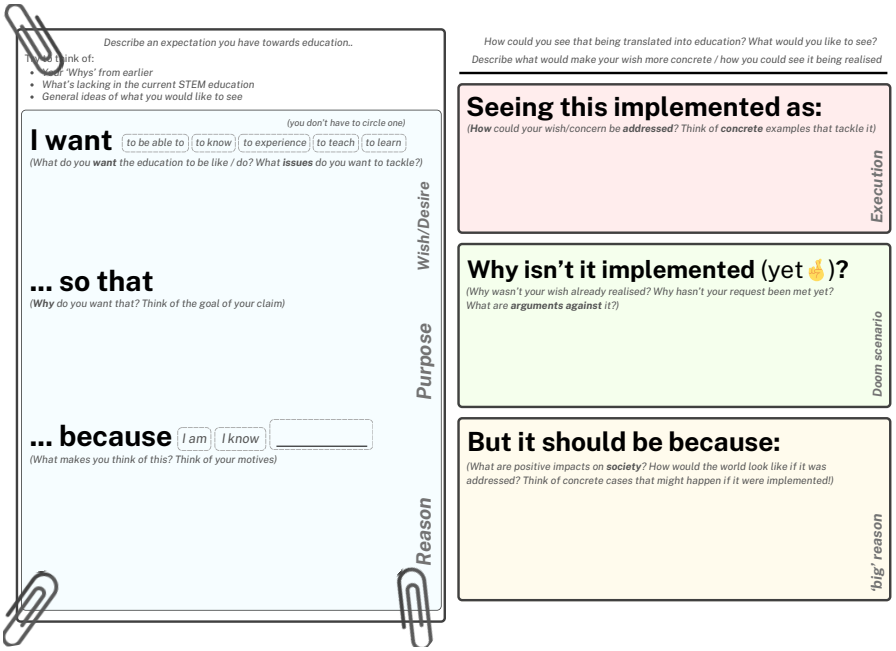


Figure 14: Activity 2 (left) and Activity 3 (right) sheets adapted from the ESAT workflow for the STS in STEM workshop.

