RWTH Aachen University – Faculty of Arts and Humanities
Chair Individual and Technology

I Learned I Ought, but Will I?

Assessing Future Developers' Perception of Ethical AI Principles in the Context of Teaching for AI Ethics Literacy

Master's Thesis

by

Ben Schultz

424244

as part of the Master's Program *Computational Social Systems*

in July 2024

1st Examiner: Prof. Dr. Astrid Rosenthal-von der Pütten
2nd Examiner: Dr. Heqiu Song

Contents

List of Figures

List of Tables

Abstract

As Artificial Intelligence systems increasingly impact society, their ethical development is viewed crucial (Hagendorff, 2020). Establishing ethical AI principles and educating future AI developers in ethical conduct are handled as a potential approaches to ensuring the ethical development of AI systems (Prem, 2023; Fiesler et al., 2020). This quasi-experimental, mixed-methods study examined differences between AI students with and without AI ethics education in regarding attitudes toward AI, attitudes toward ethical AI principles, and ethical behavior in a real-world scenario referring to Theory of Planned Behavior (Ajzen, 1991) and Attitude-Behavior Gaps (Mäses et al., 2019). Ninety-eight AI students with or without AI ethics education participated in a two-part study. They responded to questionnaires, and a subsample of 48 students participated in a mini focus group discussion. Data analysis included differential comparisons and qualitative content analysis of mini focus group transcripts. Results indicated that while both groups held positive attitudes toward AI, students with ethics education approached AI development differently, demonstrating greater knowledge and interest in AI ethics. Attitudinally, these students also viewed AI ethical principles as more important. Behaviorally, they considered a wider range of principles, while students without ethics education showed greater concern for *Reliability, robustness, and security*. Exploratively, complex attitude-behavior gaps were observed, wherein students with ethics education showed significantly fewer gaps related to the principles of *lawfulness and compliance* and *fairness*. The findings suggest that AI ethics education may be an important factor in promoting the practical incorporation of ethical principles in AI development, contributing to the growing knowledge on effective AI ethics education strategies.

Acknowledgements

This thesis has been a lot of hard work that included many eye-opening moments. On the way, several humans have been of great support.

This includes the process of the project itself as well as the masters program.

I would like to thank everyone who supported me along the way. Especially for this thesis, I'm grateful for the great feedback and ideas from fellow students and colleagues. I am grateful for the support of my friends and family, who always had a moment to listen, a place to work, or a fun activity to take my mind off things.

I thank Prof. Dr. Sandra Geisler and Reinhard Linde for providing the laboratory space, Maria Axente for insights into her work in the project for Rao et al. (2021), and Prof. Saskia Nagel and Dr. Niël Conradie for ethical advice on the selected measurements. I thank all the lecturers who were willing to give a minute of their classes or to forward the recruitment material to their students.

Finally, I thank the iTec Chair for its central role in my studies and for providing the resources to conduct this study. I thank Prof. Astrid Rosenthal-von der Pütten and Dr. Heqiu Song for their supervision.

I also thank everyone who showed interest in the project along the way. You have kept my spirits high!

Chapter 1

Introduction

Artificial Intelligence has become an integral part of modern society, influencing various parts of life, from personal interactions to critical decision-making processes in fields such as healthcare, finance, or criminal justice (Maslej et al., 2023). As AI systems continue to evolve and permeate our daily experiences, the ethical implications of their development and deployment have come under increasing scrutiny. Recent scandals and controversies surrounding AI applications have highlighted the potential for these technologies to perpetuate biases, violate privacy, and make opaque decisions with far-reaching consequences (Hinds et al., 2020; Angwin et al., 2022; Köchling and Wehner, 2020; Loyola-Gonzalez, 2019). AI ethics has emerged as a critical study area in response to these challenges. It seeks to address the complex moral and societal issues arising from the development and use of AI systems. However, despite the proliferation of ethical AI guidelines and principles (Hagendorff, 2020; Prem, 2023), there remains a significant gap between the theoretical understanding of AI ethics and its practical implementation by AI developers and practitioners (McNamara et al., 2018).

This study focuses on a crucial aspect of bridging this gap: the impact of AI ethics education on what results out of AI developers' perception on ethical AI; Their ethical attitudes and behaviors. By examining the differences between AI students who have received formal ethics education and those who have not, this thesis aims to shed light on the effectiveness of current educational approaches for ethical AI development practices.

The research is guided by three main motivations:

- The importance of understanding how AI students perceive AI and how they prioritize various ethical AI principles (EAIPs) for their development processes (Prem, 2023).
- The potential of an Attitude-Behavior Gap (ABG) in ethical decision-making,

where individuals' stated ethical attitudes may not align with their actions in practical scenarios (Blake, 1999; Ajzen, 1991).

- The need to evaluate the practical influence of AI ethics education to ensure that future AI developers are equipped with the knowledge and skills to navigate complex ethical situations (Fiesler et al., 2020; Stavrakakis et al., 2021; Brown et al., 2024).

To address these issues, this study examines multiple hypotheses, including whether AI students who have had AI ethics education (AIEE) perceive AI generally more critically, whether they rate EAIPs as 'more important', prioritize these principles differently, and whether they demonstrate more significant consideration of ethical issues in practical settings compared to AI students who did not have such education (nAIEE). Differences in interest and knowledge in AI and AI ethics are also examined for

The author of this thesis conducted a mixed-model quasi-experimental study comparing AIEE students with nAIEE students. The research design incorporates quantitative and qualitative methods, including questionnaires and mini focus groups, to provide a wide-reaching analysis of the impact of ethics education on students' perception of AI and EAIPs and their applied ethical decision-making processes. Behavior has been deductively mapped to a framework of EAIPs (Rao et al., 2021). The framework includes typical ethical topics (Human Agency, Fairness) and technical or legal aspects, allowing for an analysis closer to development practices (Hagendorff, 2020). Due to recent problems with AI in the tax system (Fidelangeli et al., 2021), a case study on AI in tax fraud detection has been used as a context.

By exploring these areas, this thesis aims to contribute to the growing body of knowledge on AI ethics education (Brown et al., 2024) and its efficiency in shaping future AI developers' ethical perspectives and practices. Ultimately, this study seeks to bridge the gap between ethical theory and practice in AI development, educating a generation of AI professionals who are not only technically proficient but also ethically courageous (Hess and Fore, 2018) in their approach to creating, deploying and maintaining AI systems (Domínguez Figaredo and Stoyanovich, 2023; Holstein et al., 2019).

This thesis study report was developed in alignment with the Publication Manual of the American Psychological Association (APA, 2020c) and their Journal Article Reporting Standards (APA, 2020b,a). Regarding references, due to the recency of the topic and the popularity in computer science (Sutton and Gong, 2017), this study also refers to sources published as arXiv (Cornell University, 2024) preprints, if sufficiently often referenced by other scholars. This allows for a broader base of knowledge on the emerging topic, though introducing issues regarding missing peer reviewing processes for these papers.

Chapter 2

Background

2.1 Literature Review

The previous introduction outlined the context, motivation, and approach of this study. Following this, the literature review shall give an overview of major previous publications that are important when investigating future developers' perceptions of ethical AI principles. First, this literature review will introduce you to the grounding context of this study by giving contextual and wide definitions of both *AI* and *perception of AI* fitting the context of this study. Afterward, this is grounded with empirical results from AI students, and due to lack of research on this recent topic, comparable groups (More on these aspects in subsection 2.1.1). These previous studies help to understand how future developers generally view the application of AI. Following this, subsection 2.1.2 allows for an overview of guidelines and principles for ethical AI development.

Within subsection 2.1.3, you receive an overview on *Attitudes towards Algorithms*, *Ethical Development Behavior*, and corresponding empirical studies as well as their intersection; the *Attitude-Behavior-Gap*. Subsection 2.1.6 answers the questions of how *Education* for students in the field of *AI Literacy* generally looks like and how *AI Ethics Literacy* is included. Additionally, there will be a short investigation of what Assessment of AI Ethics Literacy looks like. Following, you will read about the potential effects of *AI Ethics Education* on attitudes, behavior, and the attitude-behavior gap. The literature review ends with a short subsection on how and where AIEPs could be applied and which additional effects must be considered (subsection 2.1.5).

Derived from the background, the foundational hypotheses are established, and research questions of this quasi-empirical study to compare future developers with and without experience from AI ethics education regarding their attitudes, behavior, and the attitude-behavior gap in AI (section 2.2).

Overall, this chapter builds a foundation for the selection of research methods that have been applied in the study (chapter 3) and especially as a background to discuss the study's results (chapter 4) within the discussion section (chapter 5).

2.1.1 Future Developers' Perception of AI

Aside from AI students' and developers' perceptions of ethical AI principles in AI development, it is important to understand their perception of the technology itself. Thus this section will first give foundational definitions of AI, perception as well as its interdependence with attitudes and behaviors. Further, the view of AI students on AI is investigated based on multiple perspectives.

For this, shortly, a guiding understanding of AI is needed: "*Artificial Intelligence*" is a broad term that includes methods and procedures within a system that imitates human intelligence to perform real-world functions (Kok et al., 2009). This integration allows systems, it is integrated into, to act based on given data to solve particular problems (Saranya and Subhashini, 2023). Capabilities include perception, having knowledge, learning, reasoning, problem-solving, and decision-making (Kok et al., 2009). Combining multiple skills allows the system to refine itself based on the given data and new data streams (Saranya and Subhashini, 2023). AI is thus a wider selection of methods than a single technology. These methods first include rule-based systems, which rely on predefined rules and logic to make decisions or solve problems (Grosan et al., 2011). More sophisticated systems of AI include machine learning, which involves training algorithms on larger datasets to identify patterns to make predictions or decisions without being explicitly programmed (Michalski et al., 2013). This can include techniques like supervised learning, unsupervised learning, and reinforcement learning, as well as subsets of deep learning that use artificial neural networks inspired by the human brain to learn from data hierarchically (Michalski et al., 2013; Ongsulee, 2017).

While many professionals identify themselves rather with these subdomains than with the overall term of AI (Wang, 2019), it is important to consider their different backgrounds. At university, multiple different study areas offer education for their students to become involved in the broad area of AI development potentially. The university of this thesis alone lists 36 AI lectures that students from different programs can enroll in (Center for Artificial Intelligence, 2022). While this list is non-exhaustive and leaves out learning opportunities from other fields (Individual and Technology, 2023; Applied Ethics, 2023; Chair of Data and Business Analytics, 2023; Individualized Production, 2022; Institut für Kraftfahrzeuge, 2024), it already indicates the vast different possibilities students can come in contact with topics of AI.

Based on Pickens' work (2005), when investigating AI students' perceptions of ethical AI development, you would inevitably uncover their attitudes and resulting behaviors as well. This is because perception, attitudes, and behavior are closely interrelated psychological constructs. As students interpret and organize their impressions of ethical AI (perception), they will form mental states of readiness towards these concepts (attitudes) (Pickens, 2005). Attitudes' transformation, though, can be a longer process (Pickens, 2005). These attitudes will in turn influence how students further perceive AI-related information and situations. Ultimately, these perceptions and attitudes will manifest in observable behaviors, such as how students engage with AI technologies, discuss ethical concerns, or make decisions regarding AI use (Pickens, 2005). The process is cyclical, as also behaviors can modify existing attitudes and perceptions (Pickens, 2005). For instance, positive experiences with AI might lead to more favorable perceptions and attitudes (Vakkuri et al., 2019b), while new information about ethical implications could result in cautious behaviors and more critical perceptions (Bullock et al., 2021). Thus, any investigation into students' perceptions of ethical AI would naturally reveal effects regarding attitudes and behaviors.

While this study is mostly interested in AI students' views on ethical AI principles, this is closely related to the technology these ethical principles shall be applied upon, AI itself. Many studies have already investigated the attitudes of diverse stakeholders towards AI, especially for *Algorithm Appreciation* or *Automation Bias*, the tendency to excessively and unreflectively show high positive evaluations on automation and algorithms (Lyell and Coiera, 2017; Logg et al., 2019), as well as *Algorithm Aversion*, the tendency to have a reluctance or resistance to relying on computational systems (Burton et al., 2020; Jussupow et al., 2020; Mahmud et al., 2022). Within this, different strands have emerged (Mahmud et al., 2022). While the terms have mainly been used for the context of laypeople and users of AI, by their wide definition of AI, their main message, how appreciative or averse one is to automation, seems also relevant to the wide range of students learning about AI. This thesis will focus on multiple attitudes that are referable to these appreciative or aversive perspectives to give a multifaceted picture of positive and negative attitudes towards AI, adapted from Kieslich et al. (2022) and the Center for Advanced Internet Studies (2024): *Acceptance of AI* for general insights in how people and developers especially approve AI; *Risk and Opportunity Awareness of AI*, for the investigation how prevalent positive and negative outcomes are to respondents; *Usage Intention of AI* helps to understand both a users' perspective of how they want to interact with AI and a developers' perspective whether they would choose methods of AI for a certain professional development problem. The investigation of *Trust in AI* further allows for an understanding of how much developers trust their own technology. Since not much research has been done on the attitudes of (future) developers, this chapter also includes users' perspectives. This helps in the subsequent classification of the results of this study about future practitioners in the wider picture of public attitudes.

Acceptance of AI. Kieslich et al. (2022) found that the public generally mostly accepts AI, with some variations in their responses. Similarly, Bernnat et al. (2023) reported public acceptance rates ranging from 50% to 77% for applied AI, with higher ratings in criminal justice, education, and health domains but lower acceptance for AI applications in political decision processes. This indicates diverse acceptance rates towards AI based on their application. According to the study by the Center for Advanced Internet Studies (2024), there, descriptively, are differences in acceptance based on self-reported AI knowledge: The more participants knew about AI, the more they accepted AI over all domains (Horowitz et al., 2023). This goes along with the confirmation bias among developers as identified by Mohanani et al. (2018): The more people were familiar with algorithms, the more those systems are accepted. Similarly, the more comfortable people are with mathematics, an inherent part of AI development, the more they accept it (Thurman et al., 2019; Logg et al., 2019).

Risk Awareness. The public perceives high risks associated with AI, such as the spread of false information, threats to democracy, digital divide, and job loss (Bernnat et al., 2023). Most believe that the risks of AI cannot be ruled out (Bernnat et al., 2023). Kieslich et al. (2022) found that the public exhibits mid-level risk awareness regarding AI. Additionally, descriptively, participants with more knowledge on AI seem to see fewer risks of AI in comparison to the benefits of AI than participants with less knowledge (Center for Advanced Internet Studies, 2024). If developers are asked about the risks of AI, they tend to focus on physical safety while ignoring emotional and non-human harm (Vakkuri et al., 2019b). They perceive clients and end-users as unaware of the ethical complexities of AI systems. Thus, practitioners think that users may not care about harm as long as their business or direct interaction is not affected (Pant et al., 2024a). Additionally, developers tend to evaluate ethical risks in isolation instead of combined with the benefits of a system (Sanderson et al., 2022).

Opportunity Awareness. According to Bernnat et al. (2023), the public only sees lower levels of beneficial use of AI in helping to solve global challenges or improve the quality of life. The samples of Kieslich et al. (2022) and the Center for Advanced Internet Studies (2024) showed higher mid-level ratings on how large they perceive the opportunities of AI to be. Within this, descriptively, people with greater AI knowledge see greater opportunities in AI compared to the ones with less AI knowledge (Center for Advanced Internet Studies, 2024). Also, Dzindolet et al. (2002) found that expertise can help to identify possible benefits of algorithmic systems.

Usage Intentions. Generally, in a population survey, respondents showed medium levels of intention to use AI (Center for Advanced Internet Studies, 2024). Descriptively, there might have been differences between participants with higher and lower self-reported knowledge on AI. The higher the knowledge, the less people tended to stay distant from AI and the higher the levels appreciating it (Center for Advanced Internet Studies, 2024). Additionally, Önkal et al. (2019) and Araujo et al. (2020) found that training on statistical methods and algorithms increases the likelihood of using these systems.

Trust. Trust in AI is a multifaceted concept with multiple layers, such as human-like trust, as well as distrust and mistrust, and machine-related reliability (Jian et al., 2000; Marsh and Dibben, 2005). The latter is often confused with trust itself within the discourse (Lee and See, 2004). Furthermore, trustworthiness arises when discussing trust in AI (Toreini et al., 2020). While trustworthiness accounts for features in AI's design, development, application, and context, *trust* itself refers to the individual perception of trustworthiness (Toreini et al., 2020). Thus, this thesis will focus on the latter. Generally, the public seems to have a mid-level of trust in AI (Kieslich et al., 2022). Bernnat et al. (2023) found different levels of trust in AI for different domains in which AI is applied. Another public survey by the Center for Advanced Internet Studies (2024) demonstrated descriptive differences in AI's perceived performance depending on the knowledge level about the technology. The more respondents knew about AI, the more likely they were to trust AI to handle various tasks. Meanwhile, developers identify themselves as responsible for making users trust their systems, acknowledging that creating high levels of trust in AI is an important task in their work (Sanderson et al., 2022). They especially see the approach of explainable AI (more on this in subsection 2.1.2) as a method to enabling trust in decision-support systems (Sanderson et al., 2022; Lu et al., 2022). However, practitioners often mix up trust and other ethical principles (Lu et al., 2022). To the best of the authors' knowledge, there are no known studies on how developers and students of AI themselves trust their systems explicitly, even though their trust in AI allows for an understanding of whether they want to apply methods of AI or not.

2.1.2 Ethical AI Guidelines and Principles

Ethical AI Guidelines. Many academic discussions on AI also refer to the ethical problems mentioned in the introduction (chapter 1). Many discussions on how to apply, regulate or teach AI include an ethical perspective (Hagendorff, 2020; Fiesler et al., 2020). These ethical issues are diverse and heavily interconnected. This section of the thesis shall thereby introduce EAIPs as a way to conceptualize these ethical issues collected within broader Ethical AI Guidelines (EAIG). It will be explained that considering a wide range of ethical principles is more sensible for practical application than focusing on just a few (Hagendorff, 2020) but finding a balance for applicability (Whittlestone et al., 2019). Following, this section elaborates on a set of principles in between the extremes by Rao et al. (2021). Afterward, each principle is introduced by problems that occur when breaking with them, as well as practical measures of how these are considered in professional work already.

One popular approach to systematically approaching ethical problems of AI has recently been to set up guidelines and frameworks that guide users, institutions, and developers in their work with and on AI. Many frameworks have evolved over the period of the last century and can generally be viewed as non-legislative policy instruments (Jobin et al., 2019). A great section of these guidelines focuses on the ethical principles of *fairness, accountability, and transparency* (FAT) (Hagendorff, 2020). Thus, technical measures to improve transparency, explainability, and interpretability are considered necessary for experts and laypeople to judge automated systems in their design, data usage, and decision-making (Arrieta et al., 2020). This idea certainly established itself within the *Explainable AI* (XAI) research strand (Ali et al., 2023). With sufficient transparency in the systems, it should be easier to identify technical issues within the system architecture and the foundational data that led to the problem in the first place. Thus, well-trained experts should be able to identify issues and impose solutions (Busuioc, 2021; Diakopoulos, 2016). Contrary, while easily applicable, these guidelines from FAT and XAI are also criticized for being rather 'rationally', 'calculating' and 'logic-oriented' and would only solve those issues superficially when following a checklist approach to ethics (Hagendorff, 2020; Gilligan, 1993; Domingo, 2024; Wong et al., 2020)

Other guidelines go beyond this and include more aspects, which in turn are not as easily technically implementable (Hagendorff, 2020), including aspects of diversity in the field of AI, inclusion, support for whistleblowers, hidden labor costs, and environmental issues (Crawford et al., 2016; Campolo et al., 2017; Whittaker et al., 2018) or principles of *solidarity*, *democratic participation* or *prudence* (Dilhac et al., 2018). These guidelines aim to include the larger network of societal and ecological effects of AI (Hagendorff, 2020). By focusing on the societal goals of using AI, these guidelines can be understood as an abstract frame of concepts in which other, more technical application-oriented guidelines apply principles (Prem, 2023). Furthermore, Prem (2023) argues that many guidelines lack application context and do not consider aspects of practical application within a business or sociotechnical system, such as how and in which constellations AI is developed. Still, how to bring together these rather abstract *concepts of ethical AI* and more applicable *principles of ethical AI* is not always clear (Prem, 2023) and is especially complicated for practitioners (Vakkuri et al., 2019b).

Thus, another strand of guidelines termed *practically relevant EAIGs* emerges (Whittlestone et al., 2019). These try to bridge the gap between abstract constructs and technical checklists. Hagendorff (2020) found multiple guidelines to include more than eight principles. These include alternative principles s.a., *the common good or beneficence*, *human oversight and autonomy*, the *risk of dual use* as well as general *safety* or framed as *non-maleficence* as well as further differentiations of applicability (Hagendorff, 2020; Prem, 2023). This approach allows for an investigation of rather abstract ethical AI, like in Dilhac et al. (2018) in applied contexts, to make complex guidelines applicable (Prem, 2023; Zhou et al., 2020; Ayling and Chapman, 2022; Whittlestone et al., 2019). Even though investigating for adherence to predefined ethical AI development principles in a checklist manner is highly disputed in the ethical and philosophical community (Kiran et al., 2015), it is still a popular approach when bringing guidelines into practice, due to its applicability (Mökander et al., 2022; Zhou et al., 2020; Ayling and Chapman, 2022).

Ethical AI Principles. In this thesis, the *Ethical AI principles* from the responsible AI guideline by Rao et al. (2021) are used (Table 2.1), as they belong to the list of practically relevant EAIGs. The authors took technical, social, and business goals into consideration and derived nine principles out of 100 previous frameworks and guidelines. The guideline includes two general sections: *epistemic principles* as well as *general ethical principles*. Epistemic principles deal with the knowledge base and reasoning processes needed to determine ethical behavior for AI systems

Table 2.1: Nine Ethical AI Principles by Rao et al. (2021) sorted into Epistemic Principles and General Ethical AI Principles.

	Ethical AI Principles
1) Epistemic Principles	1a) Interpretability (Explainability, transparency, provability)
	1b) Reliability, robustness, security
2) General Ethical Principles	2a) Accountability
	2b) Data privacy
	2c) Lawfulness and compliance
	2d) Beneficial AI
	2e) Safety
	2f) Human Agency
	2g) Fairness

(Rudy-Hiller, 2018). This follows the argument that other ethical principles, like fairness, could not be applied without a reliable or explainable AI system. Rao et al. (2021) consider the first two principles of their list to belong to this group: The principles of 1) *Interpretability (Explainability, transparency, provability)* and 2) *Reliability, robustness, security*.

Interpretability. To tackle the multifaceted nature of a clear definition of explainability issues of AI (Felzmann et al., 2020). Rao et al. (2021) selected *interpretability* (providing information that is understandable for humans) as an overarching term (Arrieta et al., 2020). Under this term, they consider transparency (a model that is considered to be understandable by itself), explainability (the interface of the decision-making system to a human about its goals, processes, and outcomes), and provability (the design of the AI system allows experts to verify the mathematical correctness of the system) (Felzmann et al., 2020; Arrieta et al., 2020; Leitgeb, 2009). Following this, developers, users and further stakeholders should be enabled to transparently understand the work and the outcome of the automated systems for evaluation. Practically, guidelines recommend the use of special algorithms and software libraries to enhance, e.g., explainability (Prem, 2023).

Reliability The second principle of *Reliability, robustness, security* aims at a smooth operation of the system (Rao et al., 2021). Thus, reliability of an AI system is fundamental so that the automation can be relied upon (Ryan, 2020). E.g., from a practical development perspective, an inaccurate AI system would not be reliable and thus be doubted in rollout. The same holds for *robustness* and *security*: Thus, a system ought to be designed to give stable predictions even when dealing with variations in input data (Freiesleben and Grote, 2023). Adding to this, they should be secure enough to withstand attacks or accidents affecting the system (Barreno et al., 2010; Toreini et al., 2020), like hacking or energy outage. All three parts of the principle aim to improve the predictability of AI systems under usual as well as unexpected circumstances. To consider them, guidelines advise developers to learn more about the topics and (Kolter and Madry, 2018; Isaac and Reno, 2023) as well as to include compliance processes within development (Isaac and Reno, 2023; Rao et al., 2021; Prem, 2023).