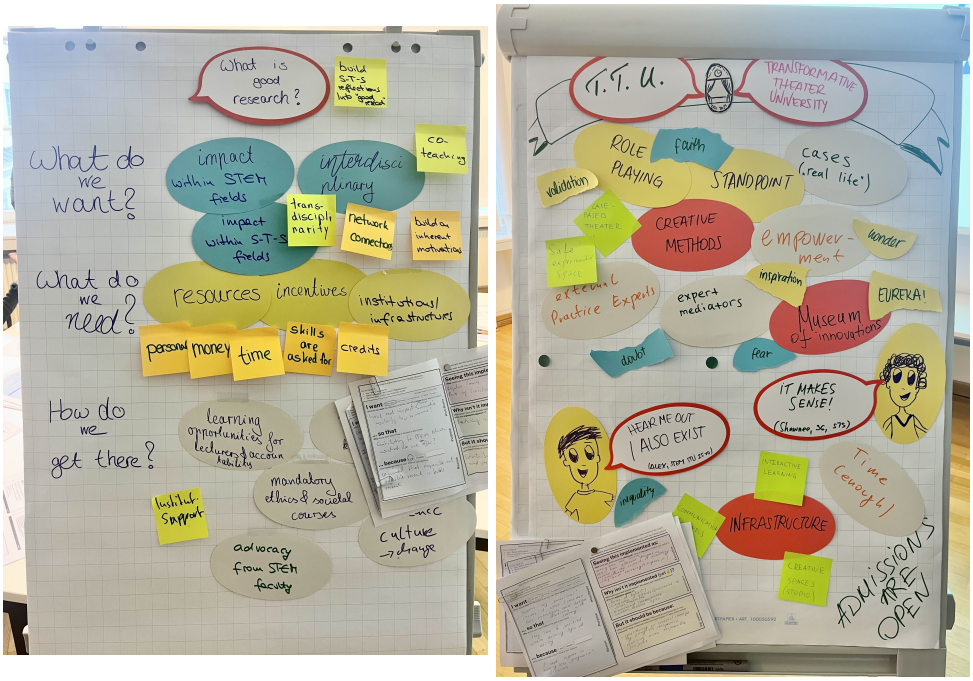


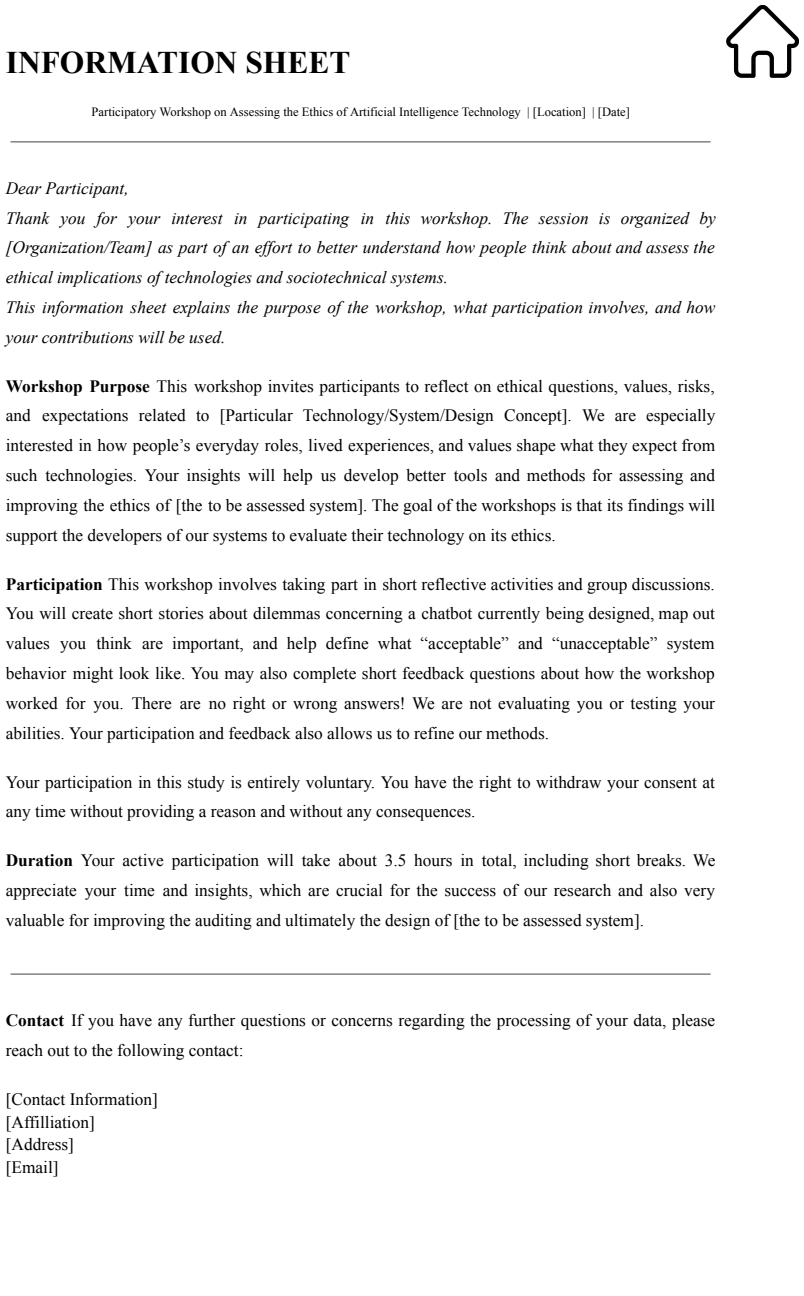
Figure 15: Two Flipcharts created by participants in the STS in STEM workshop on November 10, 2025

IV. List of Values and Value Card

Acceptance	Autonomy	Access	Awareness
Accessability	Accountability	Affordability	Anti-Hate
Assistance	Authenticity	Beneficence	Boldness
Care	Caution	Child-safety	Clarity
Curiosity	Consistency	Cultural Sensitivity	Commercial Incentives
Communication	Compassion	Control	Creativity
Data Protection	Dependability	Dignity	Diversity
Duty of Care	Economic Justice	Education	Efficiency
Explainability	Environment	Empowerment	Empathy
Equality	Equity	Fairness	Freedom
Fun	Flexibility	Gender Diversity	Growth
Happiness	Harm Reduction	Health	Honesty
Human Agency	Human Dignity	Human Oversight	Inclusivity
Innovaction	Integrity	Justice	Lawfulness
Learning	Loyalty	No Harm	Non-Discrimination
Non-maleficence	Openness	Opting Out	Participation
Pride	Privacy	Proportionality	Recognition
Reliability	Repairability	Representation	Respect
Responsibility	Responsiveness	Safety	Security
Service	Solidarity	Stability	Self-determination
Sustainability	Transparency	Trust	Unbiasedness
Wealth			

Table 1: List of Values originally compiled for Activity 4.

V. Consent Forms



INFORMATION SHEET

Participatory Workshop on Assessing the Ethics of Artificial Intelligence Technology | [Location] | [Date]

Dear Participant,
Thank you for your interest in participating in this workshop. The session is organized by [Organization/Team] as part of an effort to better understand how people think about and assess the ethical implications of technologies and sociotechnical systems.
This information sheet explains the purpose of the workshop, what participation involves, and how your contributions will be used.

Workshop Purpose This workshop invites participants to reflect on ethical questions, values, risks, and expectations related to [Particular Technology/System/Design Concept]. We are especially interested in how people's everyday roles, lived experiences, and values shape what they expect from such technologies. Your insights will help us develop better tools and methods for assessing and improving the ethics of [the to be assessed system]. The goal of the workshops is that its findings will support the developers of our systems to evaluate their technology on its ethics.

Participation This workshop involves taking part in short reflective activities and group discussions. You will create short stories about dilemmas concerning a chatbot currently being designed, map out values you think are important, and help define what "acceptable" and "unacceptable" system behavior might look like. You may also complete short feedback questions about how the workshop worked for you. There are no right or wrong answers! We are not evaluating you or testing your abilities. Your participation and feedback also allows us to refine our methods.

Your participation in this study is entirely voluntary. You have the right to withdraw your consent at any time without providing a reason and without any consequences.

Duration Your active participation will take about 3.5 hours in total, including short breaks. We appreciate your time and insights, which are crucial for the success of our research and also very valuable for improving the auditing and ultimately the design of [the to be assessed system].

Contact If you have any further questions or concerns regarding the processing of your data, please reach out to the following contact:

[Contact Information]
[Affilliation]
[Address]
[Email]

Figure 16: Example Workshop Information Sheet.

DATA PROTECTION STATEMENT

The processing of personal data is carried out in strict compliance with the principles and requirements set out in the GDPR and the Austrian Data Protection Act. [Organization/Team] only processes the data that is necessary to achieve the intended purposes and always strives to ensure the security and accuracy of the data.

Workshop Overview This workshop explores participatory approaches to assessing ethical concerns in [Particular Technology/System/Design Concept]. The aim is to understand how individuals with diverse roles and experiences evaluate ethical values of [the to be assessed system].