Both epistemic principles are considered to potentially have positive impacts on achieving trustworthy AI as proposed by the XAI research strand (Freiesleben and Grote, 2023; Toreini et al., 2020; Ali et al., 2023). Still, other researchers such as Bell et al. (2022) see issues, since the focus on explanations of AI systems may not allow for proper understandability, or that promoting accuracy as well as the design of interpretable and explainable AI might be used to manipulate users into trusting non-trustworthy systems (Gilpin et al., 2022). On a different note, the complexity of accuracy and explainability of AI becomes clear in their currently discussed direct trade-off in application (Bell et al., 2022; Felzmann et al., 2020). These debates show how hard it can be, to implement epistemic principles properly. As will be shown below, the same holds for the second set of Rao et al. (2021)'s principles:

Accountability The section on *general ethical principles* concludes the other seven principles. *Accountability* implies that all stakeholders involved are responsible for the moral implications of the use and misuse of AI (Rao et al., 2021). This also includes the often-cited concept of dual-use; the potential of artificial intelligence technologies to be used for both beneficial and harmful purposes (Brundage et al., 2018). Both beneficial and harmful use aspects are considered in more detail in principles 2d and 2e (Table 2.1). In other publications, accountability and responsibility are considered separately (Hagendorff, 2020) due to the complex nature of responsibility in its mode and direction (Van de Poel and Sand, 2021). Others only consider a forward looking version of responsibility to be able to attribute (moral) responsibility towards a party being in the process of designing, developing as well as maintaining the AI system (Bernnat et al., 2023; Santoni de Sio and Van den Hoven, 2018; Van de Poel and Sand, 2021).

Data Privacy *Data Privacy* includes the closure of information on an individual and the prevention of further distribution of already published information on an individual (Yu, 2016). Thus, it can be viewed as a variation of informational privacy (Rao et al., 2021). Popular approaches are the minimization of sharing and collecting information, anonymization of information, and more sophisticated technical measures of encryption for secure information processing, transmission, and storage (Jain et al., 2016). These shall allow users to engage in behavior without having to fear being vulnerable to blackmail (Brierley et al., 2021) or automated manipulation (Gilpin et al., 2022). Privacy is not only connected to the principle of securing human agency in interaction with AI but also to the principle of security so that privacy rights are defended against attacks on the systems and their databases. Additionally, privacy stands in conflict with the goal of reliability. The more data is collected as well as used, the more accurately an AI system can work (Machanavajjhala et al., 2011; Binns and Gallo, 2019). Even though privacy is included as a fundamental human right (Diggelmann and Cleis, 2014; EU, 2012), regions worldwide place different legislative importance on the topic. Following this, separating the principles of data privacy from the next principle in the list, lawfulness and compliance, might seem irritating initially. But when developers try to balance accuracy and privacy,

privacy must be considered separately in the analysis of ethical behavior, especially when the law requires only minimal standards.

Lawfulness and compliance Rao et al. (2021) frame the principle of *Lawfulness and compliance* as all stakeholders' adherence to the law and relevant regulatory regimes. This principle accounts for the fact that AI is used in regulated areas of high societal significance (Prem, 2023; Hagendorff, 2020). Thus, regulations of data and AI usage are being discussed and published currently all over the world (Corrêa et al., 2023), adding to the legal literature. Aside these legal foundations, often, ethical considerations are considered in self-regulatory structures (Hagendorff, 2020). Thus, reminding developers and other decision-makers in AI development to follow both the law and additional "[non-binding] corporate self-governance" (Whittaker et al., 2018, p. 30) is an important aspect included in ethical AI guidelines. Issues of neglecting this principle can be found in the example of multiple big tech companies laying off their AI ethics teams, leading to weakened internal compliance mechanisms (Duffy, 2023; Field and Vanian, 2023). In the case of being in conflict with the law over AI, e.g. responsible companies have been sued for copyright infringements, non-consensual data collection as well as discrimination (Brittain, 2024; Grynbaum and Mac, 2023; Kang, 2022; Lyles, 2024; Wiessner, 2024). Methods of minimizing these issues include improvement of internal governance (Kiran et al., 2015; Zhou et al., 2020; Corrêa et al., 2023) and auditing measures, as well as institutions for complaints and further interdisciplinary education for responsible development of practitioners (Hagendorff, 2020).

Beneficial AI *Beneficial AI* refers to the promotion and reflection of the common good through the development of AI. An AI system might follow all other ethical standards and could still not be beneficial for humanity since their goals are not aligned with the common good. While Rao et al. (2021) include sustainability, cooperation and openness in *beneficial AI*, other guidelines focus on aspects from the UN Sustainable Development Goals (International Telecommunication Union, 2017) ranging from climate change and healthcare to financial inclusion and disaster risk reduction. Still, while many guidelines include concepts of beneficial or good AI (Hagendorff, 2020; Prem, 2023), the discussion on what goodness ought to mean in the context of AI is not settled yet (Aula and Bowles, 2023; D'Acquisto, 2020).

Safety The same problem of unclear definition holds for the next principle on *Safety* (Rao et al., 2021), often also termed *"Non-Maleficence"* (Hagendorff, 2020). While the concept is well established in healthcare ethics (Gillon, 1985), the number of guidelines on ethical AI, show different definitions of non-maleficence (Prem, 2023). Here, the definition by Rao et al. (2021) is used, "not compromising physical safety and mental integrity" (Rao et al., 2021, p. 8) of human stakeholders of AI throughout the entire lifetime of the system. One prominent example of designing maleficent AI is the increasing number of deepfakes for individual and political use cases (Westerlund, 2019).

Both beneficent AI and non-maleficent AI stand in stark contrast to business practices (Taddeo and Floridi, 2018; AI HLEG, 2019) that often focus on an increasing

AI's capabilities with efficiency and technological superiority in mind (Asaro, 2019). Thus, considering both these principles is even more important in analyzing ethical conduct in AI development. Applied, methods of improvement towards beneficence and non-maleficence, especially include licensing models and certificates validating responsible conduct (Prem, 2023; Kiran et al., 2015; Cihon et al., 2021).

Human Agency While beneficence and non-maleficence are part of many EAIG, the principle of *Human Agency* is not (Zhou et al., 2020; Hagendorff, 2020; Jobin et al., 2019). Rao et al. (2021) see it as a way of possible human intervention within automated AI processes. Contrasting this, Hagendorff (2020) splits this principle into two. He identified *human agency* to be the manner of non-manipulation of AI towards stakeholders, and additionally the possibility of *human intervention* to steer the decisions of an AI system. Both concepts may occur within the same scenario. Imagine house owners using an AI system to automate their heating systems. The possibility of human intervention in the decision-making process of heating allows them to retain human autonomy in their heating decisions. Still, ethically, one could argue that a user or stakeholder might be nudged towards a certain decision by the algorithm's recommendation. Therefore, the involvement of an automated recommender system reduces their autonomy in decision-making, regardless of their level of involvement. Solving this ethical discussion goes beyond the scope of this thesis, but the conflict shall be mentioned at this point for a more profound understanding of the principle as proposed by Rao et al. (2021). To allow AI to follow individual values instead of manipulating users into certain actions, stakeholders can be involved in the development process to elucidate their positions (Prem, 2023). For human involvement, concepts such as human-on-the-loop, human-in-the-loop or human-in-command are currently evaluated regarding their benefits and drawbacks (Prem, 2023; Yurrita et al., 2023; Mosqueira-Rey et al., 2023; Wu et al., 2022).

Fairness As already mentioned, *Fairness*, as the last principle of the list, is quite popular among EAIGs (Hagendorff, 2020; Prem, 2023; Zhou et al., 2020; Jobin et al., 2019). (Rao et al., 2021, p. 8) stay rather vague within their definition that "[t]he development of AI should result in individuals within similar groups being treated in a fair manner, without favoritism or discrimination, and without causing or resulting in harm. AI should also maintain respect for the individuals behind the data and refrain from using datasets that contain discriminatory biases." Also, other frameworks similarly refer to issues of bias, discrimination, equality, when including about fairness (Prem, 2023). Sometimes, aspects of fair AI are also included in the principles of dignity or fair treatment of AI workers, as found in other guidelines (Hagendorff, 2020; Jobin et al., 2019). In cases in which fairness is not considered, AI systems could become discriminatory so that companies and individual decision-makers might even face legal consequences (Wiessner, 2024) or could face shutting down the alleged systems as a whole (Jeffrey Dastin, 2018). Thus, following fairness in AI is both morally advised (John-Mathews et al., 2022) and beneficial from a business perspective. Still, fairness stands in a technical conflict to the accuracy of predictions, thereby often being disregarded as an ethical principle compared to reliable results of responsible AI (Jui and Rivas, 2024; Kleanthous et al., 2022).

Accounting for fairness can mean to include diverse stakeholders and teams in their development process, set up technical measures of protected attributes, and try to balance weights against biased algorithms or datasets (Hagendorff, 2020; Prem, 2023; Zhou et al., 2020; Jui and Rivas, 2024).

Fully considering all principles in a project satisfactorily can sometimes be complicated or even impossible (D'Acquisto, 2020). A focus on accuracy can practically be in conflict with considering privacy and fairness (Binns and Gallo, 2019; Jui and Rivas, 2024). On the other side, higher reliability can allow for greater safety by minimizing unintended side effects. This, in turn, can also influence the lawfulness of the system. Additionally, cases of low security of a system would also be a threat to privacy when considering data breaches. These incidents could also affect the individual safety of users and stakeholders. While high transparency might be relevant to working against these issues, it could also threaten privacy. Due to this complexity and the focus on only a small set of principles in many guidelines, certain principles are more often considered than others when it comes to implementation. Additionally, principles such as data privacy, beneficence, non-maleficence, or human autonomy are often harmed by current AI applications (Hagendorff, 2020). Thus, it is important to ask how these principles can be considered in a balanced way in actual development practices (D'Acquisto, 2020).

2.1.3 Future Developers' Attitudes on Ethical AI

AI developers play an important role in AI design, development, and deployment processes. They are involved in multiple AI development steps and can influence the final system in multiple ways. Since both computer science and AI are still relatively young technologies, studies have far more frequently examined students from the more general STEM sciences. Nevertheless, to attempt to draw lessons about the knowledge, attitudes, and behavior of AI students, the studies on STEM students are supplemented in this chapter with studies on computer science and AI practitioners. As students from both computer science specifically and STEM generally (Kandlhofer et al., 2016; Meesters et al., 2022), learn the skills to potentially becoming be future AI developers (Schleiss et al., 2022), it can be assumed that they are also relevant to the analysis in this section. This section shall thus investigate professionals, and especially students, who might take positions in development or decision-making in their future careers. It gives an overview of the attitudes of these practitioners towards ethical conduct in general and ethical (AI) development practices. Building on this, this section will give insights into the tradeoffs between principles.

A violation of one or multiple ethical principles has been shown to lead to public outrage, as in the case of an automated grading system in the UK (Kelly, 2021). To understand these stark public reactions, looking at attitudes to ethical conduct is important. These attitudes are not merely personal preferences or opinions, but rather individually represent guiding principles about what individuals and society "ought to do" (Ellemers et al., 2019). Various factors shape these attitudes: The social environment, such as societal values or (sub-)cultural norms, plays a significant role in shaping perceptions of what is considered right or wrong (Pulfrey and Butera, 2016; Stoeber and Yang, 2016). The strength of personal moral beliefs, attitudes, or convictions can make individuals resilient against social pressures to act unethically (Brezina and Piquero, 2007). Strong moral convictions can be a guiding force, even in the face of opposing influences, like in a group setting with strong social norms (Hornsey et al., 2003). The individual's self-image as moral or immoral can strongly influence behavior (Zhong and Liljenquist, 2006). People who perceive themselves as "good" or "ethical" may be more likely to follow through on good intentions and act in accordance with their moral beliefs (Ellemers et al., 2019). When an individual witnesses others acting against her ethical or moral principles, it can lead to distress (Skitka and Mullen, 2002). Similarly, personal engagement in misconduct can create internal conflict and discomfort, as it might contradict with the own moral attitudes and beliefs (Graham, 2007). Attitudes toward ethics and moral behavior are not static; the interplay between thoughts and experiences shapes them. Generally, personal experiences, exposure to different perspectives, and critical reflection can all contribute to the evolution of these attitudes over time (Klein and O'Brien, 2016).

Kieslich et al. (2022) investigated in their study how important the public evaluate EAIPs to be. The participants underwent a discrete choice experiment and rated systems with different ethical features by satisfaction. Accountability was seen as the most important ethical principle. Fairness, security, privacy and accuracy were rated in the middle, while explainability was rather unimportant. Additionally, they checked for attitudes on machine autonomy, which received the lowest overall ranking. Now, various empirical studies that deal with the perspectives of developers and students in the field of AI and related will be listed.

According to studies by Finelli et al. (2012) and Kreth et al. (2022), STEM students generally exhibit low levels of knowledge and capabilities in ethical reasoning. Specifically, computing students tend to have lower social responsibility attitudes compared to students from other domains, including other STEM majors. This deficit in ethical awareness can be attributed, in part, to the emphasis placed on technical problem-solving over social dimensions in technical education (Schiff et al., 2020). Furthermore, professional internships can reinforce this development, as students often prioritize technical skills over ethical considerations (Rulifson and Bielefeldt, 2018, 2019). Additionally, students are often less strict with their own ethical conduct compared to the ethical conduct of others (Almasri and Tahat, 2018). In the study by Schiff et al. (2020), STEM students had difficulties transferring high attitudes toward social responsibility from their private lives to professional social responsibility. Ethical issues on a large societal scale, as considered in many ethical AI guidelines, were more affected by this effect than ethical questions on an individual, small or daily level (Schiff et al., 2020).

Turning to the specific domain of ethics in IT, several studies have highlighted the challenges and attitudes surrounding this topic. Almasri and Tahat (2018) found that some students perceive IT ethics as less important than general ethics. Moreover, dealing with ethical issues in AI is a relatively novel concept for many developers (Vakkuri et al., 2019a). They report only coming in contact with rudimentary ethical risk assessment frameworks, such as privacy by design, and there is often a "do once-and-forget" approach to ethical considerations (Sanderson et al., 2022). In some cases, developers consider AI ethics detached from the current issues in the field (Vakkuri et al., 2019b). A systematic literature review by Pant et al. (2024b) further highlights the varying perceptions of ethics among AI professionals. Some practitioners perceive it to be critically important, while others show the complete opposite opinion. They also found that practitioners view pro-ethical designs as being an improvement in social impact but with higher costs for resources and time. Pant et al. (2024b) conclude that these differing views shown in multiple studies hint at the fact that ethical aspects are also from the side of perception, complex to understand and to consider.

When considering ethical principles and guidelines, Sanderson et al. (2022) found that practitioners often prioritize privacy and security considerations when developing AI systems. In their study, the principle of wellbeing has only been indirectly addressed and not directly named. The developers participating have been unsure about how to ensure legal compliance and seek advice to ensure everything is done correctly from a regulatory standpoint. In the study by Lu et al. (2022), developers expressed a desire for verifiable requirements and ethical assessment tests when developing AI systems. Their focus extends beyond privacy and security to include reliability and XAI. However, human-centered values tend to be the least important to these developers. Many practitioners adopt a checklist approach when considering ethical principles in AI development (Pant et al., 2024b). This involves ensuring that they have correctly implemented and sufficiently addressed various ethical principles (Pant et al., 2024b). While there is a positive perception regarding the importance and benefits of incorporating ethics into AI development, some negative perceptions also exist (Pant et al., 2024b): The high cost associated with applying ethical principles can be a deterrent. Ethics is sometimes viewed as a non-functional requirement in AI development, potentially leading to its low prioritization (Pant et al., 2024b).

Following, the principles relevant to this thesis (Table 2.1) will be individually investigated regarding the attitudes of developers and students.

Perception of Interpretability (Explainability, transparency, probability). Students in the field are aware of the complexity of transparency and want to make explanations understandable to non-experts (Kleanthous et al., 2022). Sanderson et al. (2022) found that transparency is often an interim target for practitioners when developing AI systems to achieve high reliability, after which transparency may no longer be needed. Explainability, similarly, is considered important for increasing trust in the system, although it may only be needed temporarily until people understand how the system works (Sanderson et al., 2022). Seppälä et al. (2021) emphasize that developers give high importance to increasing trust by using measures of XAI to increase transparency on when AI is at work, as well as interpretation to detect biases. Contrary, Vakkuri et al. (2019b) found that developers perceive end-users as not being tech-savvy enough to gain anything from technical system details and thus do not see a reason to provide insights. This hints at the fact, that they are aware of the gap between explainability and interpretability. Vakkuri et al. (2020) reported that more value is put on transparency of software operations internally of the organization compared to transparency for external stakeholders. And even for internal transparency, students in the study of Kleanthous et al. (2022) did not include audits as a possible measure to increase transparency aside from more technical approaches. To the best of the author's knowledge, perceptions of the provability of AI have not been researched thoroughly so far.

Perception of Reliability, robustness, security. Developers see that ensuring AI systems' reliability, robustness, and security is an ongoing process that requires iterative testing and reworking (Seppälä et al., 2021). Model validation is a crucial aspect of this process (Seppälä et al., 2021). Developers are aware of the high odds of unpredictability associated with AI systems (Vakkuri et al., 2019b), although a study by Vakkuri et al. (2020) found that 76% of developers believed their software's operation was predictable (Vakkuri et al., 2020). However, there is a tendency among development teams to prioritize the usefulness and viability of their product over other ethical aspects (Vakkuri et al., 2019a). On the one hand, in a study by Khan et al. (2023), accuracy was rated as less important. On the other hand, multiple studies indicated that it was evaluated as more important compared to other ethical principles (Sanderson et al., 2022; Vakkuri et al., 2019b; Hadar et al., 2018; Arizon-Peretz et al., 2021). Van Stuijvenberg et al. (2024) argue that this is due to usage of accuracy as a popular performance measure of AI systems among developers. Security, as well, is consistently identified as one of the most important principles for practitioners when developing AI systems (Sanderson et al., 2023; Lu et al., 2022). The security discourse seems relatively prevalent to developers (Arizon-Peretz et al., 2021), even though that they have issues to differentiate it from the concept of privacy (Peixoto et al., 2020). This emphasis on security aligns with the need to ensure the reliability of AI systems, as security vulnerabilities could compromise the overall functioning of these systems. Still, developers also often leave their formal education without previous training on security issues and only advance in these fields as soon as they approach issues at work (Balebako et al., 2014).

Perception of Accountability. Accountability is a critical principle in the development of AI systems according to FAT machine learning and other guidelines (Hagendorff, 2020), but there are challenges in how it is perceived and approached by developers. According to Vakkuri et al. (2019b), developers' motivation for taking responsibility is often driven by pragmatic concerns rather than ethical considerations. For instance, the fear of physical harm, financial implications, customer relations, or legislative requirements may be more compelling factors than a sense of ethical duty. Furthermore, Vakkuri et al. (2019a) found that students in AI programs tend to lack knowledge on accountability issues. This knowledge gap goes along with the tendency among developers to outsource accountability to users rather than considering the complexity of responsibility in the systems they create (Vakkuri et al., 2020).

Perceptions of Data Privacy. Data privacy is considered one of the most important ethical principles among developers and software teams, as evidenced by studies such as Lu et al. (2022) and Sanderson et al. (2023). This is in line with demands by the public on privacy in AI development (Bernnat et al., 2023). However, this finding contrasts with the privacy-harming practices of some companies (Gröger, 2021; Maslej et al., 2023). Only a few developers have received formal training on privacy, with some receiving corporate certifications (Balebako et al., 2014). Thus, regardless of their region, developers are often relying on self-teaching and online forums and do not apply the full set of possible technical measures (Prybylo et al., 2024). Developers often using the vocabulary of data security to approach privacy challenges (Peixoto et al., 2020). Additionally, they rely on legal experts to help create privacy policies for their teams (Balebako et al., 2014; Prybylo et al., 2024). If developers consider privacy, several factors are found to influence their motivations in doing so (Peixoto et al., 2020; Hadar et al., 2018): When directly asked in the study by Peixoto et al. (2020), developers answer they are often motivated by ensuring users' needs or complying with privacy laws. Hadar et al. (2018) opposingly found that often, business considerations are prioritized over preserving users' informational privacy when in conflict. Thus, perceptions and behaviors of developers regarding privacy and security may not always align with the expectations and recommendations of policymakers (Arizon-Peretz et al., 2021).

Perception of Lawfulness and compliance. While most developers acknowledge that their industries are regulated (Vakkuri et al., 2020). However, many have not received enough formal education on regulation during their studies Vakkuri et al. (2020). This is problematic regarding the different languages used in the fields (Bell et al., 2023). Particularly regarding privacy regulations, developers rely on legal experts to navigate regulatory requirements or even seek help in online forums or with friends (Sanderson et al., 2023; Balebako et al., 2014; Prybylo et al., 2024). Additionally, practitioners often feel less responsible for cases of harm if their software fulfills regulatory standards (Vakkuri et al., 2020).

Perception of Beneficial AI. The principle of beneficial AI, which aims to ensure that AI systems are designed and developed for the benefit of humanity, is sometimes overlooked or treated as a mere project objective rather than a set of verifiable requirements or outcomes (Lu et al., 2022). In a study by (Khan et al., 2023), aspects of sustainability, freedom, and prosperity, all parts of beneficial AI (Hagendorff, 2020), received relatively low importance ratings.

Perception of Safety. Developers often consider safety primarily in terms of physical harm to humans (Vakkuri et al., 2019b). Further, there is a lack of skills and understanding among developers regarding the potential for AI systems to cause harm beyond physical harm, such as systemic effects or social and emotional impacts (Vakkuri et al., 2019b). Both areas of (future) practitioners' perceptions of the beneficial use of their technology and harm prevention in the sense of safety are less researched than topics like transparency, accountability, privacy, or fairness.

Perception of Human Agency. Ensuring human agency in the context of AI systems is a multifaceted challenge. There is a preference among a few developers for AI recommender systems rather than fully autonomous decision-making systems among developers (Seppälä et al., 2021), as this allows for human agency in the final decision-making process. Approaches like allowing human decision-makers to overwrite AI decisions (Sanderson et al., 2022) or incorporating "kill-switches" (Lu et al., 2022) are used to maintain human agency and control over AI systems. However, some developer also stress the potential biases that can arise from humans in the loop (Holstein et al., 2019), highlighting the need for careful consideration and mitigation of human biases when preserving human agency in AI systems. Certainly, practitioners are more concerned about human bias than e.g. lawmakers (Khan et al., 2023). The question of how much human bias is optimal goes along the lines of studies discussed previously, that some developers think that higher reliability of an AI system can be a reason to put a lower priority on transparency and human say over an autonomous system (Sanderson et al., 2022).

Perception of Fairness. Studies have shown that AI students lack knowledge on AI fairness, with many never receiving formal training on the topic (Kleanthous et al., 2022). Postgraduate students tend to identify problems in data better than undergraduates, and there is a recognition that definitions of algorithmic fairness are subjective and that human intervention is important (Kleanthous et al., 2022). Bringing diversity into development teams and focusing on training data and algorithms are seen as crucial for improving fairness in a system (Holstein et al., 2019). Gender differences in beliefs about fairness have been observed, with males rating accuracy as more important than minimizing racial disparities (Pierson, 2017). Professionals often lack the practical knowledge and skills to develop fair systems, and they view increasing fairness as challenging due to technical and organizational boundaries, even when efforts are made (Holstein et al., 2019). Compared to lawmakers, practitioners evaluate fairness as less important, potentially due to the non-technical nature of the term and a lack of understanding of how to interpret fairness technically (Khan et al., 2023).

Tradeoffs between Principles. A study by Khan et al. (2023) asked practitioners to rate the perceived importance of 21 ethical principles related to AI. When directly asked to rate the importance of ethical principles, transparency, accountability, privacy, human dignity, and non-maleficence emerged as the top priorities. Other studies consider Human dignity closely related to fairness and safety (Hagendorff, 2020). The principles that received the fewest ratings of importance were sustainability, accuracy, freedom, prosperity, and explainability. Notably, sustainability, freedom, and prosperity can be considered part of the broader principle of beneficial AI, as outlined in frameworks like by Rao et al. (2021). Overall, the participants in Khan et al. (2023) generally considered all principles to be important rather than unimportant. Developers often face situations where they must navigate tradeoffs between different ethical principles. For example, Sanderson et al. (2022) highlights potential conflicts between reliability and fairness and between reliability and privacy. In such cases, developers may need to prioritize one principle over another. Khan et al. (2023) also found that many practitioners acknowledge the issue of conflicting principles in practice.

Even though only a few studies have investigated the perceptions of ethical AI frameworks by practitioners, the comparison of multiple principles shows a trend in favor of the principles of reliability, data privacy and security. Still, the latter two are a good example to see that developers often have issues in clear differentiation of the concepts.

2.1.4 Future Developers' Ethical AI Behavior and the Attitude Behavior Gap

While developers and students of AI favor certain EAIPs over others, attitudes are not the only predictor of actual behavior to enact their favorite principles. Following the Theory of Planned Behavior (TPB), besides attitudes, subjective norms and perceived behavioral control are also relevant in predicting behavioral intention (Ajzen, 1991). Thus, not only individual attitudes towards a behavior but also the perceived social pressure around the behavior and the belief about internal as well as external factors that may facilitate or impede the behavior influence how hard an individual is willing to try and how much effort they plan to put forth to perform a behavior (Ajzen, 2014). This intention is the most proximal factor for actual behavior (Ajzen, 1991). Thus, it is often used as a proxy for behavior within behavioral ethics research (Sadeghi et al., 2022; Gino et al., 2009). As seen, attitudes must not go hand in hand with behavior directly (Ajzen, 2014). Especially in research on social and environmental behavior (Portus et al., 2024), this pattern became clear. This effect is termed the Attitude-Behavior Gap (ABG) and is especially used in research on responsible behavior, such as the usage of seat-belts (Mittal, 1988) or environmentalism (Park and Lin, 2020). Since attitude and behavior are not directly in line with each other, this section will elaborate on the ethical behavior of technical and especially AI practitioners and students, followed by a deeper contextualization of the ABG regarding the TPB with first empirical studies from the IT sector.

Ethical Behavior. Ethical behavior refers to actions that align with established ethical guidelines and principles (Ellemers et al., 2019). This stands opposed to moral behavior, which is judged against individual moral standards (Ellemers et al., 2019). According to Ellemers et al. (2019), the strength of personal moral beliefs, attitudes, or convictions can make individuals resilient against social pressures that may encourage ethical behavior. However, in domains where personal moral convictions are less strong, moral norms established by team atmosphere or principled leadership can overrule individual concerns (Ellemers et al., 2019; Osswald et al., 2010). Ellemers et al. (2019) highlights the complex interplay between personal moral convictions, social pressures, and the framing of moral dilemmas in shaping ethical behavior. While strong personal moral beliefs can promote ethical conduct, social norms and contextual factors can also significantly influence individuals' ethical decision-making and actions (Ellemers et al., 2019; Osswald et al., 2010). Reimenschneider et al. (2011) found that depending on the context of computational topics, like internet plagiarism or collaborative programming, different attitudes, and behavioral intentions are prevalent to behave (un-)ethically. Hedayati-Mehdiabadi

(2022) identified multiple factors that positively and negatively influence ethical decision-making among future developers: Relating to real-world scenarios, showing care for users or affected individuals, or applying their experiences of insecurity or confusion as users went in line with ethical behavior. The same effect occurred for recognizing biases in the arguments of other students and feeling responsible due to possessing special technical knowledge and experience compared to others (Hedayati-Mehdiabadi, 2022). On the contrary, students who showed rather unethical behavior in that study, argued with minimalist assumptions about the capabilities of users or generalizing their own (knowledgeable) technology usage to all users. Additionally, practitioners fell for fallacies e.g., that they felt incapable of acting based on their values due to external pressures (Sartre et al., 2022), that they tended to present unethical decisions as morally useful (Bandura, 1999), or that ethical problems are only relevant if part of rules, regulations, or if problems occurred (Hedayati-Mehdiabadi, 2022).

Not much research has been done on developers' ethical behavior regarding specific EAIPs. Thus, the following empirical studies exemplify based on the commonly investigated EAIPs of *Data Privacy* and *Fairness*. Regarding *Data Privacy*, Alhazmi and Arachchilage (2021) found that the minority of developers were familiar with GDPR principles; if they were, they often lacked the requisite knowledge about the implementation techniques. Developers tended to focus more on functional requirements than privacy requirements, partly due to the unavailability of online tools and lack of support from institutions and clients. Prybylo et al. (2024) observed that developers do not use anonymization techniques frequently enough. Instead of having learned about the issues beforehand, developers seek answers to their privacy questions from friends, social media such as developers' forums, in addition to legal/policy experts (Balebako et al., 2014; Prybylo et al., 2024). Regarding *Fairness*, Cowgill et al. (2020) have shown that a reminder on the fact that data can be biased increases awareness that there is a problem regarding fairness within a task. This effect did not occur by technical documentation at the hands of the developers. Furthermore, female professionals were more likely to recognize biased datasets than men. However, this gender imbalance did not translate to the level of bias in developers' development outcome (Cowgill et al., 2020). While having diverse teams has been proposed as a potential solution to promote fairness (Pant et al., 2024b), this approach has not consistently proven effective (Pierson, 2017).

The Attitude-Behavior Gap. As described above, the ABG, refers to the discrepancy between an individual's stated attitudes, or intentions and their actual behavior (Blake, 1999). External factors, such as situational constraints, social pressures, and cognitive biases, can contribute to the divergence between attitudes and behaviors (Simon, 1991; Rubinstein, 1998). The underlying theory by Simon includes that humans are not perfectly rational but seek solutions that are satisfactory rather than optimal. As already explained, psychology also considers this gap as a general phenomenon to be explained by theories, s.a. the TPB (Ajzen, 1991). The TPB can also be transferred to ethical topics (Chang, 1998) and to computational areas such as information technology or electronic piracy (Panas and Ninni, 2011; Reimenschneider et al., 2011). According to the two studies by Panas and Ninni (2011) and

Reimenschneider et al. (2011), attitude did show to predict behavioral intention in ethically complicated scenarios. However, there are also additional effects of possible legal punishment or equity in social exchanges that minimize or increase unethical behavioral intention.

While the concept of the attitude-behavior gap has been recognized in various domains (Mittal, 1988; Park and Lin, 2020), there is emerging evidence of its existence in technology ethics. The study by Sadeghi et al. (2022) have provided initial insights into this gap in the context of technology and AI ethics, especially for security behavior: Participating students from computing and/or business studies should decide whether to release an ethical worm. They had to choose either in favor of security or organizational regulation. Their decisions differed widely, and the authors assume that many students could not evaluate the situation properly before their instruction about the concept of ethical worms. Griffin et al. (2024) conducted one of the first qualitative investigations into the attitude-behavior gap among AI developers. The study identified gaps between how developers perceive themselves and their work experiences. Developers' perceived ethical agency varies: While they do have some authority to intervene for ethical reasons in the systems they work on, they often do not realize the extent of how many ethical decisions they make. Nonetheless, the study revealed a growing ethical wisdom among developers, considered beneficial to nurture by the authors. This wisdom goes in line with what Hagendorff (2020) asks for and calls *AI virtues*. Developers should be better equipped to act in line with ethical demands from ethical AI frameworks. According to Vieira et al. (2023), besides e.g., conflicting goals, a lack of knowledge can be a major barrier to reducing unethical behavior. Interventions focused on increasing awareness of problematic behavior and knowledge on behaving more ethically, as for environmentalist behavior, have increased positive attitudes and behavioral intention toward ethical conduct (Jay et al., 2019; Kwasny et al., 2022).

2.1.5 When to Consider EAIPs and Demographic Effects

Ethical Decision Points in AI Development. Even though ethical AI development underlies the complexity of bridging attitudes into behavior, Prem (2023) found that EAIGs often consider an *AI development process* to recommend when and how to include ethical principles. The approach of considering ethics during the development process of AI, can be sorted into the research strand of Ethics by Design (Brey and Dainow, 2023; Dignum et al., 2018). This idea allows incorporating values relevant to the context of the application of the system already within the development phase to allow for value-oriented automation further down the line (Brey and Dainow, 2023). Closely aligned with results from Prem (2023) analysis of ethical decision points, Rao et al. (2021) have set up a nine-step model of AI development, relevant for the methodology of this study.

1. Business and data understanding

2. Solution design
3. Data extraction
4. Preprocessing
5. Model building
6. Model deployment (Dev)
7. Transition and execution
8. Ongoing monitoring
9. Evaluation and check-in

Within the first two steps, the involved decision-makers should consider the ethical aspects of *beneficence* and *non-maleficence* for the overall system design, as well as *stakeholder participation* and *human oversight* (Prem, 2023). Within the technical steps *3, 4, 5*, developerscould extract the important data points needed for considering *fairness and biases* as well as *accuracy, reliability, robustness, and security* (Prem, 2023). They can opt to develop a model that can include technical measures for guaranteeing *transparency* and *explainability* (Prem, 2023). Further, with steps *6* and *7*, developers have to opportunity to first identify issues of the previously integrated principles. Additionally, they ought to guarantee that a diverse set of stakeholders is included in testing processes and deployment change management. Several approaches to *interpretability* could be included so that all stakeholders are able to interpret the explanations given by the AI system (Prem, 2023). Especially in integrating the final system into previous business processes, best practices of *governance and auditability*, both manual and automated (Munoko et al., 2020) could be considered. The *8th* and *9th* step, the long-term monitoring, evaluate the system. Here, goals set in the first phase are to be evaluated, and the factual *social and environmental impact* can be measured regarding *beneficence* and *non-maleficence* of the AI (Prem, 2023).

Effects by Gender, Age and Studies. While the challenges in integrating ethical considerations into AI development are widespread, research has also highlighted differences in ethical perceptions and behavior across various demographic groups. These differences underscore the need for a nuanced and inclusive approach to addressing ethical issues in AI. Several studies have explored the role of gender in ethical perceptions and behavior, particularly in the context of information technology and engineering. According to Ulman et al. (2019), men reported unethical acts more frequently than women, although they did not view the importance of IT ethics significantly differently from females. Stappenbelt (2013) found that female engineering students evaluate themselves as acting in accordance with an ethical code of conduct more than their male counterparts evaluate themselves. Additionally, more female than male students believe that these codes of conduct can always be abided by. Kreth et al. (2022) further found that male computing students have lower social responsibility attitudes than female students, with a greater gender-based gap than prevalent in other domains. The findings are less conclusive, while age and educational level may also influence ethical perceptions and behavior. In a survey conducted by Ulman et al. (2019), minor differences in self-reported IT ethics behavior were observed between age groups, but these differences were not statistically significant. However, the differences between degree years (e.g., undergraduate vs. graduate students) showed to be significant. Similarly, both Almasri and Tahat (2018) and Harris (2000) found that students later in their studies tend to have higher ethical IT intentions than freshmen order undergraduates. The findings from both these studies go against the conclusion of Schiff et al. (2020) or Cech (2014) that the further technical students are in their study, the less they consider the social responsibility of their profession. Another factor closely related to the educational level is the technical skills acquired: Ulman et al. (2019) also found in their study that the computer skill level correlated with IT ethics behavior.

2.1.6 Education for AI Ethics Literacy

Prem (2023) lists multiple measures to ensure ethical standards are followed. Aside from the standards themselves and the technical approaches to implement them, one cornerstone is decision-maker education. Without awareness of ethical difficulties and possible mitigation methods, stakeholders cannot demand consideration of the principles, nor can developers bring them into practice (Domínguez Figaredo and Stoyanovich, 2023; Holstein et al., 2019). Thus, this knowledge is seen as motivation for both sides to follow up on ethical principles. This chapter will give an overview of the ethical aspects of technological education, especially AIEE. To do so, you will read about educational frameworks and applied courses in the field. Further, the empirical effects of these educational measures on professionals' and students' attitudes and behaviors on (ethical) AI are outlined based on empirical examples.

AI Education. To comprehend *AI Ethics Literacy*, it is important to understand technical AI literacy. While the first definition of AI literacy was the ability to understand the basic techniques and concepts of AI (Kandlhofer et al., 2016), more recent research includes understanding, applying, monitoring as well as critically reflecting AI applications (Long and Magerko, 2020; Ng et al., 2021a,b). While understanding and applying AI may be sufficient for the public to deal with AI in their everyday life, students in AI must understand AI on deeper levels for evaluation and development. With the beginning of the AI literacy era, education on AI spread from higher education in computer science into general classrooms and lecture halls of other domains (Ng et al., 2023). This includes domains where AI systems can be applied, such as engineering pedagogies and many more (Ng et al., 2021b, 2023; Schiff et al., 2020; Exter and Ashby, 2019). To meet expectations from the industry, these expert AI engineers are asked to be able to understand businesses, work with data, build models, develop software, and deal with operations engineering (Meesters et al., 2022). Especially occupationally, Verma et al. (2022) found that skills in decision-making and data mining, as well as programming, statistics, and big data, are most often asked for in AI and ML professionals.

AI Ethics Education Approaches. Contrary to this, the industry does not demand the developers to have certain attitudes, knowledge, or skills regarding ethical AI (Meesters et al., 2022; Verma et al., 2022). Similarly, many competence frameworks ignore professional ethics (Dutta et al., 2022). Still, some frameworks on the ability to work and develop AI systems, like AI literacy frameworks, include AI ethics (Laupichler et al., 2022; Knoth et al., 2024) or mention it comparably as responsible AI (Domínguez Figaredo and Stoyanovich, 2023). Educators find themselves between these different demands of 'scandals' in media, ethics as part of responsible AI guidelines, industry demands, and teaching guidelines when teaching for AI ethics literacy. AI ethics literacy itself does not yet have an established definition. Since AI literacy aims at sufficient skills for understanding evaluation, usage, and creation of AI (Kandlhofer et al., 2016), the authors of this thesis views *AI Ethics Literacy* as having sufficient ethical skills in understanding, critical evaluation, and responsible development of AI systems.

Many educational initiatives that educate aspects of ethical AI literacy are centered around technology ethics, ethics of sociotechnical systems, philosophy of computational systems, or focus explicitly on AI ethics (Brown et al., 2024; Fiesler et al., 2020; Stavrakakis et al., 2021). If ethics is taught in computational areas, ethics of AI and data science are most prominent (Stavrakakis et al., 2021). Many universities rely on a single ethics course to fulfill certification requirements for a rounded education in STEM (Schiff et al., 2020), while others set up single modules or integrated/embedded the ethical topics into technical courses or curricula (Brown et al., 2024; Hess and Fore, 2018; Stavrakakis et al., 2021; Garrett et al., 2020). These course-related integrations are relatively rare (Garrett et al., 2020) and are often within courses on topics of AI, such as multi-agent systems, machine learning, or natural language processing (Stavrakakis et al., 2021). (Hess and Fore, 2018) found that ethical courses for technical students are often set around three categories of learning outcomes: 1) ethical sensitivity or awareness, 2) Ethical judgment, decision-making, or imagination, and lastly, 3) Ethical courage, confidence, or commitment. Most of the educational approaches they analyzed were aimed at the first two, while the last was only used in a small part of the interventions. The main topics considered are codes of ethics, philosophical theories, moral dilemmas, and case studies Hess and Fore (2018). Certainly, argumentation and moral reasoning skills also stand central to these courses (Stavrakakis et al., 2021; Hess and Fore, 2018). Many courses introduce EAIPs via examples of automated ethical decision-making like predictive policing, facial recognition, and autonomous weapon systems (Fiesler et al., 2020; Garrett et al., 2020). Popular principles often referred to in teaching are lawfulness and policies, especially considering privacy and surveillance issues (Fiesler et al., 2020). Also, themes of inequality, justice, and human rights are popular topics of tech ethics courses (Fiesler et al., 2020). When concerning AI especially, ethics courses discuss fairness, discrimination, and bias within the realm of algorithmic justice (Garrett et al., 2020). ~~Additionally, topics of environmental and social impacts are among the more established topics of tech ethics courses (Fiesler et al., 2020).~~ Among multiple reviews on ethical tech education, philosophical theories have been identified to be used in most courses (Fiesler et al., 2020; Stavrakakis et al., 2021; Hess and Fore, 2018; Brown et al., 2024), often as an introduction to the discipline of ethics (Fiesler et al., 2020). These interventions use different pedagogies to educate their students: Aside from lectures as a method of exposure, students also actively participate through debates, or development of heuristics, own case studies, or even codes of ethics (Hess and Fore, 2018; Brown et al., 2024). Assignments often include exams, essays, classroom participation, or presentations (Brown et al., 2024; Hess and Fore, 2018; Stavrakakis et al., 2021). To consider that future AI developers might visit other technology ethics courses, that contextually educate them on AI ethics comparable to specific AI ethics courses, this thesis will use the terms of 'technology ethics' courses and 'AI ethics' courses interchangeably.

AI Ethics Education Outcomes. When the effects of these courses are evaluated, impact evaluation has often been based on interviews, students' term papers, or feedback on the course (Brown et al., 2024). A considerable amount of research investigated students' interest in ethics, awareness of social or ethical issues, and attitudes toward ethics using qualitative and quantitative measures (Hess and Fore, 2018; Stavrakakis et al., 2021).

Qualitatively, most qualitative approaches have been based on course evaluations, observation and rarely on focus groups or homework analysis (Hess and Fore, 2018). Still, many of the studies do not name the epistemological theories underlying their analysis (Hess and Fore, 2018). Many studies map students' post-class statements to dilemmas between ethical principles (Skirpan et al., 2018) or differences in personal and professional ethics (Skirpan et al., 2018). Skirpan et al. (2018) validated a five-week human-centered computing course with a pre-post survey on the effects of embedded ethics modules. In their qualitative analysis of open-ended questions, they found that the course led to an increase in perceived relevance for ethics in future careers and an increase in ethical behavioral intention when designing a system in a future job. Bullock et al. (2021) analyzed open-response data from 24 students and their lecturer about their experience of a course on ethics for predictive algorithms. They found that students' responses differed between pre- and post-class evaluations: students' negative responses were stronger regarding the impact of predictive policing on society. The instructor stated that the module filled an important gap in the course content. Fiesler et al. (2021) qualitatively analyzed students' written feedback on an ethics course regarding personalized advertisement and college admission algorithms. The authors identified that students were engaged, especially regarding issues with direct examples, such as algorithmic bias. Still, afterward, students equated ethics with plagiarism even though it was not discussed in the course and showed low interest in theoretical explanations of principles.

Quantitatively, many studies asked students about their perception of the intervention or their civic duty (Hess and Fore, 2018; Kilkenny et al., 2022). A smaller section of studies has included *ethical reasoning instruments* (Hess and Fore, 2018). These instruments are used to assess an individual's ability to reason through ethical dilemmas and make ethical decisions. There are generic and engineering-specific instruments in which participants navigate an ethical dilemma and to rank ethical principles by importance (Rest et al., 1999; Borenstein et al., 2010). To validate the instrument, Borenstein et al. (2010) compared engineering students who took courses on the ethics of technology with those who did not. The students with ethics education showed significantly better skills in moral reasoning. Still, this effect was mostly driven by only one of the three investigated courses. The authors see that some ethics curricula are better than others for certain learning goals. Adding to the instruments for generic and engineering ethics, Horton et al. (2022) developed a scale considering general ethical attitudes in computer science students. They measured it on integrated ethics lectures within a computer science course. They compared students in the course with others from outside the intervention and explicitly asked about ethical attitudes and interests. Both groups showed no initial differences in scores in the first pre-measure, but students' interests in ethics rose significantly through the treatment. Still, Horton et al. (2022) see the threat that students selected the courses by interest, which might have influenced the results. Kilkenny et al. (2022) reviewed answers from students on civic duty based on experiences from a service learning course in computer science. Service learning is a method where students provide service to a community partner. The Likert scale questions included items like *'I feel a personal obligation to help others'*. The experimental group indicated higher civic duty scores than students not participating in the learning experience.

Also, multiple other studies investigated the effect of these integrated ethics approaches. An early example of assessment for computer ethics education is the study by Byrne and Staehr (2004). They found a significant effect caused by four ethics sessions integrated into a computer science course. Students from the course showed significant increases in moral judgment compared to the control group outside the course. Regarding AI, one literature pattern is integrating ethics into introductory AI courses. Shih et al. (2021) dedicated a section of their three-week AI program to ethics. They found that students showed greater awareness of ethical AI problems throughout the course. They also found significant positive correlations between AI understanding and Attitudes towards AI on one side and awareness of ethical AI issues on the other. Kong et al. (2023a) found in their analysis of an AI literacy course with parts on ethics that students early in their education have problems in comprehending ethical AI principles. In another study, Kong et al. (2023b) found that integrated ethics in an AI literacy course can significantly increase levels of self-assessed ethical awareness by university students. Their additional qualitative analysis after the course showed that participants used significantly more ethical terminology, real-world examples, and references to ethical principles than before the course. In 2012 already, Cramer and Toll (2012) found that a course on algorithms and ethics of computing can increase students' tendency to reevaluate their technology usage. Mäses et al. (2019) validated their full course on cyberethics with a pre-post questionnaire asking about computer attitudes and ethical views on cyber ethics. Additionally, they included dilemma-based behavioral measures in homework assignments. Mäses et al. (2019) did not find correlations between attitudinal ratings on computer ethics and their ethical behavior in the homework assignments, and see a possible connection to the TPB for this gap between attitude and behavior. Sadeghi et al. (2022) similarly taught their students about cyberethics with a focus on security. They did not find differences in computational ethics behavior instilled by their participants playing a serious game. Also here, the researchers reported a gap between attitude and behavior. McNamara et al. (2018) implemented a very short intervention and verified that when developers connect with real-world stories, it increases their ethical decision-making. Contrary to their hypothesis, the developers in their study did not change their behavior when having the ACM code of ethics at hand instead of the real world news article. Mahmud et al. (2022) found that multiple other studies investigated the effects of awareness of problematic aspects of algorithms. Awareness of problems that are inherent to algorithms and direct experiences with them in education can increase the negative attitudes on AI (Fenneman et al., 2021; Liu et al., 2019). The same also holds for the trust in the systems, so that this is also decreased by these negative experiences (Merritt and Ilgen, 2008). Still, to this point, to the author there are no known studies that investigated ethics education on algorithm aversion and algorithm appreciation of AI students.

Overall, analyzing the effects of ethics education for technical students has quite a long history, emerging from engineering (Rest et al., 1999). Still, educational interventions on AI and AI ethics are fairly new (Garrett et al., 2020). Many previous studies found significant effects of ethical interventions, be it stand-alone courses or ethics integrated into other computer science or ethics courses, to significantly affect students' interest in ethics, perceived relevance of ethics for future careers, and their intended technology usage as well as their usage of ethical terminology and their abilities in moral judgment (Skirpan et al., 2018; Garrett et al., 2020; Byrne and Staehr, 2004; Kong et al., 2023b; Cramer and Toll, 2012). Still, issues occur with students not being interested in theoretical aspects or unable to delineate ethical issues clearly (Kong et al., 2023a). Additionally, certain interventions seem to be more effective than others (Kong et al., 2023a; Borenstein et al., 2010) Especially the effects of this education on AI students from a perspective of attitudes and behavior have only been investigated in a minority of cases, such as by Mäses et al. (2019). The same research gap occurs for AI students' attitudes to 'their' technology.