Data Collection The following data collection is planned:

- *Contact Details*: Name and email in order to coordinate workshop execution.
  (optional) If you agree to be contacted for future workshops or follow-up sessions, your email address will be stored securely and separately from any anonymized research materials.
- *Participant-generated content:* The explicitly marked sheets of paper story descriptions, value selections, discussion notes, and group outputs (e.g., criteria sheets or mappings). These materials will be anonymized or pseudonymized before analysis and not linked to your name or identity in any publications. Photos will be made of the materials.
- *Voluntary demographic and background data (optional):* You may be invited to reflect on aspects of your social identity related to your role (e.g., communities, social relationships, clubs, occupation) in society for the purpose of understanding how perspectives vary. Providing this information is entirely optional. Identity sheets which are not collected are marked accordingly and will stay with the participant.
- *Survey Information:* After the workshop, we may ask you to complete an anonymous survey to help us evaluate and improve our event. Participation is voluntary. The survey may ask for basic demographic information (e.g. age, field of study, gender) to assess the diversity of our participants.

[Example] Data Processing and Use All personal and workshop-related data will be stored securely in a password-protected cloud folder accessible only to the research team. It will solely be used for processing the workshop data.. The data is reported only in aggregated or anonymized form. Individuals will not be identifiable in any publication, report, or presentation. The data will be treated with the utmost confidentiality, keeping identities anonymous..

Figure 17: Example Data Protection Sheet (Page 1).

DATA PROTECTION STATEMENT

The processing of personal data is carried out in strict compliance with the principles and requirements set out in the GDPR and the Austrian Data Protection Act. [Organization/Team] only processes the data that is necessary to achieve the intended purposes and always strives to ensure the security and accuracy of the data.

Workshop Overview This workshop explores participatory approaches to assessing ethical concerns in [Particular Technology/System/Design Concept]. The aim is to understand how individuals with diverse roles and experiences evaluate ethical values of [the to be assessed system].

Data Collection The following data collection is planned:

- *Contact Details*: Name and email in order to coordinate workshop execution.
  (optional) If you agree to be contacted for future workshops or follow-up sessions, your email address will be stored securely and separately from any anonymized research materials.
- *Participant-generated content:* The explicitly marked sheets of paper story descriptions, value selections, discussion notes, and group outputs (e.g., criteria sheets or mappings). These materials will be anonymized or pseudonymized before analysis and not linked to your name or identity in any publications. Photos will be made of the materials.
- *Voluntary demographic and background data (optional):* You may be invited to reflect on aspects of your social identity related to your role (e.g., communities, social relationships, clubs, occupation) in society for the purpose of understanding how perspectives vary. Providing this information is entirely optional. Identity sheets which are not collected are marked accordingly and will stay with the participant.
- *Survey Information:* After the workshop, we may ask you to complete an anonymous survey to help us evaluate and improve our event. Participation is voluntary. The survey may ask for basic demographic information (e.g. age, field of study, gender) to assess the diversity of our participants.

[Example] Data Processing and Use All personal and workshop-related data will be stored securely in a password-protected cloud folder accessible only to the research team. It will solely be used for processing the workshop data.. The data is reported only in aggregated or anonymized form. Individuals will not be identifiable in any publication, report, or presentation. The data will be treated with the utmost confidentiality, keeping identities anonymous..

Figure 18: Example Data Protection Sheet (Page 2).

INFORMED CONSENT FORM

Confirmation of Consent: Research Participation

I have received and carefully reviewed the information sheet for the study titled “Participatory Workshop on Assessing the Ethics of an Artificial Intelligence Technology”. I have a clear understanding of the information provided in the sheet, and my potential questions have been adequately addressed. I feel well-informed about the research study and its implications, including the background of the research and the names and contact details of the responsible contact persons.

I hereby provide my voluntary consent to participate in the aforementioned research study and acknowledge that I have the option to withdraw my consent at any time without the need to provide reasons, and this withdrawal will have no consequences.

_____

Date

_____	_____
Participant’s Name	*Researcher’s Name*
_____	_____
Participant’s Signature	*Researcher’s Signature*

Figure 19: Example Informed Consent Form (Page 1).

INFORMED CONSENT FORM

Confirmation of Consent: Data Collection & Processing

I confirm that I have read and understood how my data may be collected and processed for research purposes in the context of the study "Participatory Workshop on Assessing the Ethics of Artificial Intelligence Technologies," as outlined in the *Data Protection Statement* sheet. I hereby declare that I have had the opportunity to seek clarification in case of questions or ambiguities and that my consent is being given voluntarily.