2.2 Hypotheses and Research Questions

Ethical AI guidelines are used as a measure when setting up systems that follow ethics by design (Prem, 2023). Schiff et al. (2020) and Harding et al. (2013) found that generally, technical and especially AI students do not have high levels of ethical literacy from their education nor interest in them - especially regarding questions on a societal level. If they consider these aspects, these students and their fully educated counterparts focus on principles of reliability, security, and privacy (Lu et al., 2022; Vakkuri et al., 2019b; Sanderson et al., 2022) with issues in actual implementation (Hedayati-Mehdiabadi, 2022). Most technological, computational, or AI ethics courses have shown positive results in raising awareness regarding AI ethics (Fiesler et al., 2020; Hess and Fore, 2018; Stavrakakis et al., 2021; Brown et al., 2024). Such courses often focused on concepts of privacy, fairness, and lawfulness

(Fiesler et al., 2020; Garrett et al., 2020). While these interventions seem to have effects on attitudes regarding the importance of AI ethics (Skirpan et al., 2018; Shih et al., 2021), more research is needed to understand the effects on individual principles as well as on behavior and the gap in between (Mäses et al., 2019).

2.2.1 Hypotheses on Attitudes on AI

There is not much research on AI students' or AI developers' attitudes towards AI. While the public has been considered in multiple studies (Kieslich et al., 2022; Bernnat et al., 2023), both algorithm aversion and algorithm appreciation/bias have been identified for users or laypeople (Lyell and Coiera, 2017; Burton et al., 2020; Logg et al., 2019). Still, there is only a small amount of research on people knowledgeable on AI (Center for Advanced Internet Studies, 2024; Mahmud et al., 2022). Generally, these show some positive tendency towards accepting, trusting, and using AI and identifying positive use cases for AI due to familiarity with the systems (Mahmud et al., 2022). Through their education, AI students have even higher familiarity and knowledge with AI systems. Additionally, they most often selected the topic purposefully, to learn more on AI (Barretto et al., 2021). Still, the effects of ethics education have not been investigated specifically. Ethics education has been shown to make people more aware of the ethical problems of AI and change their intention to use AI (Cramer and Toll, 2012). This adds to findings, that negative experiences and knowledge of problematic aspects of automation go along with aversion of these systems (Fenneman et al., 2021; Liu et al., 2019; Merritt and Ilgen, 2008). To investigate this area for future developers, this thesis takes questions from Kieslich et al. (2022) and the Center for Advanced Internet Studies (2024) and investigates whether there are differences between AI students who took a course on ethics and those who did not. For this, the variables of acceptance of AI, risk and opportunity awareness, usage intention AI in their projects, and trust in AI are used. It is hypothesized that more negative attitudes show itself among AI students, who have been confronted with critical societal questions in technology ethics courses (Schiff et al., 2020; Balebako et al., 2014; Vakkuri et al., 2019a). Referring back to the specific variables, it is hypothesized that,

- ***H1:*** AI students who took a course on technology ethics accept AI less than AI students who did not take such a course.
- ***H2:*** AI students who took a course on technology ethics show a greater awareness of the risks of AI than AI students who did not take such a course.
- ***H3:*** AI students who took a course on technology ethics show a lower awareness of the opportunities of AI than AI students who did not take such a course.
- ***H4:*** AI students who took a course on technology ethics show a lower usage intention for AI than AI students who did not take such a course.

- ***H5:*** AI students who took a course on technology ethics show lower levels of trust in AI than AI students who did not take such a course.

2.2.2 Hypotheses and Research Questions on Attitudes and Behavior Regarding Ethical AI Principles

Next, it is hypothesized that students with ethical education have developed positive attitudes toward considering ethical questions in AI development. Since courses on technology ethics often go beyond the typical focus on reliability and privacy (Skirpan et al., 2018; Fiesler et al., 2020), a difference is assumed for importance levels so that AI students give to different ethical AI principles. From this, it is hypothesized that

- ***H6:*** AI students who took a course on technology ethics rate Ethical AI Principles as more important than AI students who did not take such a course.
- ***H7:*** AI students who took a course on technology ethics give different attitudinal priority to individual Ethical AI Principles compared to AI students who did not take such a course.

While self-assessment questionnaires help to understand these attitudes, a more nuanced approach is needed for behavioral measures. Whereas there are established tools for moral judgment and decision-making among the public and engineers (Rest et al., 1999; Borenstein et al., 2010), so far, there is no comparable instrument for decision-making, especially for ethical AI. Thus, one solution to investigating the decision-making of future developers is to observe them when working on a real task. AI development often goes through many stages in the process and involves a multitude of developers (Prem, 2023). To consider these different ethical decision points (Prem, 2023) and perspectives, here a scenario is proposed, where AI students discuss how to design an AI system in groups. This discussion behavior can then be qualitatively analyzed and quantitatively mapped to understand the consideration of ethical AI principles. For this behavioral measure, the same argument as for attitudes is followed: Since ethical courses consider more and other ethical principles than usually prevalent among AI professionals, it is hypothesized that, on a group level, AI students with a background in ethics consider AI topics in their discussion behavior more often than students without. Additionally, it is asked whether diversified prioritization of ethical AI principles shows itself in behavior so that, e.g., students who took ethical courses consider ethical principles differently in their behavior than students without this educational background. I additionally set the research questions on how this different prioritization shows itself.

- ***H8:*** Groups of AI students who all took a course on technology ethics consider ethical AI principles more often than groups of students who all did not take a course on ethics.

- ***RQ1:*** Do AI students who took a course on technology ethics behaviorally prioritize ethical AI Principles differently than AI students who did not take such a course?
- ***RQ2*** What are the differences in the behavioral consideration of ethical AI principles between AI students who took a course on technology ethics and those who did not?

Regarding the gap between attitudes and behavior, Mäses et al. (2019) identified a missing correlation between ethical attitudes and ethical behavior regarding cyber ethics and Sadeghi et al. (2022) tried to close this gap with a single intervention without success. To investigate this for ethical AI development, this thesis asks whether there is an ABG among future AI developers. Additionally, it is important to ask whether there is a difference between AI students with and without experience from technology ethics courses.

- ***RQ3:*** Is there an attitude-behavior gap in the prioritization of ethical AI principles among AI students?
- ***RQ4:*** Are there differences between AI students who took a course on technology ethics and those who did not, with regard to an attitude behavior gap in prioritization of ethical AI principles?

Chapter 3

Methods

3.1 Overall Research Design

This experimental study combines hypothesis-based research with exploratory questions in a one-factor, two-level cross-sectional quasi-experimental subjects design. The method section explains the practical implementation of the research design to analyze the hypotheses and research questions on AI students' attitudes towards AI and EAIP, behavior regarding EAIP, and their ABG. While the order of topics as proposed in section 2.2 provides a good framework for a line of argumentation regarding the research literature to date, the order of this chapter deviates from this. The methods chapter first introduces the general research design and how the topic of this thesis is investigated with two studies. Thus, the design of *Study 1* for attitudes towards the importance of ethical AI principles and questions on attitudes towards AI is introduced. Afterward, the exploratory study design on discussion behavior is introduced for *Study 2* (Figure 3.1). The questionnaire-based Study 1 includes all participants (N = 98), while Study 2 has been conducted as a mini focus group (Kamberelis and Dimitriadis, 2005) lab study with a subsample (N = 48) of the first study (Figure 3.2. This means that participants of Study 2 also participated in Study 1: First, they were part of the mini focus group (*Study 2*) and then filled out the questionnaire (*Study 1*). Both samples differed in means on only 9 of 43 measured variables and are thereby sufficiently comparable. The split of both studies allows for a thorough investigation of future developers' attitudes toward AI ethics and an exploratory analysis of their ethical behavior in contrast to this. The results of the correlational analysis are uploaded to this thesis' project folder (Ben Schultz, 2024) in the Open Science Framework (OSF). Due to the two different studies, the both within the description of the sample as well as in the results (chapter 4), the presented data refers to varying total response sizes (represented by 'n'). Additionally, means (M) and standard deviations (SDs) of demographics and results are presented.

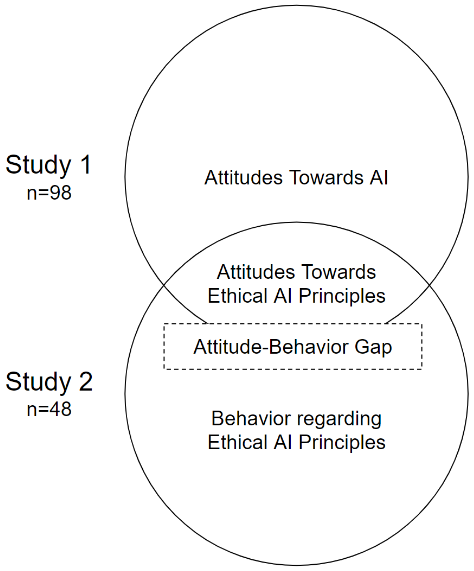

Study 1
n=98
Attitudes Towards AI
Attitudes Towards Ethical AI Principles
Attitude-Behavior Gap
Study 2
n=48
Behavior regarding Ethical AI Principles

Figure 3.1: *Study 1* investigates *Attitudes towards AI* as well as *Attitudes towards Ethical AI Principles (EAIP)*. Study 2 also takes the latter into account and investigates *Behavior regarding Ethical AI Principles* as well as a potential *Attitude-Behavior-Gap* regarding the Ethical AI Principles.

Even though the order of this chapter is not in line with the procedural nor argumentative order, it allows a better understanding of the sample and helps to discuss the comparison of self-assessment and attitudes with behavior for an exploratory investigation of the attitude-behavior-gap in ethical AI development. The Independent Ethics Committee at Rheinisch-Westfälische Technische Aachen University (RWTH Aachen University) reviewed and approved the study design under ID *EK 23-339* (Ben Schultz, 2024).

3.1.1 Overall Participation

Before participating, candidates had to confirm that they were enrolled in a study program that included courses on data science, machine learning, and other topics of AI. Additionally, participation requirements included being over 18 years old and speaking English fluently on a level similar to or above B2. Two different groups of students are compared with each other: Students in the experimental condition have had educational experience with ethics of technology, data, AI or comparable by taking a dedicated full-length course on the topic (n = 44). As a comparison, the control group did not take such a course (n = 54). A small section of students from the latter group heard something about technology ethics in their jobs or integrated within technical courses at university (n = 28). Students who had taken a short dedicated course on the topics (n = 3) or did not finish a full-length course (n = 4) have been considered for the group with a dedicated learning experience, even though these individuals must be considered with caution. Whether the students had experience with such a course has been asked after investigating experimental items. This order was selected to minimize socially desirable answers of participants (Randall and Fernandes, 1991). Students who took a dedicated course on technology ethics have been asked to indicate which course this has been. The largest group of students (n = 32) had participated in the course 'Ethics Technology and Data' and four students had taken the seminar on 'Social and Technological Change'. Six had individually visited other courses. Only two participants visited courses outside RWTH Aachen University: One had taken a course at another German university, and one had a training at his company. The experimental group with such education experience is termed by the previously introduced abbreviation *AI Ethics Education (AIEE)* for the remainder of this thesis. The control group is similarly termed *no AI Ethics Education (nAIEE)*. In both studies, participants had to be excluded. More on this will be discussed within the respective subsections (subsection 3.2.2 and subsection 3.3.2). Students have been recruited via convenience sampling. To achieve a balance between the conditions, the selection of recruitment channels was an important decision point. Lecturers of 31 courses have been contacted, of which 20 have been on technology ethics and comparable, to forward the information to their students or allow for a two-minute presentation of the study, at the beginning of their courses. Students did not receive benefits within these courses for participating in the study. Aside from the recruitment via selected courses, respective student councils and initiatives forwarded information on the study to their members. Flyers have been placed and handed out in the IT faculty building of RWTH Aachen University (Appendix A). Online platforms have been used for spreading a digital flyer and a short informational text (Appendix A). This includes instant messengers like whatsapp (WhatsApp LLC, 2024), telegram (Durov, 2024), or discord (Discord Inc., 2024), as well as social networking systems like linkedin.com (LinkedIn Ireland Unlimited Company, 2024) or studydrive.net (Studydrive GmbH, 2024). In a signup procedure they could decide which study to participate in (Appendix B). In case they participated in person, they could select a free slot via the web service of meetergo (2024) After signup, students received a confirmation mail for their future participation. Additionally, they received a reminder mail one day in advance. In the mails, future participants have been thanked for their interest in the study and received information on the location of the lab with the possibility to cancel or reschedule (Appendix C).

For participation, students had the opportunity to enter a lottery on three cash prices worth 200€ (100€, 50€, 50€) sent out after the study's analysis. 102 participants signed up for the lottery. The prize money was sent out after analysis of the study. The recruitment process included snowball sampling. This means that participants who recruited acquaintances received an additional lottery ticket for themselves as well as for their recruitees. Coming with its own downsides, this sampling technique aimed to achieve higher participation rates among the small population of students with AIEE. These courses are usually smaller and have fewer students than general technical courses (Fiesler et al., 2020; Hess and Fore, 2018).

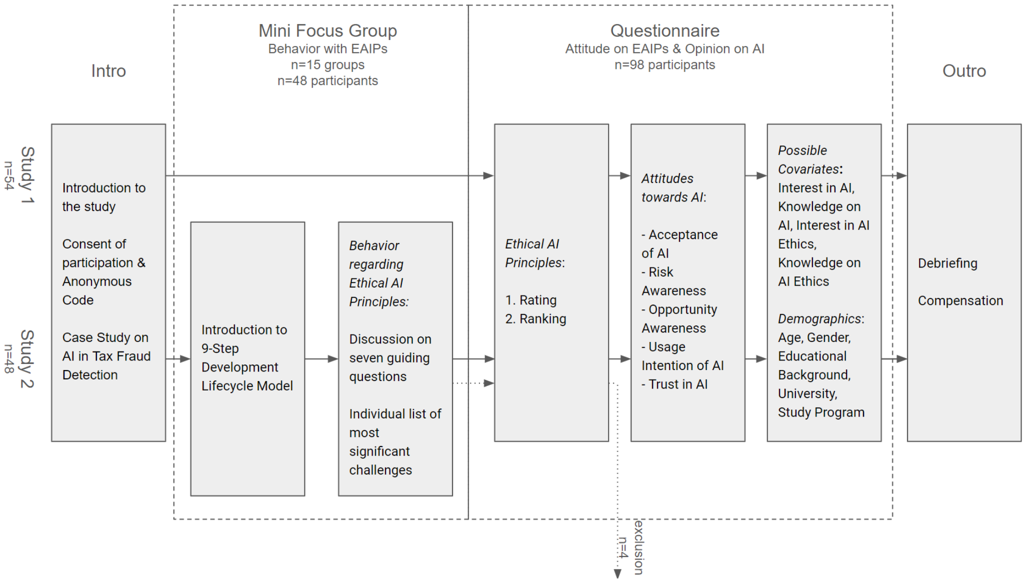

Intro
Mini Focus Group
Behavior with EAIPs
n=15 groups
n=48 participants
Questionnaire
Attitude on EAIPs & Opinion on AI
n=98 participants
Outro
Study 1
n=54
Study 2
n=48
Introduction to the study
Consent of participation & Anonymous Code
Case Study on AI in Tax Fraud Detection
Introduction to 9-Step Development Lifecycle Model
Behavior regarding Ethical AI Principles:
Discussion on seven guiding questions
Individual list of most significant challenges
Ethical AI Principles:
1. Rating
2. Ranking
Attitudes towards AI:
- Acceptance of AI
- Risk Awareness
- Opportunity Awareness
- Usage Intention of AI
- Trust in AI
Possible Covariates: Interest in AI, Knowledge on AI, Interest in AI Ethics, Knowledge on AI Ethics
Demographics: Age, Gender, Educational Background, University, Study Program
Debriefing
Compensation
exclusion
n=4

Figure 3.2: The study procedure: 54 participants only took the questionnaire of *Study 1*. 48 participants additionally participated in the mini focus groups of *Study 2*. Study 1 investigates the explicit Attitudes towards AI. Study 2 focuses on the evaluation of Ethical AI Principles within questionnaires and the consideration of Ethical AI Principles within the mini focus groups. Thereby it compares both in regard to an Attitude-Behavior-Gap in AI Development

3.1.2 Overall Material

AI systems and their application can differ widely. In this study, participants should gain a common understanding of AI, so that measures are comparable. Therefore, a cover story in the form of a case study was given to participants at the beginning of the study. The researcher introduced participants to the topic of AI in tax fraud investigation with a 620-word text (as in Kieslich et al. (2022)). Participants read about tax fraud in Germany, methods of tax fraudsters as well as tax authorities, the general legal process and different potentials of AI in the area of fraud investigation.

3.2 Methods of Study 1: The Questionnaire

3.2.1 Research Design of Study 1

Study 1 was conducted using an online questionnaire within the software SoSci Survey (Leiner, 2024). As with the research design of the overall thesis, individual students were compared in terms of their participation in AIEE. They rated and ranked EAIP by importance and indicated their attitudes towards AI. These included Acceptance of AI, Risk Awareness, Opportunity Awareness, Usage Intention of AI, Trust in AI.

3.2.2 Participation in Study 1

This subsection is meant to describe the recruitment as well as the sample of the study by their demographics (Table 3.1) as well as their knowledge and interest regarding AI and AI ethics (Table 4.1).

Participants were able to take part in the questionnaire in person or online. In person, they could come to the lab and directly participate at one of the given laptops. Online, they could use the link or QR code from the flyer to access the questionnaire.

The overall sample for Study 1 consisted of 98 students emerging from 369 clicks on the landing page. 23 participants had to be excluded from this study. 10 had little experience with AI (below a threshold of 3.33 on a 7-point Likert scale, see subsection 3.2.4), 14 failed the attention checks on two items within the questionnaire (Not selecting *'...choose "completely against".'* when asked for it). 42 students have been in the group with AIEE and 56 students have been in the control group with nAIEE. 3 students assigned themselves to the wrong condition and had to be reassigned based on qualitative analysis of the alleged technology ethics courses. While the majority of students were from RWTH Aachen University (n = 94), only four came from other *Universities*. 39 participants *studied* Computer Science/Computer Engineering/Computational Engineering. 6 students studied Data Science. 31 students came from the group of students in Data Analytics and Decision Science/Computational Social Systems. Another 22 students studied other fields s.a. 'Automation' (n = 3), 'Electrical Engineering' (n = 4) or 'Materials Engineering'/'Materials Data Science' (n = 5). The studies have been measured in groupings to optimize participants' privacy. This turned out to have a negative effect on the analysis as written in section 5.2. For the *highest educational qualification*, most students answered to have a bachelor's degree (n = 60) or A-levels (n = 20). A smaller section of students previously achieved a master's degree (n = 16). Only one had a specialist degree and one did not want to answer. The *age* range of students was measured within categories. Ten students indicated be between 18 and 21 years old. 40 have been in the category from 22 to 25 years, and 44 in the one from 26 to 30 years. Only 4 stated to be older than 30. 29 identified their *gender* as female, 64 as male and two as diverse. Three students did not indicate their gender. The percentage of female students in the sample was approximately 30% and comparable with the international average of females working in AI (Pal et al., 2023).

For *compensation*, in addition to the lottery ticket, students had the opportunity to receive 1 participation hour for their accreditation within their studies if needed. 16 students selected this option. If they did not and if they participated in this study in the lab, they had been eligible for additional compensation with 10€ cash

Table 3.1: Study 1 sample demographics, split by conditions of *AI Ethics Education* (AIEE) and *no AI Ethics Education* (nAIEE) for the university they are in, the studies they are enrolled in, their previous educational qualification, their age, and gender.

Variable	Characteristics	AIEE		nAIEE		Sum	
		n	%	n	%	n	%
University	RWTH Aachen University	41	41.84	53	54.08	94	95.92
	Other Universities	1	1.02	3	3.06	4	4.08
Studies	Computer Sc./Eng.	7	7.14	32	32.65	39	39.80
	Data Science	4	4.08	2	2.04	6	6.12
	DADS / CSS	27	27.55	4	4.08	31	31.63
	Other	4	4.08	18	18.37	22	22.45
Educat. Qual.	A-levels	1	1.02	19	19.39	20	20.41
	Bachelor's degree	28	28.57	32	32.65	60	61.22
	Master's degree	12	12.24	4	4.08	16	16.33
	Specialist degree	0	0.00	1	1.02	1	1.02
	No answer	0	0.00	1	1.02	1	1.02
Age	18-21 years	3	3.06	7	7.14	10	10.20
	22-25 years	11	11.22	29	29.59	40	40.82
	26-30 years	24	24.49	20	20.41	44	44.90
	> 30 years	4	4.08	0	0.00	4	4.08
Gender	Female	14	14.29	15	15.31	29	29.59
	Male	24	24.49	40	40.82	64	65.31
	Diverse	2	2.04	0	0.00	2	2.04
	No answer	2	2.04	1	1.02	3	3.06
Total		42	41.16	56	54.88	98	100

Note. Abbreviations are: Educat. Qual. = Educational Qualification, Univ. = University, Sc. = Science, Eng. = Engineering, DADS = Data Analytics and Decision Science, CSS = Computational Social Systems

(n = 51). The latter compensation was given out only to participants in person, due to both the benefit of a standardized procedure in the lab and bureaucratic reasons for money handling at RWTH Aachen University. This differentiation was considered and accepted by the ethics committee. All personalized information on compensation has been stored separately from the experimental data and cannot be merged without significant effort.

3.2.3 Procedure and Materials of Study 1

The questionnaires for Study 1 were conducted both online and in the lab. In the lab, participants either booked a slot or came by spontaneously. At the beginning of the questionnaire study (Appendix D), participants were informed about the context and procedure of the study. Aside from the summary of terms of participation, they could open and download a more detailed and elaborated version. After accepting the terms of participation and declaring the fulfilment of participation criteria, participants received an anonymous random code. This code could be noted down and sent to the researchers to request the deletion of the data. After this, participants have been informed that the study begins and read the use case of AI in tax fraud detection. Following this, they are asked to imagine being in the role of developers within the financial tax fraud investigation office. Next, they introduced the set of nine EAIPs, which are termed 'design principles' in the material so that participants do not know about the investigation of ethicality upfront (Gino et al., 2009). The text explained that these principles usually have to be weighed against each other to come to decisions within an AI development process. The principles are introduced with a brief description that fits the topic of the cover study of AI in tax fraud investigation (see Table 3.2) as done by Kieslich et al. (2022). These principles are adapted from the framework of Rao et al. (2021) for the simulation of an industrial work environment for which this framework of principles has been developed. Additionally, the EAIP set covers a wide range of aspects that need to be considered during a working project. Aside from aspects typical to discussions of ethics in AI like fairness, human agency, interpretability, benevolence (beneficial AI) and malevolence (safety) or accountability and data privacy (Prem, 2023), it also considers reliability and lawfulness. This wide range is especially helpful for a scientific comparison of the attitudes towards with behavioral consideration of different aspects that come up during development. Descriptions are adapted from the study by Kieslich et al. (2022) and the framework from Rao et al. (2021) and verified for proper formulation by a professor and a post-doc at an applied ethics chair.

After time for reading the EAIP, participants began to fill out the scales. They rated and ranked the EAIP and gave insights into their attitudes towards AI in regard to their *Acceptance of AI*, *Risk Awareness of AI*, *Opportunity Awareness of AI*, *Usage Intention*, *Trust in AI* as well as. Questions on knowledge and interest regarding AI and AI ethics followed this to investigate the general difference between the conditions. Lastly, they answered the demographic questions. To conclude, participants read a debriefing, which indicated that the study intended to investigate the effects of ethics education on their evaluations. This fact has not been disclosed nor indicated until this point to minimize participants acting in a socially desirable way (Randall and Fernandes, 1991). The participants were thanked for their participation and received their compensation. If they decided on monetary compensation, they signed a receipt.

3.2.4 Measurements & Method of Analysis of Study 1

As the first step of the questionnaire, participants gave a *Rating* of how important they find each EAIP (Table 3.2) to be (7 point-Likert-scale, 1 = not important at all, 7 = very important). Rating reached a close to acceptable reliability mean index (M = 5.904; SD = .129; α = .614).

Table 3.2: Ethical AI Principles with a description written within the use case of AI in tax fraud investigation.

Ethical AI Principles*	Description**
Interpretability (Explainability, Transparency, Provability)	*Explanation of the decision:* Each/any person concerned is explained in a generally understandable way why the system has classified him/her as a potential tax fraudster.
Reliability, Robustness, Security	*Reliable and secure tax fraud detection:* The automated identification of tax fraud by the computer system works consistently almost without errors, even in edge cases. It utilizes advanced security measures against hacker attacks and is always kept up to date with the latest security technology.
Accountability	*Full responsibility with the tax authority:* Should the automated tax investigation system lead to false accusations, the responsible tax authority bears full responsibility for any damage incurred.
Data privacy	*Use of data for a specific purpose only:* Only the necessary data is used by the automated tax investigation system. Any other use of the considered data is excluded.
Lawfulness and compliance	*Adherence to legal and regulatory requirements:* All stakeholders involved in the design and implementation of the algorithmic tax fraud detection system strictly comply with the law and relevant regulatory regimes. They ensure that the tax fraud detection systems' procedures and decisions are lawful and in accordance with established guidelines.
Beneficial AI	*Promoting the common good:* Both the process of development and the tax fraud detection system itself consider the common good of the society for safeguarding economic resources in an open, cooperative and sustainable way.
Safety	*Preservation of human well-being:* The algorithmic tax fraud detection system prioritizes (physical and psychological) safety of accused tax fraudsters throughout its operational lifespan, ensuring that it does not compromise their well-being until they are found guilty.
Human agency	*Appropriate human intervention:* The degree of human intervention required in the identification of tax fraud is dictated by the seriousness of ethical risks associated with the individual accusation.
Fairness	*No systematic discrimination:* No individuals (or groups) are systematically disadvantaged by the automated tax investigation.

Note. The set and description have been verified by a professor and a postdoc in applied ethics.

*The Ethical AI Principles have been framed as "Design Principles" in the study material to lower social desirability bias. EAIP taken from Rao et al. (2021)

**Descriptions adapted by Kieslich et al. (2022)

Next, participants *Ranked* the EAIP by importance on ranks from '1 = most important' to '9 = least important'. For both Rating of EAIP and Ranking of EAIP, they received the table with description on EAIP underneath the questionnaire interaction. Following, participants provided insights on their attitudes on AI. The set of questions is adapted from Kieslich et al. (2022).

For *Acceptance of AI*, participants stated whether they are rather for or against the use of AI in 12 different domains (e.g. financial institutions, court, health care, . . .) on a 5-point Likert scale (1 = completey against, 5 = completely in favor; No 6 = answer). The scale was adapted and translated from Došenović et al. (2022) and achieved an acceptable reliability mean index (*M* = 3.586; *SD* = .530; α = .798).

For *Risk Awareness of AI*, participants were asked to answer 'Completely independent of how big you think a possible benefit is, how great do you think is the risk posed by artificial intelligence?'. They were able to give their attitude 'for themselves', ' for their friends and family' as well as 'for the whole society' on a 10-point Likert scale (1 = 'no risk at all' to 10 = 'very high risk'). The scale was taken from Kieslich et al. (2022) who adapted it from Liu and Priest (2009). It received an acceptable reliability mean index (*M* = 5.648; *SD* = 1.039; α = .787).

Comparably, for *Opportunity Awareness of AI*, participants were asked to answer 'Completely independent of how big you think a possible risk is, how great do you think is the benefit posed by artificial intelligence?'. They could answer on their attitude 'for themselves,' 'for their friends and family', and 'for the whole society' on a 10-point Likert scale (1 = 'no opportunity at all' to 10 = 'very high opportunity'). The scale was taken from Kieslich et al. (2022) who adapted it from Liu and Priest (2009). It received an acceptable reliability mean index (*M* = 7.614; *SD* = .15; α = .801).

For *Usage Intention of AI*, the researcher included a translated version of the measurement by Došenović et al. (2022). It consists of 7 items in a 5-point Likert scale (1 = does not apply at all to 5 = applies completely, 6 = No Answer) asking for applicability of statements like 'I will always try to use the advantages of Artificial Intelligence.' or 'I will stay away from artificial intelligence wherever possible.'. The original version of the scale led to a low-reliability index mean (*M* = 3.725; *SD* = .258; α = .577). After the exclusion of two items (items 5 and 7), the advanced 5-item scale had an acceptable reliability mean index (*M* = 3.820; *SD* = .287; α = .709).

Trust in AI has been measured with a 12-item, 7-point Likert scale (1 = not at all to 7 = extremely) from the Checklist for Trust between People and Automation (Jian et al., 2000) adapted to AI by replacing 'they system' with 'AI". Participants answered, 'Please mark the point which best describes your feeling or your impression.' for items like 'AI provides security.' or 'I am confident in AI.'. Overall, the scale includes items both from literature on human-to-human trust and items on human-machine trust. The researcher calculated an acceptable reliability mean index (*M* = 4.15; *SD* = .343; α = .724).

Additionally, participants answered a scale *Responsibility attribution for actions of AI*. This possible independent variable is interesting for further investigation of the principle of *Accountability* and developers' self-perceived role in the AI development process but does not fit the focus of this report on developers' attitudes towards AI and EAIP. Thus, the variable is excluded from further analysis.

Additionally, four measures were taken to explore general group differences regarding their knowledge and interest in both AI and AI ethics. This was initially planned to investigate potential covariate-effects.

Knowledge on AI might affect the consideration of EAIP and evaluation of AI (Borenstein et al., 2010). Thus, the adjusted three-item AI steps knowledge subscale by Pinski and Benlian (2023) has been utilized for this. It is a self-assessment 7-point Likert scale (1 = strongly disagree to 7 = strongly agree) for items like 'I have knowledge of the input data requirements for AI'. The reliability mean index of the scale was good (*M* = 5.13; *SD* = .004; α = .805).

Interest in AI has been one factor in the studies by Shih et al. (2021) and Cramer and Toll (2012). In this study, the four-item 5-point Likert scale for *Interest in AI* from Kieslich et al. (2022) (1 = does not apply to 5 = applies completely) was enriched with two items for students in AI (e.g. 'In my studies, I take great interest in courses on artificial intelligence.'). The reliability mean index of the scale was good (*M* = 3.981; *SD* = .052; α = .846) within the study at hand.

Knowledge on AI ethics can be considered as a possible outcome of education in AI ethics (Kasinidou et al., 2021) but might also influence evaluation and usage of EAIP as well as attitudes on AI (Skirpan et al., 2018; Mäses et al., 2019). To verify the correlation of course experience with knowledge of EAIP, a scale from Kasinidou et al. (2021) has been adapted to the 9 EAIP used within this study. It is a self-assessment 5-point Likert scale for the question 'Please think about how knowledgeable you were before participating in this study' (1 = not at all, 5 = very knowledgeable). Underneath the scale, the description of principles was listed in Table 3.2. A good reliability mean index has been calculated for this scale (*M* = 3.346; *SD* = .170; α = .822).

Interest in AI ethics has been measured as well. Certainly, this interest in a topic can lead to better learning outcomes (Schiefele, 1992). At the same time, it could be a reason for students who show general interest to form strong attitudes on the topics as well as to behave differently; even without dedication to education on the topics. For this reason, we added an ad hoc four-item 7-point Likert scale (1 = not at all to 7 = completely) on the agreement of participants to statements like 'I enjoy discussing how technology affects society.' or 'I am interested in ethics of technology, data, algorithmic systems or autonomous systems.'. The scale showed to have a good reliability mean index (*M* = 5.894; *SD* = .016; α = .793). Still, the interest in a topic, as well as the knowledge, can also rise due to exposure to it (Horton et al., 2022).

Additionally, participants answered measures on *Length of Engagement With AI*, *Exposure to AI Development* as well as *Intention to Work in AI* out of the scope of this thesis. They also provided insights on the *Attended AI Ethics Course* and *Last Time a Technology Ethics Course Was Visited* and their *Position in the Political Spectrum*. In future, these insights can be used in future deeper and further analysis of the sample, especially after additional recruitment for a larger group of AIEE students. For this, study, the selected variables were sufficient to answer the aim of the study. All questions, including the excluded ones, are listed within the questionnaire that students received in Appendix D. While the appendix allows insights into the study material (Appendix D), different versions of the material can be found in the Open Science Framework (OSF) in Ben Schultz (2024).

All scales were presented to the participants in the order outlined above. Items within all scales for independent variables have been randomized. An a priori power analysis was conducted using G*Power 3.1.9.7 (Faul et al., 2020) to determine the sample size required to detect a large effect size (f = 0.4) with 80% power to run a one-way ANOVA's with the two conditions using an α of .05. The required sample size was 84 participants (power = .9518269). The final sample of 98 participants was sufficient to achieve the desired power Appendix E.

3.3 Methods of Study 2: The Mini Focus Group

3.3.1 Research Design of Study 2

Study 2 aims to exploratively investigate the ethical behavior of students in a discussion environment and compare them with the attitudinal ratings and rankings on EAIP from Study 1. Semi-structured mini focus groups have been set up to simulate a kick-off meeting of a development team. Each mini focus group only consisted of participants from the same condition. So, there were AIEE groups and nAIEE groups. Contrary to Study 1, Study 2 explored behavior both on an individual and on a group level: On a group level, discussion statements have been considered. On the individual level, individually written statements on the most significant challenges of AI development have been analyzed. This section of the thesis describes the sample of Study 2, the procedure followed, and the materials used, as well as the measurements and analysis methods.

3.3.2 Participation in Study 2

The sample for study 2 consisted of 48 students emerging out of 117 clicks on the sign-up landing page. Both conditions consisted of 24 participants. Four additional participants had to be excluded from the mini focus group study due to too little knowledge of AI (n = 2) or a mixed group constellation of AIEE and nAIEE (n = 2). Four more (n = 4) failed the attention check in the questionnaire on Attitudes

Toward AI. Still, since they neither showed outlying behavior within the mini focus group nor gave outlying ratings or rankings of EAIP, they have been included in Study 2 and for measurements on ABG. This resulted in 15 mini focus groups, with seven groups in the condition of AIEE and eight groups in the control sample of nAIEE. While the first one was conducted as a trial run, four mini focus groups had been excluded due to the exclusion of one of their participants each (Groups #1, #5, #12, #15). The group size ranged from two to four participants (*M* = 2.875, *SD* = 1.0246). The length of sessions ranged from 30:15 minutes to 00:55 minutes *M* = 40:01 minutes, *SD* = 07:17 minutes).

All 48 participants of study 2 (Table 3.3) came from RWTH Aachen University. 17 participants studied Computer Science/Computer Engineering/Computational Engineering. Two students studied Data Science. 12 students came from the group of students in Data Analytics and Decision Science/Computational Social Systems. Another seven students came from individual other studies s.a. 'Automation' or 'Materials Engineering'. For Highest Educational Qualification, most students had a bachelor's degree (*n* = 31) or came from school (*n* = 10). A smaller section of students previously received a master's degree (*n* = 5) or are masters of crafts (*n* = 2). The Age range of students was measured within categories. Three students indicated be between 18 and 21 years old. 16 have been in the category from 22 to 25 years, and 28 in the one from 26 to 30 years. Only one person stated to be older than 30. 16 identified as female, 31 as male and one as diverse. The percentage of female students in the sample was 33.3% and is comparable with the international average of 30% females working in AI as reported by Pal et al. (2023).

As for Study 1, students who participated in Study 2 had the opportunity to enter a lottery and to decide between compensation of cash or participation hours. As they participated in Study 1 and Study 2 together, they received either 20€ or 2 participation hours. 16 students decided for participation hours for curriculum, 27 for cash compensation. As in Study 1, participants gave their contact information for the lottery as well as participation hours. All this personalized information has been stored separately from the experimental data and cannot be merged for privacy reasons. The recruitment was similar, focusing on the universities in Aachen and respective lecturers and university buildings. Still, a few students (*n* = 9) have been recruited outside Aachen for three mini focus groups via the video calling software Zoom ('Zoom Video Communications, Inc.', 2024) and received their compensation in person in an arranged meeting after participation. Differences between online and offline participation cannot be reliably investigated due to the small sample size of students online. To establish groups split by conditions during sign up, participants have been asked to select courses/topics they had taken in their educational past. This list of 37 classes/topics on AI based on courses from the RWTH Aachen University module list included six that fit into the area of AIEE especially like e.g. 'Ethics, Technology, and Data', 'Ethics of Artificial Intelligence and Robotics' or 'Privacy Enhancing Technologies for Data Science'. This splitting procedure pre-sign up is shown in Appendix B. Within this procedure, the webservice of meetergo (2024) was integrated and linked for participant signup and management (section B.2). The participants selected a date and signed up for it,

Table 3.3: Study 2 sample demographics, split by conditions of *AI Ethics Education* (AIEE) and *no AI Ethics Education* (nAIEE) for the university they are in, the studies they are enrolled in, their previous educational qualification, their age, and gender.

Variable	Characteristics	AIEE		nAIEE		Sum	
		n	%	n	%	n	%
University	RWTH Aachen University	24	50	24	50	48	100
Studies	Computer Sc./Eng.	2	4.17	15	31.25	17	35.42
	Data Science	2	4.17	0	0.00	2	4.17
	DADS / CSS	18	37.50	4	8.33	22	45.83
	Other	2	4.17	5	10.42	7	14.58
Educat. Qual.	A-levels	1	2.08	9	18.75	10	20.83
	Bachelor's degree	19	39.58	12	25.00	31	64.58
	Master's degree	4	8.33	1	2.08	5	10.42
	Specialist degree	0	0.00	2	4.17	2	4.17
Age	18-21 years	0	0.00	3	6.25	3	6.25
	22-25 years	8	16.67	8	16.67	16	33.33
	26-30 years	15	31.25	13	27.08	28	58.33
	$>$ 30 years	1	2.08	0	0.00	1	2.08
Gender	Female	9	18.75	7	14.58	16	33.33
	Male	14	29.17	17	35.42	31	64.58
	Diverse	1	2.08	0	0.00	1	2.08
Total		24	50	24	50	48	100

Note. Abbreviations are: Educat. Qual. = Educational Qualification, Univ. = University, Sc. = Science, Eng. = Engineering, DADS = Data Analytics and Decision Science, CSS = Computational Social Systems

following the procedure described in subsection 3.1.1.

3.3.3 Procedure & Materials of Study 2

This subsection will elaborate on the procedure and materials used in Study 2.

After arrival, participants were greeted by the researcher and introduced to the study procedure as shown on Figure 3.2. Similar to Study 1, participants read and confirmed the participation requirements. The participants had time to read the case study on AI in Tax Fraud Investigation as well as into an AI development process framework. This framework is adapted from Rao et al. (2021). It is used to frame the mini focus group as a design discussion and to keep participants involved in the upcoming discussion. The participants also had the opportunity to know how far the discussion had gone and what was still to be discussed. Participants received these pages as well as the following questions in the form of a mini focus group guidebook (The researchers version can be found in Appendix F while the version for participants is uploaded in Ben Schultz (2024)).

There was time to answer questions, and the researcher started the mini focus group, introducing himself and his reason for being involved with AI. Participants have been asked to follow up with an introduction to create a comfortable, friendly environment. Participants consented to the audio recording of the following part: the researcher started the recording and began with the first question regarding the general understanding of the use case: 'What is your understanding of tax fraud detection systems and their technical complexities?'. He followed up with a question on how participants would approach data collection and algorithm selection. Participants were asked what they'd consider for rollout and integration into previous processes, as well as how they'd be measuring performance and the overall impact of the system. Afterward, they were asked to note in the guidebook and share what they considered to be the three most significant challenges when creating a good tax fraud detection system. As an ending question, participants followed up with some advice they'd give to one of the managers overseeing a development project for AI in tax fraud detection. Questions have been framed to minimize anchoring on ethical questions. The researcher guided the group through the discussion, aiming for a fair share of contributions, and took notes in the mini focus group protocol template Appendix F. The discussion of the mini focus group took between 28 minutes to 55 minutes (*M* = 39min, *SD* = 8min). The researcher thanked each participant for their input and referred the participants to the laptops, to follow up with *Study 1*. After finishing the steps of the questionnaire, participants were debriefed and informed about deception, that the reason for this study was the focus on ethical attitudes as well as ethical behavior. Before receiving their compensation, they had time to ask questions for deeper understanding. The combined procedure of study 1 and study 2 together took around 90 minutes.

3.3.4 Measurements & Method of Analysis of Study 2

Study 2 is concerned with the ethical behavior of future developers. Since the development of a real AI system might last for multiple months or years, a simulation of a working environment is one way to make this investigation scientifically feasible. Thereby, this study setup aimed to simulate a kickoff meeting of an AI development team coming together to discuss their upcoming project. To measure the behavior of participants, the mini focus group discussions have been transcribed and statements have been deductively coded onto the EAIP Table 3.2 via qualitative content analysis (Pearse, 2019). One exemplary transcript from each homogenous AIEE and nAIEE group can be found in Appendix G. The transcripts of all groups are uploaded to Ben Schultz (2024). The length of a statement considered in coding is defined by the length of a person speaking about one topic. If this topic matches the topic or subtopic of an EAIP as described in subsection 2.1.2, a counter is added to a list of considerations for each EAIP. If the statements refer to multiple principles at one time, the statement has been counted for each principle individually. If another participants refers to the same argument, another counter is added, indicating a higher relevance of the topic due to multi-participant consideration. Other themes and new emerging codes have not been taken into account due to the focus of the study on the EAIP in Rao et al. (2021). It has to be noted that the codings are interpretive and influenced by the author's expertise in the areas of computer science, as well as ethics and their interpretation of the EAIPS (Rao et al., 2021). Due to the scope of this thesis on, this has been conducted on a group level (Limitations are discussed in section 5.2). Additionally, the researcher deductively coded the written list of the three most significant challenges when creating an AI tax fraud detection system onto the EAIP.

Participants were only introduced to the EAIP in the form of "Design Principles" after the discussion, unaware of the principles' importance for the mini focus group. Together, both group and individual investigations allow for explorative insights into the ethical behavior in normatively driven development group settings and the developers themselves. For analysis regarding the ABG, the top three ranked EAIP from the questionnaire are compared with the coded EAIP from the written statements on the three most significant challenges. The rankings and statements are counted as either *mentioned* or *not mentioned* as an ordinal value. Each EAIP is investigated individually between and within the samples. The approach is elaborated on with more detail in section 4.4 due to decision-making in the procedure of data analysis.

The a priori power analysis for Study 2 was conducted to determine the sample size required to detect a large effect size (w = .5) with 80% power to run a Cramers V on the correlation between attitude and behavior over groups using an α of .05. The required sample size was 39 participants (power = .8074304). The final sample of 48 participants was sufficient to achieve the desired power. (Appendix E)

3.4 Overview on Methods

This study employed a mixed-methods approach combining quantitative and qualitative data collection through two interconnected studies. Study 1 utilized an online questionnaire to assess participants' attitudes towards AI and EAIPs, while Study 2 involved mini focus groups to explore ethical behavior in AI development discussions. The sample consisted of 98 students for Study 1, with a subset of 48 participating in Study 2. Participants were primarily from RWTH Aachen University, studying computer science, data science, and related fields with experience in AI. The sample was divided into two groups: those with AI ethics education experience (AIEE) and those without (nAIEE). In Study 1, participants rated and ranked the importance of nine EAIPs and completed questionnaires on AI attitudes, including acceptance, risk awareness, opportunity awareness, usage intention, and trust. The study also collected demographic information and assessed participants' knowledge and interest in AI ethics. Study 2 simulated a kickoff meeting for an AI development project focused on tax fraud detection. Participants engaged in group discussions, which were audio-recorded and later coded using the EAIPs framework. They also individually wrote down the three most significant challenges in creating an AI tax fraud detection system. The research design allowed for comparison between AIEE and nAIEE groups, as well as exploration of the ABG in ethical AI development. Data analysis involved both quantitative methods for the questionnaire responses and qualitative coding of discussion statements and written challenges.

Chapter 4

Results

This chapter presents the studies' results for hypotheses and research questions. First, as check for general influence of AIEE, the results on the measures of knowledge and interest in AI and AI ethics are presented. Secondly, the AI students' attitudes on AI are investigated inferentially. Afterward, the results for attitudes toward ethical AI principles are described, and inferential statistics are explained. The qualitative results of focus group behavior are shown and quantified. Afterward, the differences between students with and without education in technology ethics are investigated. This is followed by an analysis of the attitude-behavior gap. As in the methods section, due to the two different studies, the results presented have varying total response sizes (represented by 'N' for the whole response size and 'n' for a subsize). Additionally, means (M) and standard deviations (SD) are presented. Significance is assumed below a threshold value of $p = 0.05$ sufficient for the research topic (Andrade, 2019). Effect sizes are reported to indicate the magnitude of the effects, respectively. The hypotheses are approached with individual statistical methods to account for the specific circumstances of each measure. These decisions are explained in the corresponding chapter. Overall, there has been no opportunity to conduct a one-way ANOVA, due to missing normality in the data. The calculations are uploaded to Ben Schultz (2024). Statistical analysis has been conducted with SPSS (IBM Corp., 2021).

4.1 Results on Interest and Knowledge

In the questionnaire, answered on how much they know and how much they are interested in both AI and AI ethics. These variables have been analyzed for description of the two-group samples.

There have only been significant differences between the groups for knowledge on and interest in AI ethics, thus being described at this point. Students who did

Table 4.1: Differences between groups of AI students who had AI ethics education (AIEE) and students who did not (nAIEE) regarding their Knowledge on AI and on AI ethics as well as their interest in AI and AI ethics.

Variable	AIEE		nAIEE		Inferential Statistics		
	M	*SD*	*M*	*SD*	*t*(96)	*p*	Δ
	Sample of Study 1 (N = 98[a])						
Knowledge on AI	5.4	.795	5.35	.942	.298	.383	
Interest in AI	4.131	.530	4.012	.698	.923	.179	
Knowledge on AI ethics	3.706	.543	3.226	.695	3.708	.001****	.695
Interest in AI ethics	6.304	.746	5.563	1.043	3.914	.001****	.711
	Sample of Study 2 (N = 48[b])						
Knowledge on AI	5.5	.7986	5.278	.7965	.965	.17	
Interest in AI	4.243	.494	4.042	.525	1.368	.089	
Knowledge on AI ethics	3.880	.512	3.356	.742	2.848	.003	.742
Interest in AI ethics	6.594	.520	5.531	.851	5.218	.001****	.851

Note. *n* = sample size, *M* = Mean, *SD* = Standard Deviation, *t* = students t,
*p***** <.1, Δ = Glass's delta,
[a] *n* of AIEE = 42, *n* of nAIEE = 56,
[b] *n* of AIEE = 24, *n* of nAIEE = 24.

take a technology ethics course before (n = 42, *M* = 5.4, *SD* = .795) and students who did not (n = 56, *M* = 5.35, *SD* = .942), did not significantly differed in their self-assessments on AI knowledge. (*t*(96) = .298, *p* = .383). Students who did take a technology ethics course before (n = 42, *M* = 6.304, *SD* = .746) gave significantly higher self assessments on their interest on ethics of AI (*t*(96) = 3.914, < .001, *Glass's* Δ= 0.711) than who did not (n = 56, *M* = 5.563, *SD* = 1.043). Students who did take a technology ethics course before (n = 42, *M* = 4.131, *SD* = .53) and students who did not (n = 56, *M* = 4.012, *SD* = .698), did not significantly differ in their self-assessments on Interest in AI (*t*(96) = .923, *p* = .179). Students who did take a technology ethics course before (n = 42, *M* = 3.706, *SD* = .543) gave significantly higher self assessments on their knowledge on EAIPs (*t*(96) = 3.708, < .001, *Glass's* Δ= 0.691) than who did not (n = 56, *M* = 3.226, *SD* = .695).