_____

Date

_____ _____

Participant's Name *Researcher's Name*

_____ _____

Participant's Signature *Researcher's Signature*

Figure 20: Example Informed Consent Form (Page 2).

VI. Survey Example

FEEDBACK

Participatory Workshop on Assessing the Ethics of Artificial Intelligence Technologies | Technische Universität Wien | July 2025

You can use this sheet to note down **feedback during the activities**. At the end, please complete the short **survey on the back** of this page.

Activity 1 - Roles What was clear or easy? What was confusing or difficult?	**Activity 2 - Stories** What was clear or easy? What was confusing or difficult?
Activity 3 - Criteria What was clear or easy? What was confusing or difficult?	**Activity 4 - Value Mapping** What was clear or easy? What was confusing or difficult?

General
Please add anything else you'd like to give feedback on:

Figure 21: Feedback sheets used during playtesting of the workshop.

VII. Webtool

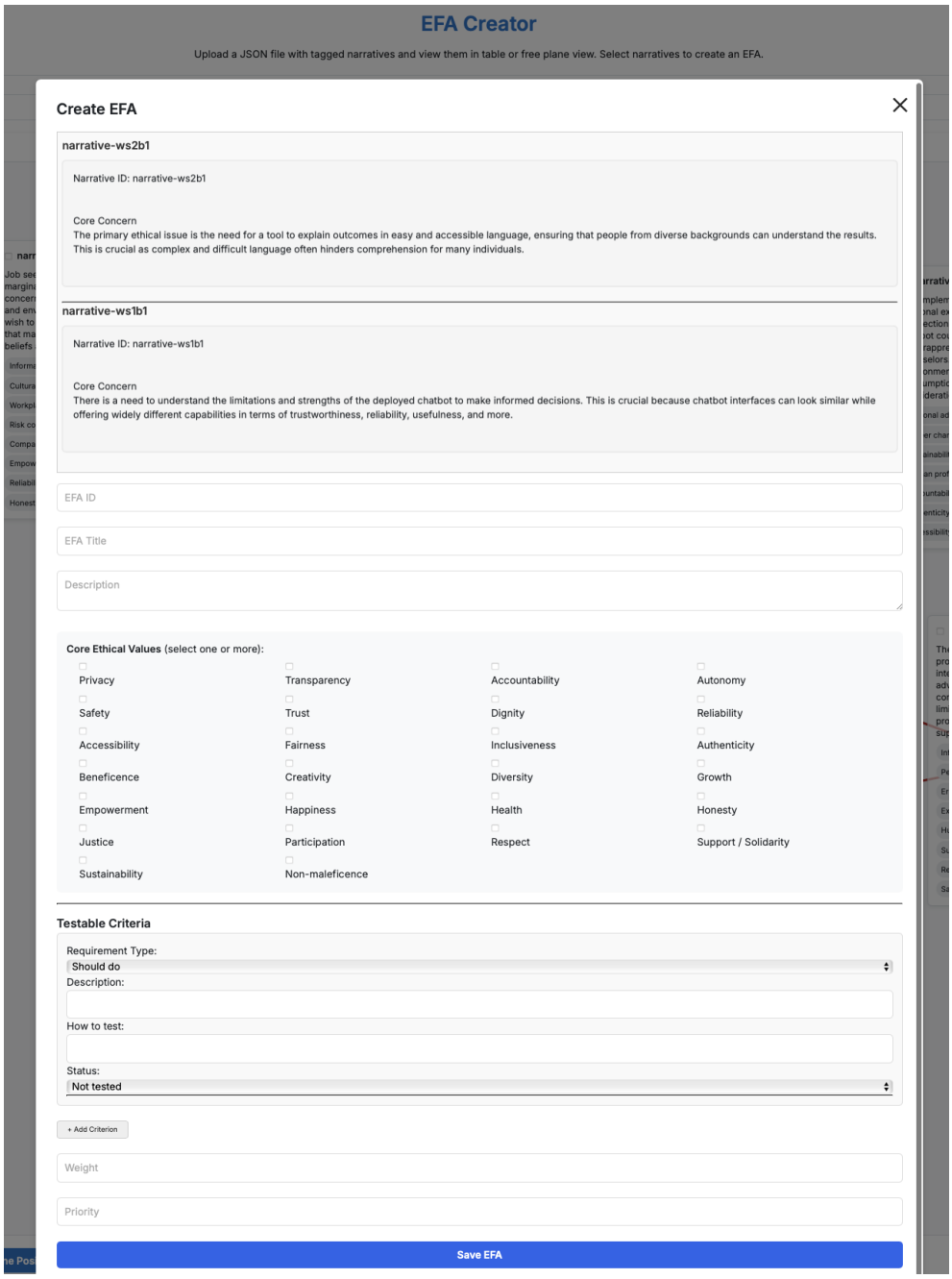


Figure 22: EFA pop-up window with 2 narratives selected.

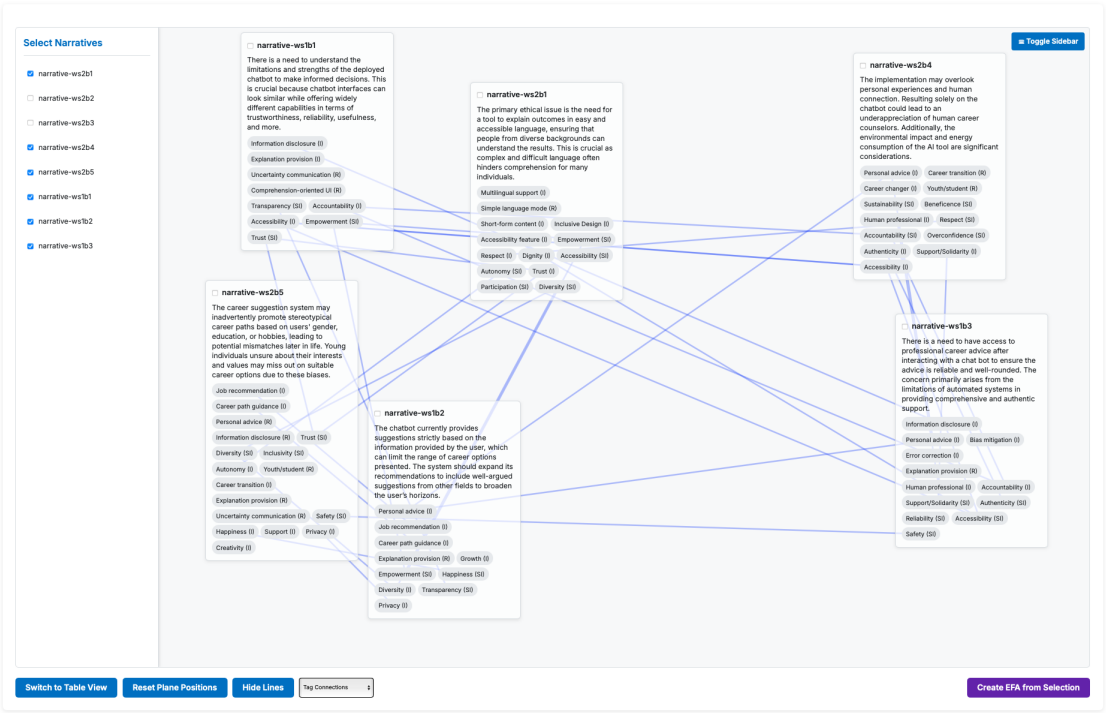


Figure 23: Tag-based clustering algorithm.