Also, within the subsample that took study 2 there have been significant differences between the groups of students who did have and did not have education on technology ethics before regarding their knowledge on AI ethics (*t*(46) = 2.848, < .003, *Glass's* Δ= 0.742) as well as interest in AI ethics (*t*(46) = 5.218, < .001, *Glass's* Δ= 0.851).

4.2 Results on Attitudes towards on AI

All results on general attitudes towards AI are calculated by running a Welch's t-test to compare the ratings of students who took a technology ethics course with

Table 4.2: H1-H5 results. Differences between students with AIEE and nAIEE on Attitudes towards AI.

Variable	AIEE (n = 42)		nAIEE (n = 56)		Inferential Statistics.			
	M	SD	M	SD	t	*df*	*p*	Δ
Acceptance of AI	3.567	.590	3.637	.542	-.605	96	.273	
Risk Awareness	5.746	1.75	5.548	1.872	.534	96	.297	
Opportunity Awareness	7.698	1.459	7.702	1.404	-.014	96	.495	
Usage Intention	3.756	.721	3.881	.783	-.811	96	.21	
Trust in AI	4.137	.781	4.232	.600	-.683	96	.248	

Note. M = Mean, SD = Standard Deviation, t = students t, *df* = degrees of freedom, $p^* < .05$, $p^{**} < .025$, $p^{***} < .01$, Δ = Glass's delta.

those who did not. The data was non-normal, the variances showed to be non-homogeneous, and standard deviations differed (< .05). This led to the decision to use a Welch's t-test with an effect size analysis of Glass's Δ to account for these deviations from requirements for a regular student t-test and different sample sizes (Delacre et al., 2017). Results are displayed in Table 4.2.

H1. Regarding the ratings on *Acceptance of AI*, there was no significant difference between the two groups of students with (N = 42, M = 3.567, SD = .590) and without (N = 56, M = 3.637, SD = .542) education on technology ethics (t(96) = -.605, p = .273). The students gave positive ratings on the scale.

H2. Regarding the ratings on *Risk Awareness*, there was no significant difference between the two groups of students with (N = 42, M = 5.746, SD = 1.75) and without (N = 56, M = 5.548, SD = 1.872) education on technology ethics (t(96) = .534, p = .297). The students gave ratings in the mid-range of the scale.

H3. Regarding the ratings on *Opportunity Awareness*, there was no significant difference between the two groups of students with (N = 42, M = 7.698, SD = 1.459) and without (N = 56, M = 7.702, SD = 1.404) education on technology ethics (t(96) = -.014, p = .495). The students gave ratings in the positive range of the scale.

H4. Regarding the ratings on *Usage Intention of AI*, there was no significant difference between the two groups of students with (N = 42, M = 3.756, SD = .721) and without (N = 56, M = 3.881, SD = .783) education on technology ethics (t(96) = -.811, p = .210). The students gave ratings in the positive range of the scale.

H5. Regarding the ratings on *Trust in AI*, including aspects of human-to-human and human-to-machine trust, there was no significant difference between the two groups of students with (N = 42, M = 4.137, SD = .781) and without (N = 56, M = 4.232, SD = .6) education on technology ethics (t(96) = -.683, p = .248). The students gave ratings in the positive mid-range of the scale.

Table 4.3: H6 results. Differences between students with AIEE and nAIEE on attitudinal importance ratings of EIAPs

Variable	AIEE (n = 42)		nAIEE (n = 56)		Inferential Statistics.			
	M	SD	M	SD	t	df	p	Δ
Interpretability (Explainability, transparency, provability)	6.12	1.109	5.41	1.745	2.449	93.714	.012**	.406
Reliability, robustness, security	6.4	.885	6.13	1.251	1.235	96	.11	
Accountability	6.12	1.347	5.64	1.299	1.767	96	.04*	.366
Data privacy	5.76	1.574	5.86	1.354	-.321	96	.374	
Lawfulness and compliance	6.24	.958	6.46	.934	-1.174	96	.122	
Beneficial AI	5.71	1.132	5.11	1.603	2.196	95.702	.015**	.379
Safety	6.12	1.152	5.96	1.334	.602	96	.27	
Human Agency	5.83	1.286	5.04	1.684	2.559	96	.006**	.474
Fairness	6.52	.707	6.07	1.425	2.061	84.754	.021**	.317
All EAIPs	6.09	.530	5.74	.69	2.74	96	.004*	.508

Note. M = Mean, SD = Standard Deviation, t = students t, df = degrees of freedom, p* < .05, p** < .025, p*** < .01, Δ = Glass's delta.

4.3 Results on Attitudinal Consideration of EIAPs

H6. Similarly to the statistics for H1-5, Welch's t-test was selected due to the non-normality of the data, non-homogeneity, and standard differing deviations (< .05) to investigate whether the hypotheses were met. Effect size was estimated by using Glass's Δ to account for different sample sizes between the groups.

There was a significant difference between the two groups for rating the importance of ethical AI principles overall (t(96) = 2.74, p = .007) with a medium effect size (*Glass's* Δ= 0.508). This means that AI students who took a course in technology ethics (N = 42, M = 6.08, SD = .53) view ethical AI principles as more important than students who did not take such a course (N = 56, M = 5.74, SD = .69). This effect mainly shows itself in five principles: Interpretability (Explainability, Transparency, Probability) (t(93.714) = 2.449, p = .008, *Glass's* Δ= 0.406), Accountability (t(96) = 1.767, p = .04, *Glass's* Δ= 0.366), Beneficial AI (t(95.702) = 2.196, p = .015, *Glass's* Δ= 0.379), Human Agency (t(96) = 2.559, p = .006, *Glass's* Δ= 0.474) and Fairness (t(84.754) = 2.061, p = .021, *Glass's* Δ= 0.317). Generally, both groups gave high-importance ratings to the EAIPs as indicated in Table 4.3.

H7. To investigate future developers' priority of ethical AI principles' importance, their rankings of EAIPs had been reversed. For each participant, the principles

Table 4.4: H7 results. Differences between students with AIEE and nAIEE on attitudinal importance rankings of EIAPs

Variable	AIEE (n = 42)	nAIEE (n = 56)	Mann-Whitney U Test			
	MR	MR	U	z	p	η^2
Interpretability (Explainability, transparency, provability)	51.51	47.99	1091.5	-.611	.541	
Reliability, robustness, security	46.98	51.39	1070	-.769	.442	
Accountability	52.82	47.01	1036.5	-1.009	.313	
Data privacy	47.83	50.75	1106	-.506	.613	
Lawfulness and compliance	43.71	53.84	933	-1.758	.079	
Beneficial AI	49.74	49.32	1166	-.073	.942	
Safety	46.44	51.79	1047.5	-.930	.353	
Human Agency	54.85	45.49	951.5	-1.627	.104	
Fairness	49.96	49.15	1156.5	-.141	.888	

Note. MR = Mean Ranks, U = U statistic, z = z statistic, p* < .05.

on the rank with the greatest importance received the value 9, the second highest received the value 8, and so on until the last priority with value 1. A Mann-Whitney U test was performed to evaluate whether an AI student's ranking of EAIPs (ordinal) differed by taking a technology ethics course or not (nominal). The preconditions were met by having ordinally scaled variables and the independent variable of AIEE or nAIEE. The results indicated no significant difference between the groups for any ranked EAIP. Only the principle of *Lawfulness and Compliance* was close to a significant difference (z = -1.758, p = .079, η^2 = .032), ranked lower by AIEE (n = 42, *M rank* = 43.71) than by nAIEE (n = 56, *M rank* = 53.84). The results are listed in Table 4.4.

4.4 Results on Behavioral Consideration of EAIPs

All statements of students have been transcribed to compare the different behavioral considerations of EAIPs by groups of AI students with and without education in technology ethics (Two exemplary transcripts can be found in Appendix G. All other transcripts are available, uploaded to Ben Schultz, 2024). Afterward, deductive content analysis (Pearse, 2019) was run on statements of the AI students within each homogenous group (students either had taken or had not taken a course on technology ethics). The length of a statement is defined by the length of a person speaking about one topic. If this topic matches the topic or subtopic of an EAIP as described in subsection 2.1.2, a counter is added to a list of mentions for each EAIP. If the statements refer to multiple principles at one time, the statement has been counted for each principle individually. Other codes and themes have not been reported due to the focus of the study on the EAIPs in Rao et al. (2021). The codebook can be found in Table 4.5 with counts on mentions for each EAIP as well as examples for statements on the topics. It has to be noted that the codings are interpretive and influenced by the author's expertise in the areas of computer science, as well as ethics.

After transcription and coding of the focus group discussions and the individual statements as described in the methods subsection 3.3.4, the statements have been quantified regarding consideration of EAIPs. The codebook with themes, counts on mentions for each EAIP as well as examples for statements on the topics can be found in Table 4.5. In total, there have been 681 statements on EAIPs on the group level. Most of these statements focused on the EAIP of *Reliability, robustness, security* (*n* = 265), followed by Human Agency (*n* = 106). The least mentioned considered EAIPs have been *Accountability* (*n* = 15), *Lawfulness and compliance* (*n* = 28) as well as *Interpretability (Explainability, Transparency, Probability)* (*n* = 37). Individually, EAIPs have been considered 144 times. Again the most considered principle was *Reliability, robustness, security* (*n* = 70). The least considerations regarded *Accountability* (*n* = 3) and Safety (*n* = 3). The themes mentioned regarding each EAIP are outlined first in this section. Group ID is abbreviated with "G" and Speaker ID with "S". Afterward, you read about the results for the hypotheses and research questions on the behavioral consideration of EAIPs, first on a group and then on an individual level.

Considerations of Interpretability (Explainability, transparency, provability). Only a small section of statements were on *Interpretability (Explainability, transparency, provability)* often through referring to "explainability". The principle was mentioned 33 times in group discussion, with participants individually addressing it 8 times as in "Well, for me it's actually the issue of the black box" (G11, S5). Arguments centered around the need for transparency in AI systems for various stakeholders, including decision subjects, decision-makers, and developers. Participants emphasized the importance of understanding AI decision-making processes and the ability to explain these processes to affected individuals. One participant the necessity of explainability for developers on the example of eliminating bias, "And so the explainability of the [...] algorithm we will create [...] is a challenge, I think because sometimes without realizing it, you might have bias in it without knowing why or how or realizing it too late" (G18, S2). Another participant emphasized the right to explanation, noting, "if an AI system makes a, a judgement you have a right [... to] explanation for that" (G10, S5).

Considerations of Reliability, robustness, security. During the mini focus group sessions, participants most frequently discussed the principle of *Reliability, robustness, and security* (group *n* = 230, individual *n* = 70), often referring to it in terms of accuracy, false positive/negative rates, and system performance. Participants emphasized the importance of developing accurate and reliable AI systems for tax

Table 4.5: Focus group codebook coding EAIPs including themes, examples and counts of statements regarding each EAIP in group and individual work.

Code*	Theme	Example**	Group Count (n = 15)	Indiv. Count (n = 48)
1a	Transparency for decision-subjects Transparency for decision makers	"sometimes without realizing it, you might have bias in it without knowing why or how or realizing it too late" (G18, S2) "if an AI system makes a [...] judgement you have a right [... to receive] explanation for that" (G10, S5)	33	8
1b	Focus on accuracy False positives and negatives Optimize for Beneficial AI & Safety Multimodel approach Decision support system Security against hacking	"I think that [the percentage] should be the only criteria [...] of the success of this model" (G6, S2) "expect that there will be loopholes. Expect that it will be false positives" (G7, S3)	230	70
2a	Human accountability Decision support system	"it should be very carefully checked if it really works as we expected and does not, for example, has a bias or there's a lot of mispredictions" (G8, S3) "I would make sure that a responsibility gap is avoided by ensuring that decision making is always made by humans" (G17, S2)	13	3
2b	Context-dependent privacy	"Is the privacy point, really an issue? Because doesn't the government has all the data anyways [...]?" (G3, S4)	46	8
2c	AI Regulations Inclusion of law experts	"I mean because AI now doesn't even have the legal whatever power to do anything like this, like to make a decision on its own" (G6, S3) "it would require technical people as well as subject experts" (G3, S5)	28	9
2d	Economical benefit Procedural benefit Focus on major taxpayers Societal benefit	“So it's about like low effort, high result let's say” (G3, S3) “So it just saves time and you don't need to look at a few million pieces. Only at a few hundred” (G4, S4) “I think in that case it should be above certain tax crackers, because like, if I like stole like €50 from that” (G6, S2) “But it also creates a sense of fairness, because if everyone were paying their tax, everyone would have to pay less tax as well, right” (G19, S3)	44	11
2e	Needed accuracy Prevent wrong accusations Individual harm Economic harm Trustworthiness	“You need to have a high confidence level when you go with these results” (G7, S2) “Like prosecuting a person who is innocent, it's yeah. Like, not a good idea“ (G3, S2) “I'm wondering this is going to scare off all the small businesses in Germany” (G6, S3) “We should be made sure among ourselves that this is trustworthy” (G14, S1)	52	3

Continued on next page

Note. *EAIPs are used as codes as numbered in Table 2.1. **G = Focus Group ID, S = Speaker ID within the Focus Group.

Table 4.5 continued from previous page.

Code*	Theme	Example**	Group Count (n = 15)	Indiv. Count (n = 48)
2e	Needed accuracy Prevent wrong accusations Consider Individual harm Consider Economic harm Trustworthiness	“You need to have a high confidence level when you go with these results” (G7, S2) “Like prosecuting a person who is innocent, it’s yeah. Like, not a good idea“ (G3, S2) “I’m wondering this is going to scare off all the small businesses in Germany” (G6, S3) “We should be made sure among ourselves that this is trustworthy” (G14, S1)	52	3
2f	Human Involvement Decision support system Increase accuracy Trust in AI Job losses	“hopefully people are looking over it before it goes to court” (G4, S4) “I don’t trust AI to make the decisions, but it can point us in the right direction, yeah” (G4, S3) "Are we delegating the decision making to a machine or are we delegating [...] some groundwork to the machine [...]?" (G18, S4)	93	17
2g	Discrimination "Ethical AI" Protected attributes	“Because AI can also discriminate if you don’t train it, so it is not“ (G4, S1) “the ethical considerations regarding the data and the data collection. So to make sure that there is no discrimination basically“ (G4, S3) “if [...] it’s [...] just a linear regression and we are like trying to account for [...] protected categories, I think it’s easier to control“ (G17, S2)	69	15
Total			608	144

Note. *EAIPs are used as codes as numbered in Table 2.1. **G = Focus Group ID, S = Speaker ID within the Focus Group.

fraud detection, with many suggesting that high accuracy should be a primary criterion for success. One participant stated, “I think that [the percentage of flagged cases that were actually fraud] should be the only criteria [...] of the success of this model“ (G6, S2). Others recognized the imperfect side of AI models and the need for ongoing evaluation and improvement. As one participant noted, “expect that there will be loopholes. Expect that it will be false positives“ (G7, S3). Discussions also centered around the challenges of maintaining system reliability, including the need for robust data, handling missing or anomalous data, and adapting to changing laws and evolving fraud techniques. Security concerns were also raised, particularly regarding data protection and cybersecurity. Participants discussed the importance of secure data storage and the need for authorities to oversee cybersecurity measures.

Considerations of Accountability. Participants only occasionally referred to the principle of *Accountability*, although it was mentioned less frequently compared to other principles (group n = 13, individual n = 3). If it was mentioned, arguments centered on the responsibility associated with AI systems and their decision-making processes. Participants often mentioned the interconnection of accountability to other principles: One participant thus argued on with the problem of bias, stating, "it is a very big responsibility for the AI model, so it should be very carefully checked if it really works as we expected and does not, for example, has a bias or there's a lot of mispredictions" (G8, S3). Another participant suggested an approach to addressing accountability by human involvement, proposing, "I would make sure that a responsibility gap is avoided by ensuring that decision making is always made by humans and manually review what the model says instead of letting the model decide by itself because you cannot blame a model, but you can blame people" (G17, S2).

Considerations of Data privacy. More frequently (group n = 48, individual n = 8), participants discussed the principle of *Data privacy*, often in the specific context of its relevance to governmental systems like tax fraud investigation. The discussions revealed a range of perspectives on the importance and applicability of privacy in this context. Some participants questioned the relevance of privacy concerns in governmental systems, with one stating, "Is the privacy point, really an issue? Because doesn't the government has all the data anyways [...]?" (G3, S4). However, others expressed strong concerns about privacy, particularly regarding the use of personal financial information. One participant noted, "I actually have some ethical concerns also because we are taking personal data" (G4, S1). Another raised questions about consent and data usage, asking, "What about consent? Do we, do we ask people consent to use their financial data, I feel like that wouldn't likely happen, although it should" (G19, S2). The discussions also touched on legal aspects of data privacy, with participants mentioning GDPR and other data protection laws. One participant observed, "With regards to data protection law in Germany, I think it might be a little bit different to gather some information or yeah" (G10, S5).

Considerations of Lawfulness and compliance. A smaller set of statements referred to the principle of *Lawfulness and compliance* on both a local and international level (group n = 29, individual n = 9). Participants discussed the need for AI systems to operate within legal boundaries. One participant mentioned, "I mean because AI now doesn't even have the legal whatever power to do anything like this, like to make a decision on its own" (G6, S3). Based on this, the discussions included the complexity of adhering to legal standards, especially internationally. One participant noted, "No, it is quite difficult to track the shell company, especially when it is outside of your legality" (G6, S4). In this line, participants also mentioned the need for legal expertise in AI development. One participant stated, "it would require technical people as well as subject experts" (G3, S5). This sentiment was repeated in discussions about understanding and implementing specific regulations, such as GDPR.

Considerations of Beneficial AI The participants discussed the principle of *Beneficial AI*, often in the context of economic benefits, efficiency improvements, and societal impact of the AI system for tax fraud detection (group n = 44, individual n = 11). Statements surrounded the potential for AI to optimize processes and reduce workload for tax authorities. One participant noted, “So it’s about like low effort, high result let’s say” (G3, S3). Another participant emphasized the time-saving aspect, stating, “So it just saves time and you don’t need to look at a few million pieces. Only at a few hundred“ (G4, S4). Some arguments centered on the economic benefits of the system, particularly in terms of increased tax revenue and improved performance metrics. A participant suggested evaluating the system’s success by asking, “do we have more tax revenue or no or less? How many appeals at court do we have? How many of those succeed or not?“ (G4, S3). Also, there were statements about focusing the system on major taxpayers or high-value cases. One participant argued, “I think in that case it should be above certain tax crackers, because like, if I like stole like €50 from that” (G6, S2), suggesting that the system should prioritize larger-scale tax fraud for greater impact. One participant considered the broader societal implications of such a system and noted, “But it also creates a sense of fairness, because if everyone were paying their tax, everyone would have to pay less tax as well, right” (G19, S3).

Considerations of Safety The principle of *Safety* was considered a little more often (group n = 52, individual n = 2), especially in the context of preventing harm to individuals and maintaining trust in the system. The arguments revealed various concerns about the potential negative consequences of AI implementation in this domain. Central was the emphasis on the need for high accuracy of an AI system to ensure safety. One participant stated, “You need to have a high confidence level when you go with these results” (G7, S2). A primary concern was the risk of incorrectly accusing innocent individuals of tax fraud. One participant emphasized this point, stating, “Like prosecuting a person who is innocent, it’s yeah. Like, not a good idea“ (G3, S2). Another participant echoed this thought, noting, “Bringing a person into the court, and they did not do anything is kind of worse” (G16, S3), highlighting the potential for serious consequences resulting from false accusations. Participants also discussed the broader societal impacts of implementing such an AI system. One participant expressed concern about potential economic consequences, saying, “I’m wondering this is going to scare off all the small businesses in Germany” (G6, S3). The issue of job displacement was also raised by a participant, noting, “You can fire most of your tax Fraud Department office. So the model makes a lot of people unemployed because the the AI model is more efficient and cost less than people“ (G17, S2),. The issue of losing stakeholders’ trust was explicitly mentioned, with one participant stating, “We should be made sure among ourselves that this is trustworthy” (G14, S1).

Considerations of Human Agency More often participants argued on the principle of Human Agency (group n = 93, individual n = 17), often in the context of maintaining human control and decision-making power in AI-assisted tax fraud detection systems. The statements emphasized the importance of human involvement and oversight in various stages of AI development and deployment. Participants stressed the need for human review and intervention in the AI decision-making process. One participant stated, “hopefully people are looking over it before it goes to court” (G4, S4). Another participant emphasized, “I don’t trust AI to make the decisions, but it can point us in the right direction, yeah” (G4, S3), indicating a preference for AI as a decision-support tool rather than an autonomous decision-maker. The concept of a human-AI collaborative approach was frequently mentioned. One participant suggested, “Use both humans and AI together don’t rely on AI completely“ (G4, S3), while another proposed, “the AI does like 80 to 90% of the work and then the hard cases are still done by humans, but it’s still so much faster” (G13, S2). Participants also discussed the importance of human involvement in the development and deployment processes. One participant noted, “Like, what are we delegating? Basically, are we delegating the decision making to a machine or are we delegating [...] some groundwork to the machine, so it’s very important to kind of differentiate those“ (G18, S4).

Considerations of Fairness In the mini focus group sessions, participants frequently considered the principle of *Fairness* (group n = 69, individual n = 15), often in the context of bias, discrimination, and equitable treatment in AI systems for tax fraud detection. Participants expressed concerns about the potential for bias in AI systems. One participant noted on AI being fair, “Because AI can also discriminate if you don’t train it, so it is not“ (G4, S1). Another participant emphasized the need to consider “the ethical considerations regarding the data and the data collection. So to make sure that there is no discrimination basically“ (G4, S3). Participants also discussed the potential for AI systems to manifest existing societal inequalities. One participant noted, “So what this causes is, the IRS to forego targeting rich people who have the ability to combat these cases in court and just target the regular tax payer“ (G11, S3). Some participants suggested approaches to mitigate bias, such as using simpler models or accounting for protected categories. One participant proposed, “if you use more complex models you could be doing, you could be incurring like some kind of discrimination“ (G17, S2), while another suggested, “if [...] it’s [...] just a linear regression and we are like trying to account for like biases like gender and other protected categories I think it’s easier to control“ (G17, S2).

H8. After counting statements on EAIPs, a Mann-Whitney U test was run to compare how often the different mini focus groups with (n = 7) and without (n = 8) technology ethics education considered these principles (Table 4.6). Due to the small sample size and violation of the assumptions of normality ($<$ .05) as well as homogeneity of variance ($<$ .05), the non-parametric Mann-Whitney U test was chosen instead of a parametric student’s t-test. There were significant differences with large effects in the principles of *Reliability, Robustness, Security* (z = -2.613, p = .006, η^2 = .488) as well as *Data Privacy* (z = -2.217, p = .029, η^2 = .35) and *Beneficial AI* (z = -2.468, p = .014, η^2 = .435) between the conditions. Groups of students who took a technology ethics course considered *Data Privacy* (n = 7, M *rank* = 10.71) significantly more often than students who did not take such a course (n = 8, M *rank* = 5.63). The same held for *Beneficial AI* where AI students with such educational experience (n = 7, M *rank* = 11) mentioned it more often than students without (n = 8, M *rank* = 5.38). Contrary, the EAIP of *Reliability, Robustness, Security* was considered significantly more often by students who did not take a course on technology ethics(n = 8, M *rank* = 10.81) compared to students who did (n = 7, M *rank* = 4.79). When comparing the means of all EAIPs, these differences do not repeat themselves in the sum of all statements on EAIPs: Even though there was a tendency for more EAIP statements in groups

Table 4.6: H8 results. Group level differences between students with AIEE and nAIEE on behavioral consideration of EIAPs.

Variable	AIEE (n = 7)	nAIEE (n = 8)	Mann-Whitney U Test			
	MR	MR	U	z	p	η^2
Interpretability (Explainability, transparency, provability)	9.71	6.5	16		.189	
Reliability, robustness, security	4.79	10.81	5.5		.006***	.488
Accountability	8.29	7.75	26		.867	
Data privacy	10.71	5.63	9		.029*	.35
Lawfulness and compliance	9.64	6.56	16.5		.189	
Beneficial AI	11	5.38	7		.014**	.435
Safety	10	6.25	14		.121	
Human Agency	9.07	7.06	20.5		.397	
Fairness	9.93	6.31	14.5		.121	
Total	9.71	6.5	16		.189	

Note. MR = Mean Ranks, U = U statistic, z = z statistic (not reported due to small sample size),
p* < .05, p** < .025, p*** < .01, η^2 = effect size.

of students who had technology ethics education (n = 7, *M rank* = 9.71), the mean ranks were not significantly different (z = -1.391, p = .189) to groups of students without (n = 8, *M rank* = 6.5).

RQ1 and RQ2. To understand whether individual students with and without education technology ethics prioritize EAIPs differently, the three greatest challenges in designing a good AI system, as stated by students in the discussion individually, have been analyzed. As for H3, the researcher ran a deductive content analysis and categorized each statement onto an EAIP. Each statement on an EAIP incremented the count of considerations of this EAIP by 1. By this, each participant had three mentions that have each been mapped to an EAIP. Participants might have addressed the same EAIP within several statements. These have been counted as two considerations of an EAIP. With this, it is possible to analyze how often an EAIP is considered one of the top three important challenges in AI development. The counts can be viewed in table Table 4.5. The most popular EAIPs were *Reliability, robustness, security*, considered in 62 statements, and next *Human Agency* in 15 statements, as well as *Fairness* in 13 statements. *Accountability* (n = 3) and *Safety* (n = 2) have been addressed in the least statements on the greatest challenges in developing a good AI. Both groups mentioned *Reliability, robustness, and security* much more often than the other EAIPs. This effect is strongly influenced by considerations of AI students who did not take a course on technology ethics. Similarly, considerations of *Human*

Agency are driven by the nAIEE condition. Students who did take a course on technology ethics put the highest priority on *Interpretability (Explainability, Transparency, Probability)* and *Fairness*. Among students with no experience from technology ethics education, the EAIPs of *Interpretability (Explainability, Transparency, Probability)*, *Accountability*, *Data Privacy*, *Safety*, and *Fairness* have only been mentioned either once or never as the greatest challenges to solve for developing a good AI system. For students with ethics education, this only occurred for the EAIP of *Lawfulness and Compliance*. Overall, descriptively, there is a smaller standard deviation between EAIPs in ratings from students who took a course on technology ethics (*SD*=6.928) compared to students who did not (*SD*=12.61). This indicates for a more balanced consideration of EAIPs in behavioral prioritization by students who had taken a technology ethics course.

After counting statements, a Mann-Whitney U test was run to compare how often the students with (n = 24) and without (n = 24) technology ethics education considered these principles highly important during their behavioral interaction. Again, this test was selected due to non-normality of the data (< .05). The results (Table 4.7) indicate a significant difference for four EAIPs: *Interpretability (Explainability, Transparency, Probability)* was considered as important significantly more often (z = -2.299, p = .021, η^2 = .112) by students who took a course on technology ethics (n = 24, *M rank* = 27.5) than by students who did not (n = 24, *M rank* = 21.5). The same direction showed for *Fairness* being rated significantly more often as important (z = -3.769, p = <.001, η^2 = .302) by students who took a course on technology ethics (n = 24, *M rank* = 30.52) than by students who did not (n = 24, *M rank* = 18.48). The opposite direction was prevalent for *Reliability, robustness, security*, where students who took courses on technology ethics (n = 24, *M rank* = 19.58) considered the principle significantly less often as important (z = -2.523, p = .012, η^2 = .135) than students who did not take such a course (n = 24, *M rank* = 29.42). Similarly, *Human Agency* was considered significantly less often as important (z = -2.688, p = .007, η^2 = .154) among students from the AIEE condition (n = 24, *M rank* = 20) compared to from students from the nAIEE condition (n = 24, *M rank* = 29). Among the other five EAIPs two showed non-significant differences for *Accountability* (z = -1.77, p = .077) or *Data Privacy* (z = -1.533, p = .125), with greater consideration among students who took a technology ethics course, as shown in the results table.

In summary, the significant differences in behavioral prioritization of EAIPs between the conditions of AIEE and nAIEE indicate that AI students who took a course on technology ethics and those who didn't prioritize EAIPs differently in their behavior. Generally, students with education in technology ethics considered more different EAIPs as important and did focus less on *Reliability, robustness, security* as done by students without this education.

Table 4.7: RQ1 & RQ2 results. Individual differences between students with AIEE and nAIEE on behavioral consideration of EIAPs

Variable	AIEE (n = 24)	nAIEE (n = 24)	Mann-Whitney U Test			
	MR	MR	U	z	p	η^2
Interpretability (Explainability, transparency, provability)	27.5	21.5	216	-2.299	.021**	0.112
Reliability, robustness, security	19.58	29.42	170	-2.523	.012**	.135
Accountability	26	23	252	-1.77	.077	
Data privacy	26.5	22.5	240	-1.533	1.25	
Lawfulness and compliance	23.44	25.56	262.5	-.813	.416	
Beneficial AI	25.58	23.42	262	-.760	.447	
Safety	25	24	276	-.590	.555	
Human Agency	20	29	180	-2.688	.007***	.154
Fairness	30.52	18.48	143.5	-3.769	<.001****	.302

Note. MR = Mean Ranks, U = U statistic, z = z statistic (not reported due to small sample size),
$p^* < .05$, $p^{**} < .025$, $p^{***} < .01$, $p^{****} < .001$, η^2 = effect size.

4.5 Results on Attitude Behavior Gap

RQ3. To identify whether there is an ABG for consideration of EAIPs, the three highest-ranked EAIPs from the attitudinal considerations are compared with the three most significant challenges considered in the discussion. For this, the attitudinal ranking is cut off at the threshold of rank three. Only the three highest-ranked EAIPs by each participant are compared with the three EAIPs from statements on most significant challenges (as coded for R1 and R2 in section 4.4). Because the behavioral data from RQ1 and RQ2 includes EAIPs, which could be considered more often than once, and the attitudinal data does not, the behavioral data has multiple levels (ordinal) while the attitudinal data does not (nominal). Thus, since the attitudinal values are nominal (1='the EAIP is considered important', 0='the EAIP is not considered important') the ordinal behavioral values are turned into nominal measures to make them comparable. For this, multiple considerations are handled as if there has only been one consideration. The downsides of losing information on multiple considerations of individual principles are discussed in section 5.2.

Camer's V was calculated to investigate possible correlations in attitudinal and behavioral nominal data for each EAIP. This calculation includes both students with

Table 4.8: RQ3a results. Results on Correlation Analysis with Cramer's V between attitudes and behavior for EAIPs considered as important. No significant correlations between attitude and behavior among all participants who did study 1 and study 2 (N = 48)

Variable	Cramer's V	$\chi^2(1)$	p
Interpretability (Explainability, transparency, provability)	.133	.854	.356
Reliability, robustness, security	.205	2.026	.155
Accountability	.183	.206	.206
Data privacy	.082	.323	.57
Lawfulness and compliance	.079	.3	.584
Beneficial AI	.059	.168	.682
Safety	.012	.006	.936
Human Agency	.031	.046	.831
Fairness	.004	.978	.978

Note. $\chi^2(1)$ = Chi-Square value (degrees of freedom), $p^* < .05$.

(n = 24) and without education (n = 24) in technology ethics to account for an overall ABG. The results did not indicate any significant correlation between attitudinal and behavioral correlation for any EAIP. Correlational results are collected in Table 4.8.

Additionally, besides investigating similarities with Cramers V, due to the small amount of groups (n < 30) exact McNemar's tests have been conducted to test for significant differences of each EAIPs between the attitudinal and behavioral nominal measures. In this case, McNemar's test investigates how likely it is that an AI student who showed attitudinal consideration of an EIAP (A=1) did also show behavioral consideration (B=1) of the same EAIP (A=1 → B=1) against how likely it is that an AI student who showed attitudinal consideration of an EAIP did not show behavioral consideration of the same EAIP (A=1 → B=0). The results indicate significant differences between the attitudinal and behavioral measurements for all principles except of *Reliability, Robustness, Security*. These results are listed in Table 4.9.

Interpretability (Explainability, Transparency, Probability) is considered significantly more often ($\chi^2(1)$ = .854, p = <.001, ϕ =-.133) attitudinally than behaviorally with a small effect. This means that there is statistically significant ABG for *Interpretability (Explainability, Transparency, Probability)*, in the sense that when participants considered the principle attitudinally, they were less likely to consider the principle behaviorally.

Table 4.9: RQ3b results. Results on Differential Analysis with Mc Nemar's test between attitudes and behavior for EAIPs considered as important.

Principle	$\chi^2(1)$	p	ϕ
Interpretability (Explainability, transparency, provability)	0.854	<.001****	-.133
Reliability, robustness, security	2.026	.108	
Accountability	1.600	<.001****	-.183
Data privacy	0.323	<.001****	.082
Lawfulness and compliance	0.300	<.001****	-.079
Beneficial AI	0.168	<.001****	-.059
Safety	0.006	<.001****	.012
Human Agency	0.046	.006***	.031
Fairness	0.001	.002***	.004

Note. $\chi^2(1)$ = Chi-Square value (degrees of freedom),
p* < .05, p** < .025, p*** < .01, p**** < .001, ϕ = effect size.

Reliability, Robustness, Security is not considered significantly more often attitudinally than behaviorally ($\chi^2(1)$ = 2.026, p = .108). This means that there was no ABG by AI students on the EAIP of *Reliability, Robustness, Security*.

Accountability is considered significantly more often ($\chi^2(1)$ = 1.6, p = <.001, ϕ =-.183) attitudinally than behaviorally with a small effect. This means that there is statistically significant ABG for *Accountability*, in the sense that when participants considered the principle attitudinally, they were more likely not to consider the principle behaviorally.

Data Privacy is considered significantly less often ($\chi^2(1)$ = .323, p = <.001, ϕ =.082) attitudinally than behaviorally with a very small effect. This means that there is statistically significant ABG for *Data Privacy*, in the sense that when participants considered the principle attitudinally, they were slightly more likely to consider the principle behaviorally.

Lawfulness and Compliance is considered significantly more often ($\chi^2(1)$ = .3, p = <.001, ϕ = -.079) attitudinally than behaviorally with a very small effect. This means that there is statistically significant ABG for *Lawfulness and Compliance*, in the sense that when participants considered the principle attitudinally, they were less likely to consider the principle behaviorally.

Beneficial AI is considered significantly more often ($\chi^2(1)$ = .168, p = <.001, ϕ = -.059) attitudinally than behaviorally with a very smal effect. This means that there is

Table 4.10: RQ4 results. Differences between students with AIEE and nAIEE for ABGs for EAIPs.

Variable	$\chi^2(1)$	p	ϕ
Interpretability (Explainability, transparency, provability)	.091	.763	
Reliability, robustness, security	.000	1	
Accountability	.000	1	
Data privacy	1.061	.303	
Lawfulness and compliance	6.454	.024*	-.367
Beneficial AI	2.021	.155	
Safety	.403	.525	
Human Agency	2.021	.155	
Fairness	7.111	.017**	-.385

Note. $\chi^2(1)$ = Chi-Square value (degrees of freedom), $p^* < .05$, $p^{**} < .025$.

statistically significant ABG for *Beneficial AI*, in the sense that when participants considered the principle attitudinally, they were less likely to consider the principle behaviorally.

Safety is considered significantly less often ($\chi^2(1)$ = .006, p = <.001, ϕ = .012) attitudinally than behaviorally with a very small effect. This means that there is statistically significant ABG for *Safety*, in the sense that when participants considered the principle attitudinally, they were more likely to consider the principle behaviorally.

Human Agency is considered significantly less often ($\chi^2(1)$ = .046, p = .006, ϕ = .031) attitudinally than behaviorally with a very small effect. This means that there is statistically significant ABG for *Human Agency*, in the sense that when participants considered the principle attitudinally, they were less likely to consider the principle behaviorally.

Fairness is considered significantly less often ($\chi^2(1)$ = .001, p = .002, ϕ = .004) attitudinally than behaviorally, with an extremely small effect. This means that there is statistically significant ABG for *Fairness*, in the sense that when participants considered the principle attitudinally, they were less likely to consider the principle behaviorally compared to considering it attitudinally.

RQ4. To investigate the AGB between groups of students who took a technology ethics course and students who did not, changes in attitudinal and behavioral considerations are compared. Whenever a participant shifted consideration of an EAIP, this was noted down. If they shifted from considering an EAIP attitudinally to not considering it behaviorally, this was marked as "1" for an indicator of an individual instance of an ABG. If they shifted from not considering the EAIP attitudinally to considering it behaviorally, this was marked as "0" as no ABG occurred. If they showed the same consideration in attitudinal and behavioral measures, this was marked as "0", as no ABG occurred.

A Chi-Squared test with a contingency table was calculated to compare the two groups of AI students who did and did not take a course on technology ethics Table 4.10. The results indicate a significant difference for the two EAIPs of *Lawfulness and Compliance* as well as *Fairness* between the two groups. For *Lawfulness and Compliance* AI students who had ethics education showed significantly less ABGs than those who did take an ethics course ($\chi^2(1) = 6.454$, p = .024, ϕ = -.367). Similarly, for *Fairness* AI students who had ethics education showed significantly less ABGs than those who did take an ethics course ($\chi^2(1) = 7.111$, p = .017, ϕ = -.385). This means taking an ethics course accounts for a moderate effect on showing less occurrences of ABGs regarding both *Lawfulness and Compliance* and *Fairness*. There were no significant differences for any of the other EAIPs. Especially for *Reliability, robustness, security* and *Accountability*, there were no differences for ABG between the groups.

Chapter 5

Discussion of Results

This study aimed to investigate the differences between students who had received AI ethics education (AIEE) and those who had not (nAIEE), with a particular focus on attitudes on AI and their ethical attitudes and behavior related to AI development. This chapter considers previous research to contextualize the findings from this study. Firstly, the foundational differences between the AIEE and nAIEE regarding interest and knowledge in AI and AI ethics are discussed. This is followed by a contextualization of missing differences in attitudes towards AI, hinting at an algorithm appreciation among AI students.

5.0.1 Higher Knowledge and Interest on AI Ethics Among Students with AI Ethics Education

As part of the study, four variables have been measured as potential descriptors of the study: self-assessed knowledge of AI, interest in AI, knowledge of AI ethics, and interest in AI ethics. These measurements were intended to provide context and explain for potential confounding factors in our primary analysis. The results revealed significant differences between the AIEE and nAIEE groups in two of these variables: knowledge of AI ethics and interest in AI ethics. These findings, while not of main interest in this study, offer valuable insights into the impact of AI ethics education on students' self-perceptions regarding ethical issues in AI. More importantly, these differences also provide an important backdrop against which to interpret the main results on attitudes toward AI as well as attitudes and behaviors towards EAIPs. In this discussion, we will first examine the results on knowledge and interest and their implications, drawing on relevant literature in the field of AI ethics education. We will then proceed to analyze our primary findings on attitudes and behaviors, considering how they may be influenced by or related to these differences in ethical knowledge and interest. This approach allows us to present a more nuanced understanding of the complex interplay between AI ethics education, ethical awareness, and AI students' perception of AI technologies.

The results on knowledge and interest in AI did not indicate significant differences between students who had AI ethics education and those who did not. At the same time, in their self-assessed knowledge and interest in AI ethics, students with AIEE gave significantly higher ratings compared to their counterparts without AIEE. These findings align with previous research indicating that ethics education can enhance students' awareness and engagement with ethical issues (Skirpan et al., 2018; Kong et al., 2023a). This suggests that while AI ethics education significantly impacts AI students' understanding and interest in ethical issues, it does not necessarily add to their previous interest or knowledge in AI itself. This finding is consistent with the literature, which often highlights the challenge of integrating ethics into technical education without diluting the technical content (Kong et al., 2023a). The significant differences in AI ethics knowledge and interest underscore the importance of incorporating ethics education into AI curricula. However, the lack of significant differences in AI knowledge and interest suggests that ethics modules should be carefully designed to complement rather than compete with technical content. This balance is crucial to ensure that students gain a holistic understanding of AI and its ethical implications.

5.0.2 No Difference on Attitudes Toward AI

Investigating AI students' attitudes toward AI is an approach aiming to understand, whether they tend to apply AI-based technology in their future work, potentially as developers. While some applications of AI can lead to negative outcomes, it is worthwhile investigating whether students who have been taught about these complex ethical outcomes tend to show more negative attitudes than others without this education. Overall, in this study, there have been no significant differences between the two groups for any of the five measures, rejecting Hypotheses 1-5. The implications of tendencies of algorithm aversion and algorithm appreciation are discussed below.

No Differences by AIEE for Acceptance of AI. Previous studies indicated that (future) AI developers tend to have positive attitudes toward AI systems: People with greater AI familiarity or skills in mathematics show higher acceptance of AI (Fenneman et al., 2021; Thurman et al., 2019; Logg et al., 2019). In this study, AI students were asked to give acceptance ratings on applying AI in different domains. Generally, they indicated relatively high scores. Furthermore, there was no difference between students with and students without AIEE. This means even though that students in the group with AIEE indicated to have more knowledge of AI ethics and more interest in ethical and societal implications of AI, this did not translate to lower acceptance of AI in applied domains. This goes along with findings, that computer science students have issues hard translating ethics to a broader societal picture (Schiff et al., 2020). It could also be explained by their regular exposure and familiarity with AI (Mohanani et al., 2018), leading to a positive evaluation that could not be irritated by negative exposure through AIEE.

No Differences by AIEE for Risk Awareness of AI. Some practitioners, investigated in other studies, seemed to have a simplified understanding of AI risk, (Sanderson et al., 2022; Vakkuri et al., 2019b) like by viewing risks of AI as risks to business objectives (Pant et al., 2024a) instead of risks for humans or sociotechnical systems. People with higher knowledge of AI tend to see fewer risks in AI than benefits (Center for Advanced Internet Studies, 2024). While other studies indicated that AIEE increases awareness of societal and social issues of AI (Fenneman et al., 2021; Liu et al., 2019), this effect did not show here. The sample rather gave ratings of perceived risk around the mid-point of the scale, comparable to ratings by the public in Center for Advanced Internet Studies (2024). It might be the fact, that the high interest in the technology itself, that more technological solutions could easily solve the risk. This effect could be then strong enough so that AI students who had AIEE don't show greater risk awareness than their fellows without AIEE.

No Differences by AIEE for Opportunity Awareness of AI. The question of this study was, whether experiences from AIEE make a difference in how many opportunities future developers see in AI. The general sample showed relatively high opportunity awareness, and there was no difference between the groups of students with and without AIEE. The overall high ratings go along with the previous findings that high knowledge of AI appears together with high opportunity awareness (Center for Advanced Internet Studies, 2024). Learnings on the complexities that sociotechnical solutions come with by AIEE (Fiesler et al., 2020; Garrett et al., 2020; Hess and Fore, 2018) do not seem to go along with reevaluation of opportunities of the technology. This might indicate, that if AIEE affected the perceived opportunity of AI, this might not counter the very positive view AI students have on opportunities of AI.

No Differences by AIEE for Usage Intention of AI. Previous studies found that high knowledge of AI and training in statistics and algorithms go along with high usage intentions of AI. Separately, participating in a technology ethics course can lead to lower usage intentions of the discussed technology Cramer and Toll (2012). The sample of this study showed relatively high usage intentions of AI, higher than the general population surveyed on the same questions from the Center for Advanced Internet Studies (2024). Also, for this measure, there have not been any differences between the groups of students with and without AIEE. As observed by, Cramer and Toll (2012) the effects of ethics education might not be strong enough to overrule the overall positive usage intention. Still, AI students often select their study domain out of interest in learning and using AI (Barretto et al., 2021). Thus, they do not seem willing to step back from using the technology, when they find out about downsides of AI in AIEE, contrasting their otherwise positive usage intention.

No Differences by AIEE for Trust in AI. For trust, Lu et al. (2022) found that practitioners tend to mix up trust with other ethical principles, meaning they lack a clear understanding of the topic. Still, people with higher knowledge and familiarity with AI seem to give relatively high ratings on technical trust in AI systems (Center for Advanced Internet Studies, 2024; Gillath et al., 2021). In this study, AI students have been asked how much they trust AI, including both a technical and a social perspective within one scale. The sample indicated to have mixed, above-mid-level trust in the technology. Comparing the groups, there have been no indicators that AI students with and without AIEE differ in their trust towards AI. So, even when having higher knowledge of AI ethics, including societal risks, this does not go along with lower ratings on either technical or social aspects of trust. The level of trust seems more strongly determined by other factors, such as being a student with an interest in AI in the first place (Barretto et al., 2021). Another reason might lie within AI students' own involvement in AI development, so that not trusting AI would mean not trusting one's own domain. Still, the students referred to maintaining trust multiple times during the mini focus group discussions. Additionally, they mentioned, that they would not rely on AI systems completely, putting higher emphasis on human involvement, to account for flaws in the system. It is up for deeper qualitative investigation, whether these themes occurred more often among AIEE students than nAIEE students.

In this study, the tested attitudes toward AI have not been different between students with and without AIEE. Rather, these students do not differ in their rather positive attitude of accepting AI, and in seeing rather medium risks and high opportunities. The same holds for their intention of using AI and trusting it. These results indicate a rather positive view of AI students on AI, which is not altered by AIEE or knowledge of AI ethics, respectively. These courses often include topics like sociotechnical complexities as well as negative social, societal or environmental side effects of AI (Fiesler et al., 2020; Hess and Fore, 2018). Thus, these negative experiences have the potential to inflict a sense of algorithm aversion or reevaluation of the technology (Mahmud et al., 2022; McNamara et al., 2018). With this contradiction, it can be assumed that future AI developers show tendencies of *confirmation bias* so that they do not translate this knowledge to changes in attitude on the technology (Schiff et al., 2020), even when directly confronted. This does not directly fit the definition of automation bias or algorithm appreciation (Lyell and Coiera, 2017; Logg et al., 2019), since it is usually applied to users. Still, it might be beneficial to further investigate this tendency of what I frame *developers' algorithm affinity*. More research is especially needed to understand, in detail, why AIEE did not account for a difference in the attitudes toward AI of future AI developers in this study.

5.0.3 Complex Differences in Attitudinal Consideration of Ethical AI Principles

Generally, STEM and especially AI students have shown in previous studies to have relatively low knowledge and interest in topics of technology ethics (Harding et al., 2013; Schiff et al., 2020). But if they did, they focused in their evaluations of importance on a small set of principles such as reliability, security, and privacy (Vakkuri et al., 2019b; Lu et al., 2022; Sanderson et al., 2022). Multiple courses have shown to result in positive outcomes regarding the perceived importance of AI ethics and EAIPs (Skirpan et al., 2018; Fiesler et al., 2021; Kasinidou et al., 2021; Pierson, 2017). Still, these often focused on individual principles such as transparency, accountability, privacy, lawfulness, or fairness (Fiesler et al., 2020;

Garrett et al., 2020). Adding to this, this thesis checked for a wider set of ethical AI principles like those proposed by holistic *Ethics by Design* approaches (Prem, 2023).