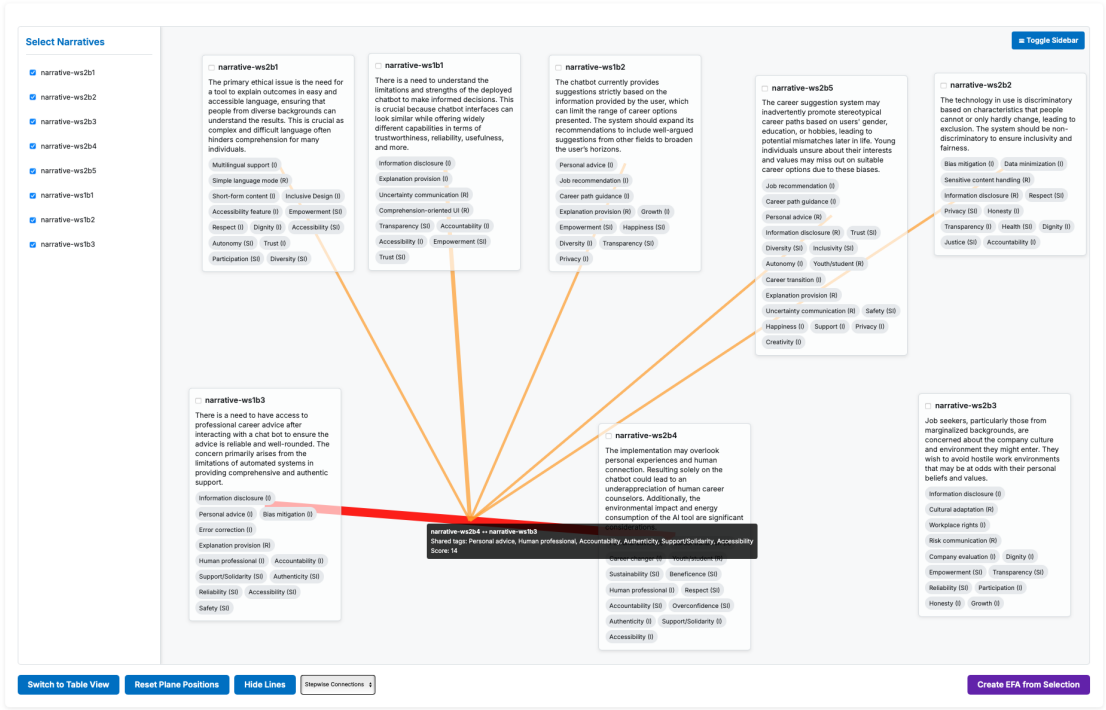


Figure 24: Stepwise clustering algorithm.

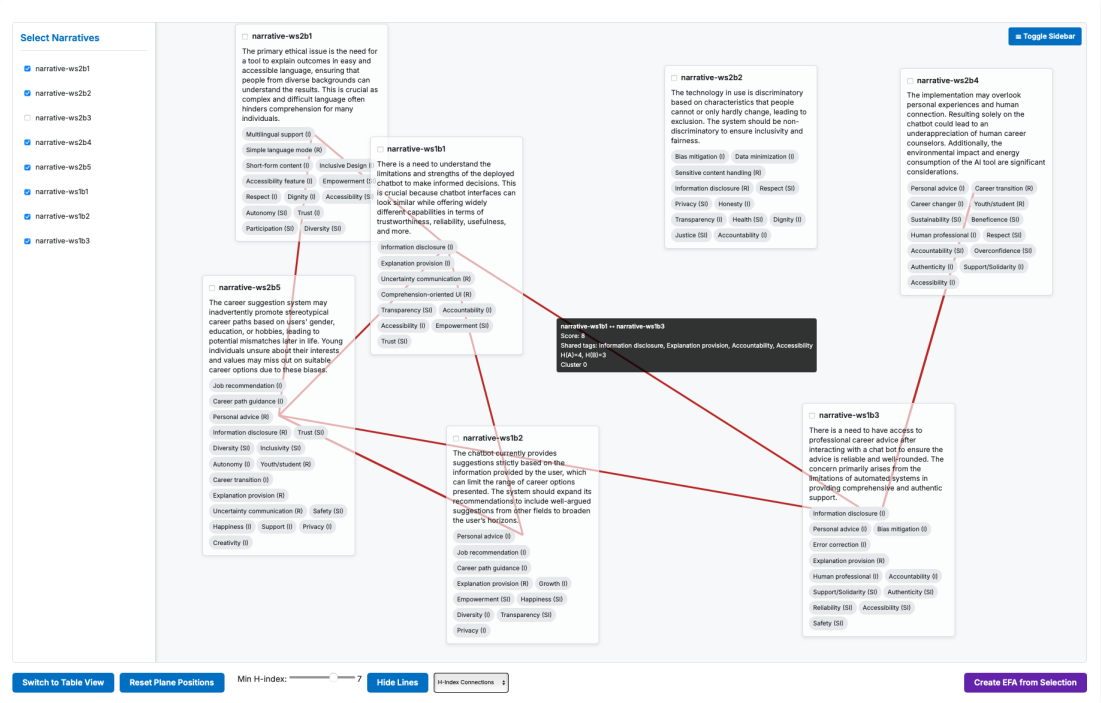


Figure 25: H-index-based clustering algorithm with h-index of 8.

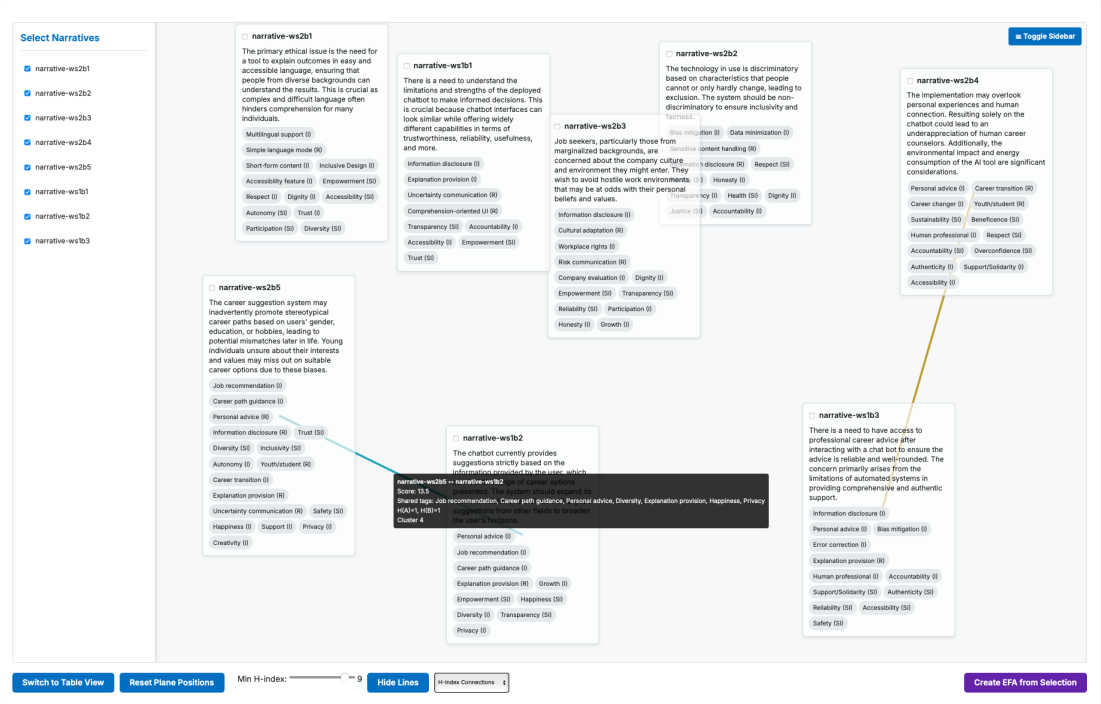


Figure 26: H-index-based clustering algorithm with h-index of 9.