Differences in Importance Ratings for Ethical AI Principles. This study's results indicate significant differences between students who received AI ethics education and those who did not in terms of their attitudes towards ethical AI principles. Specifically, students with AIEE rated ethical AI principles as more important overall compared to their counterparts without such education. This finding supports the hypothesis (H6) that AI students who took a course on technology ethics would rate ethical AI principles as more important than those who did not. The significant differences were particularly evident in principles such as *Interpretability (Explainability, Transparency, Provability)*, *Accountability*, *Beneficial AI*, *Human Agency*, and *Fairness*. The significant effects on *Interpretability (Explainability, Transparency, Provability)*, *Accountability*, and *Fairness* might be explainable by the extensive focus of many academic educators from the FAT community (Fiesler et al., 2020; Hagendorff, 2020), translating the principles to their course design. Courses like the one from Kasinidou et al. (2021) are exemplary for significant positive effects resulting from education in this domain.

While aspects of *Beneficial AI* (such as sustainability, freedom, and prosperity) received relatively low importance ratings in other studies (Khan et al., 2023; Lu et al., 2022), ethics courses typically emphasize the broader societal impacts of AI (Fiesler et al., 2020). This could lead AIEE students to rate Beneficial AI as more important. There have been no differences for the principles of *Reliability, robustness, security*, *Data Privacy*, *Lawfulness and Compliance* or *Safety*. For the first three, this goes in line with previous findings that accuracy, security, privacy, as well as law, are topics familiar already without ethical education (Sanderson et al., 2022; van Stuijvenberg et al., 2024; Vakkuri et al., 2020). This may explain their consistent importance ratings across both AIEE and nAIEE groups. The aspect of attitudes towards safety or, more ethically taken, 'non-maleficence' (Hagendorff, 2020) requires further investigation for detailed understanding. It might be due to the greater complexity under this term that especially also includes a more distanced view on AI (Schiff et al., 2020).

No Differences in Importance Rankings for Ethical AI Principles. Despite the overall higher importance ratings for ethical AI principles by AIEE students, this study found no significant differences in the ranking of these principles between AIEE and nAIEE students, rejecting H7. This suggests that while AIEE students may generally value ethical principles more, the relative priority they assign to different principles does not significantly differ from nAIEE students. This seems surprising at first. But previous studies showed that AI developers and AI students have varying perceptions of ethics (Pant et al., 2024a) and that education on the topics does not always allow for correct understanding of the taught concepts (Fiesler et al., 2021; Lu et al., 2022). This overall complexity of AI ethics may lead to unexpected rankings as students struggle to differentiate between principles. This might also be a reason why the results do not support previous findings on popularity and awareness of *Reliability, robustness, security* (Sanderson et al., 2023; Vakkuri et al., 2019a; Hadar et al., 2018; Arizon-Peretz et al., 2021) or *Data Privacy* (Lu et al., 2022; Sanderson et al., 2023), *Lawfulness and Compliance* (Vakkuri et al., 2020). A different reason might lie in the limited effectiveness of certain interventions. Kong et al. (2023a) and Borenstein et al. (2010) suggest that certain educational interventions are more effective than others. The ethics education received by AIEE students in this study may not have prioritized on individual principles, thus not leading to differences in prioritization.

Overall, the study found significant differences in importance ratings between AIEE and nAIEE students. However, these differences did not translate to rankings, suggesting that students may view principles as important without necessarily prioritizing them in practice.

5.0.4 Complex Differences in Behavioral Consideration of Ethical AI Principles

This study took three approaches to investigate behavioral consideration of AI ethics by students with and without AIEE. In the mini focus group setting, 48 students, each in groups from their condition (AIEE or nAIEE) discussed how to implement an AI system for tax fraud detection, following the development lifecycle steps by Rao et al. (2021) to allow for proper consideration of ethical principles as proposed by Prem (2023). This session aimed to simulate a development meeting on solution design as students were asked to imagine being part of the development team from the case study. First, their discussion statements referring to EAIPs have been quantified on a group level. The results were reported descriptively as well as checked for differences between the groups. Additionally, at one point in the discussion, the students were asked to individually note down the three most significant challenges to developing a good AI system. These notes have been used for investigation on an individual level.

Group Behavior Focuses on Reliability. Based on TPB, it is to be expected, that while attitudes are rather individual, behavior occurs in a stronger interplay of (perceived) social norms and (perceived) behavioral control (Ajzen, 1991). In group settings, these factors can lead to favoring certain principles over others based on perceived social pressure and beliefs about the feasibility of implementing different EAIPs. This suggests that in group settings, certain principles may be favored due to social pressures and established norms within the AI development community. The high frequency of mentions for *Reliability, robustness, security* aligns with findings from Sanderson et al. (2022) and Vakkuri et al. (2019a), who noted that accuracy and reliability are often considered as more important compared to other ethical principles by (future) developers. This technical focus may be a normative tendency, reinforced by social pressure of group settings. In this sample, nAIEE showed greater tendency to favor the principles. The difference between the groups on *Reliability, robustness, security* indicates that AIEE might alter either subjective norms or perceived behavioral control (Ajzen, 1991), so that AI students with AIEE break away from this sole focus on one principle by also considering a wider range of principles.

The low frequency of mentions for principles like *Interpretability (Explainability, Transparency, Provability)*, *Accountability* and *Lawfulness and compliance* may be due to practical challenges in implementation. Sanderson et al. (2022) found that transparency is often considered an interim target for achieving high reliability. Once reliability is achieved, transparency may no longer be prioritized. This interim nature has also been used as an argument in the mini focus groups of this study and could have led to less frequent consideration of interpretability (Table 4.5). The students may have also perceived other stakeholders as not being tech-savvy enough to benefit from detailed explanations of the discussed AI system (Vakkuri et al., 2019a). This perception can reduce the consideration of interpretability, as developers may not see the value in providing extensive explanations. *Accountability* in AI systems involves complex and often ambiguous responsibilities, which can be challenging to define and implement (Prem, 2023). The AI students might have found it difficult to assign clear accountability due to the collaborative nature of AI development and the involvement of multiple stakeholders (Vakkuri et al., 2019b). This complexity could have led to less frequent consideration of accountability in the group discussions. Bell et al. (2023) and Prybylo et al. (2024) noted that developers often struggle with implementing legal measures due to a lack of resources and support from institutions and clients. This goes along with statements of students in this study, demanding legal input from experts to properly navigate the development of AI in tax fraud detection (Table 4.5). With the stark focus on technical performance, especially these three principles have been overlooked the most in group work.

Bidirectional Effects in Group-Based Behavior Regarding Ethical AI Principles. The significant differences observed between AIEE and nAIEE groups for group-level consideration of EAIPs indicate, that in groups with students who had AIEE, EIAPs had been considered significantly more often during their work on the topic, accepting H8. As already discussed above, the focus on *Reliability, robustness, security* might be explainable by the normative focus on accuracy among AI practitioners (Sanderson et al., 2022) as it is center to general AI education in which ethics usually is treated with a secondary attached role (Knoth et al., 2024).

On the other hand, groups of AIEE students did behaviorally consider *Data Privacy* and *Beneficial AI* significantly more often than nAIEE students. Previous studies indicate that only a minority of fully educated developers are familiar with GDPR principles, and those who are, often lack the requisite knowledge about implementation techniques (Alhazmi and Arachchilage, 2021). The results of our study indicate, that ethics education might increase awareness of *Data Privacy* even more, make it more prevalent as well as has the ability to fill the knowledge gap. By this, AIEE students may feel more confident and equipped to address *Data privacy* concerns, leading to more frequent consideration of this principle, so that there can be AI considering privacy (Gröger, 2021). Additionally, Hedayati-Mehdiabadi (2022) found that relating to real-world scenarios affects future developers positively towards ethical decision-making. Ethics courses often include case studies and real-world examples (Hess and Fore, 2018; Fiesler et al., 2020), which could have enhanced the AI students' ability to recognize and consider the broader societal impacts of AI, leading to more frequent consideration of *Beneficial AI*.

Bidirectional Effects in Individual Behavior Regarding Ethical AI Principles. In the group work, students wrote down the three most significant challenges of developing a good AI system in the context of tax fraud detection. The scope of this thesis did not allow for individual tracking in the multi-person group work setting, so the discussion included this part of individual work. With this exercise, RQ1 and RQ2 were to be answered: Whether and how, in behavior, EAIPs are differently prioritized among individual students with and without AIEE. Generally, as discussed in the previous part of this section, the TPB can have a large effect in turning attitudes into behavior. The individual task in this study allows for an exploratory investigation of how attitudes on EAIPs transfer to behavior in individual work settings (Ajzen, 1991; Ellemers et al., 2019). Since all participants attitudinally rated all EAIPs relatively high there was a significant difference for the majority of EAIPs between groups, but in ranking it was unclear, it is important to understand, how this plays out in practice.

Descriptively, as for the behavioral considerations on the group level, *Reliability, robustness, security* received the most attention with 70 mentions and by this far more than other EAIPs (Table 4.5). As above, this also extends the previous research on attitudinal prioritization of AI practitioners (Sanderson et al., 2022; Vakkuri et al., 2019a) on accuracy by the behavioral dimension.

Interestingly, there have been bidirectional significant differences between the groups. Students with AIEE considered *Interpretability (Explainability, transparency, provability)*, *Fairness* significantly more often in their behavior than nAIEE students. The significantly higher considerations by AIEE students go along with these two topics being central to both AI ethics education (Fiesler et al., 2020) and of the AI ethics discussion in general (Hagendorff, 2020). Thus students AIEE may have been more regularly confronted with these issues compared to nAIEE students. This is interesting, since on a group level, these two principles did not show to differ between groups. This indicates a potential normative pressure on not raising issues of *Interpretability (Explainability, transparency, provability)* and *Fairness* in groups, not affecting individual behavior. Nevertheless, other significant differences in higher attitudinal ratings by AIEE students did not translate to individual behavior. Either normative pressure or lower perceived behavioral control might have minimized the greater consideration of EAIPs compared to students without AIEE.

At the same time, students from nAIEE did show more behavior regarding *Reliability, robustness, security* as well as more consideration of *Human Agency*. For the *Reliability, robustness, security*, the effect is consistent with behavioral patterns on a group level. Since it is repeated on the individual level, with less external normative pressure involved (Ellemers et al., 2019), the effect might be based on an underlying attitudinal prioritization (Sanderson et al., 2022). The difference between the groups indicates, that AIEE might be a relevant factor in minimizing AI students' focus on

Reliability, robustness, security, both in individual and group behavior.

Surprisingly, *Human Agency* also received higher levels of consideration by nAIEE students than by AIEE students. This could be explained by multiple discussion statements referring to *Human Agency* as a measure to increase accuracy. While the tendency of using other ethical principles as interim targets for better predictions was identified by Sanderson et al. (2022) for the principle of transparency, this thesis is able to give evidence, that this effect might also occur for *Human Agency*. If there is an interim effect for transparency among nAIEE students, it might either be smaller than the significant difference by AIEE or be prevalent for both conditions in the argument that greater transparency allows for better reliability and trust by users (Table 4.5, G4, S4; G4, S3).

5.0.5 Inconclusive Results for Attitude Behavior Gaps

Previous studies found that ethics education may lead to increased awareness and understanding of ethical principles, potentially resulting in different prioritization patterns (Skirpan et al., 2018; Fiesler et al., 2020). Students with ethics education might behaviorally prioritize those principles more than nAIEE students, as attitudinally rated (H6 in section 4.3). While there have been no differences between groups for attitudinal rankings (H7 in section 4.3), the investigation of differences on the level of ABGs can still hold interesting findings. Mäses et al. (2019) did not find a correlation between cyber-ethical attitudes and behavior among IT students in their study. This might be a hint for an ABG as also discussed by Sadeghi et al. (2022) and Bada et al. (2019) for the ethical concept of *security*. (Griffin et al., 2024) just recently published a study, qualitatively indicating the existence of the ABG among AI developers for ethical topics. Sadeghi et al. (2022) checked for effects of an educational intervention on the gap but did not find positive effects. To combine this, the quasi experimental study in this thesis investigated two aspects for RQ3 and RQ4: Whether an ABG exists among AI students and if, whether AIEE might be a factor influencing its prevalence. The researcher ran a correlational analysis Table 4.8 as well as a Mc Nemar's test Table 4.9 for possible similarities as well as differences between attitudes and behavior. The combination of both analyses allows for a bidirectional view on the ABG.

No Correlation between Attitude and Behavior. While there have been no indications for correlations between the attitudes and behavior of AI students on EAIPs, the differential analysis showed differences for eight of the nine EAIPs, with only Reliability, Robustness, Security showing no significant ABG (Table 4.9). Interestingly, these differences go either way: Three variables have been considered more often attitudinally than behaviorally. This is in line with the general understanding of the ABGs that attitudes appear more pronounced than behavior (Bada et al., 2019; Blake, 1999; Chang, 1998).

Evidence of Typical and Reversed ABG. Contrastingly, four other principles show significant differences in the other direction. These principles have been considered significantly more often behaviorally than attitudinally. The applied measure compared those three out of nine EAIPs, which have been considered by participants as most important (both attitudinally and behaviorally). When investigating changes from one principle to another on a set of nine principles (Rao et al., 2021) and only investigating three answers, it is to be expected to find changes in both directions. While limitations of this are discussed below (section 5.2), these effects give important insights: It would also mean that there exists a significant ABG for the EAIPs of *Interpretability (Explainability, Transparency, Provability)*, *Accountability*, *Lawfulness and Compliance*, and *Beneficial AI*. Even though, there also has been a significant effect on *Fairness*, the effect size is negligible. The significant differences suggest, that while AI students recognize the importance of these principles in attitudinally, they may face challenges in applying them in practice. This could be due to a lack of practical knowledge or resources to implement these principles effectively (Alhazmi and Arachchilage, 2021; Prybylo et al., 2024). The complexity and ambiguity of these principles may also contribute to the gap. For example, accountability involves complex responsibilities that are difficult to define and implement (Vakkuri et al., 2019a). Similarly, lawfulness and compliance require a deep understanding of regulatory requirements, which many developers often lack (Vakkuri et al., 2020).

The absence of a significant ABG for *Reliability, robustness, security* suggests that, taking both educational conditions together, AI students show relatively comparable attitudinal and behavioral prioritization of the principle. While there has not been a significant correlation, the p value for *Reliability, robustness, security* was closest to significance compared to the other EAIPs. This consistency may be due to the principle's popularity (Sanderson et al., 2022) especially among nAIEE is seen above in combination with the larger sample size of nAIEE in this study. It can also be seen as a reason for the omnipresence of the argument, that the focus on accuracy might influence close to every measure in this study.

For four principles, students exhibited more behavioral consideration than their attitudinal prioritization would suggest. This could be due to practical or situational factors, (Ajzen, 2014) or simply due to the choice of measure as described above.

Still, for *Privacy* this tendency suggests that while students may not prioritize *Data privacy* as highly in their attitudes, practical considerations such as compliance with regulations and the high awareness of the topic among practitioners could drive their individual behavior (Lu et al., 2022; Sanderson et al., 2023). This effect thus might have been strong enough to counter statements from the group work phase, in which participants regularly discussed whether privacy is important at all to the use case in tax investigation (Table 4.5). Still, these results contradict results from multiple studies, that, even fully educated, developers often lack the knowledge of how to consider privacy practically (Prybylo et al., 2024; Peixoto et al., 2020; Balebako et al., 2014). It might be due to progress in the field and progress in education in general, that the students have some different tendencies than their working counterparts (Fiesler et al., 2020). Further investigation is advised to understand the interplays of ABG regarding *Data privacy*.

For *Safety* the reversed ABG indicates that practical concerns about preventing harm and ensuring the safe operation of AI systems may lead to more frequent behavioral consideration of safety, even if it is not a top attitudinal priority (Vakkuri et al., 2019a). In the working phase, multiple participants mentioned that breaking with this principle would undermine taxpayers' trust in the system and thus might strongly influence whether is could be operating in the first place. This argumentative influence from the discussion might have been a major factor for *Safety* to be considered with higher priority in individual behavior than attitudinally.

The greater behavioral consideration of *Human Agency* could be due to two reasons. Firstly, it might also be based on the practical necessity of maintaining human oversight and control in AI systems. Going along with results from Seppälä et al. (2021), the participants argued that especially for a use case in tax fraud, the system should be a decision support system rather than a fully autonomous one. This argument might have been especially prevalent by the discussion compared to individual attitudinal evaluation. Secondly, the question used as a measure of individual behavioral consideration directly followed the group discussion on how to release of the model into practice. There, many students argued for human involvement as an interim target for achieving both trust and higher reliability of the system Table 4.5. Thus, the greater behavioral consideration could be explained by the fact, that the students still had this argument in mind.

Differences on Attitude Behavior Gaps by AI Ethics Education. This study also took into account the differences elicited by technology ethics education on the ABG of AI students. For this, the occurrences of ABGs have been used. In this case, only classical ABGs with a switch of prioritization away from the attitudinal focus have been investigated.

From previous studies and the TPB, two outcomes could be expected. Either the increased interest and knowledge in AI ethics (See section 4.1) shows itself in a shift in perceived social and perceived behavioral control, which both, in turn, increase chances of ethical behavior (Ajzen, 1991, 2014). This would lead to a smaller ABG for AIEE students. Contrary, the complexity of practically following ethical conduct as seen by some practitioners (Pant et al., 2024b), and the identification of practical challenges (Prybylo et al., 2024) would minimize developers' perceived behavioral control, thus not translating their high attitudes into behavior. The same effect could have occurred, as described above, regarding taking over social norms from group work towards favoring e.g. or disregarding other ethical principles (Sanderson et al., 2023). If these tendencies are stronger than the facilitating effects of ethics education, no differences in ABG might show. It is also possible to only see effects for certain ABGs: AIEE students might show a smaller gap for principles which are more often emphasized in ethics courses (Fiesler et al., 2020).

The results indicate significant differences in the ABG for the EAIPs of *Lawfulness and Compliance* and *Fairness* between AI students who took a course on technology ethics and those who did not. There were no significant differences in the ABG for other ethical AI principles, including *Reliability, robustness, security*, and *Accountability*.

This means that for *Lawfulness and Compliance* and *Fairness* there is a reduced ABG among AIEE students. This is certainly interesting for these two principles, since in the previously investigated measurement (RQ3), there was an indication of a very small reversed ABG for *Fairness*. So, while AI students of both groups together considered fairness just a little more often behaviorally than attitudinally, AIEE showed fewer ABGs than nAIEE in this measure.

5.1 Overarching Discussion

Summarizing, we can assume that AIEE students did not only have increased AI ethics knowledge and greater interest in AI ethics subsection 5.0.1, but they were also more often following up on their attitudes with aligned behavior. While ethicists would frame this as ethical virtue (Hagendorff, 2020), the TPB would explain it with an alignment of attitudes, subjective norms, and perceived behavioral control among AIEE students for these two principles. While the effect did not show for the majority of principles, it could be an indicator that AIEE students have developed a greater sense of for *Lawfulness and Compliance* and *Fairness* (Griffin et al., 2024; Hagendorff, 2020). Since nAIEE did not show to have fewer occurrences of ABGs than AIEE, it can be assumed that the ethical courage (Hess and Fore, 2018) to consider *Lawfulness and Compliance* and *Fairness* could stem from technology ethics education. While there was no difference between conditions for the ABG on the popular EAIP of *Reliability, robustness, security*, there is an overall tendency that AIEE students focus less on the principle than students without ethics education. While EAIGs have different approaches, from rather broad approaches (Dilhac et al., 2018; Floridi et al., 2018), that can be complex (Prem, 2023; Hagendorff, 2020), over really short lists like the one by Microsoft Corporation (2019), that might be too superficial (Hagendorff, 2020) to be practically relevant EAIGs (Whittlestone et al., 2019).

Even though that the guideline from Rao et al. (2021) used in this study, falls into the latter category by combining wide considerations from rather complex guidelines with a checklist manner from the shorter ones (Hagendorff, 2020), the AI students in this study still had issues clearly differentiating the principles. The (shallow) analysis of themes in the mini focus group already indicated, that multiple students saw the interconnection of topics that are considered in the EAIG. Thus, a deeper analysis of these arguments in regard to pro and contra statements might yield further insights. Still, it can be said, that the epistemic principles in Rao et al. (2021) of the study received more attention in students' behavior than the general ethical principles of the guideline. Especially *Reliability, robustness, security* drove this effect, while *Interpretability (Explainability, transparency, provability)* fell short in consideration. Still the students in the mini focus group recognized the effect it can have, to allow for *Fairness* (Table 4.5, G18, S2) or *Safety* (Table 4.5, G10, S5) going along with findings by Gilpin et al. (2022), that transparency can have an intermediate use for other principles. Contrary, this interplay seems not exclusive to the direction from epistemic to general ethical principles(Rudy-Hiller, 2018). Similarly, the mini focus group students saw the relevance of legal support and understanding to be able to navigate *Data Privacy*, thus mentioning also *Lawfulness and Compliance* as well as *Human Involvement* in the same vein (Table 4.5, G4, S4). The latter has been mentioned as a necessary factor to create a trustworthy system (Table 4.5, G4, S3). Thus, the omnidirectional direction from epistemic principles being a necessary precondition for general ethical principles to apply (Rudy-Hiller, 2018) does not hold and needs rethinking.

Overall, the study of this thesis showed that, while AI students perceive AI comparably with and without AIEE, there are differences in how they approach *Ethics by Design* (Prem, 2023; Brey and Dainow, 2023). Students with AIEE show to have more knowledge and interest in AI ethics, giving greater importance to EAIPs overall. They might even have developed some ethical courage (Hess and Fore, 2018) to practically consider a wider range of EAIPs and, in some cases, to bridge the ABG for these principles more often than their fellow students without ethics education. Still, there are also larger gaps for students who took AIEE.

5.2 Limitations

This study faces several limitations that should be considered when interpreting the results, building on the discussion, and planning future research.

Limitations in the Study Design Overall, this study compared two groups of students in a quasi-experimental study design. There has not been a direct experimental manipulation but a comparison of true independent groups comparable over the variable of AIEE. Thus results of this study might imply that differences could come from AIEE, but due to the amount of possible additional factors, causality cannot be guaranteed. Rather, this study resulted in correlational insights. It is advised for future projects to apply these measures at the beginning and end of a semester to experimentally verify differences that can be verifiably related to AIEE.

Limitations in the Sample The relatively small sample size limits the generalizability of findings to the broader AI student population. Especially, the small number of groups and participants for the behavioral analysis did affect the choice of statistical methods, which meant losing important interpretable results. Additionally, effects might be overestimated due to the small sample. The diversity in backgrounds and experiences across different studies further complicates this issue (Schiff et al., 2020; Harding et al., 2013). Still, it allows for greater external validity since future developers come from different backgrounds with different variations in curriculum (Fiesler et al., 2020; Shih et al., 2021) instead of only one monopolized educational program.

Additionally, students' self-selection into majors and ethics courses may have significantly influenced their current attitudes and behaviors, potentially skewing the results (Kreth et al., 2022). Especially, a large proportion of students (see Table 3.1 and Table 3.3) indicated to study either Data Analytics and Decision Science (DADS) or Computational Social Systems (CSS) at RWTH Aachen University. While DADS focuses on a technical and business perspective of data science, CSS considers ethics and psychology (RWTH Aachen University, 2024a,b). The curriculum shows interwoven ethical courses throughout the length of the studies. This is both an example of integrated ethics education, and it might be affected by a self-selection bias of students in the program. To protect participants' privacy, the researcher decided to combine studies into categories within the demographic questions. Thus, this effect cannot be clearly differentiated from the studies in DADS. Further, while the first intention of the project was to solely consider one compulsory ethics course (n = 32), due to time constraints in recruitment, other ethics courses have also been included in the study (n = 10). Future studies should consider the potential individual effects of educational background when designing their demographic questions and draw clear lines to account for educational variability.

Limitations in Methods. Methodologically, the reliance on self-reported data, as well as the group setting, introduces a social desirability bias (Randall and Fernandes, 1991), which can affect the accuracy of the reported attitudes and ABG. While the researcher intended to control these effects and include them in the analysis of group-level behavior, this bias is a common limitation in behavioral ethics research and cannot be ruled out completely. Thus, it is advisable for future research to include measures checking for the prevalence of social desirability effects. Additionally, the mode of participation (in person, online) might have induced effects. Still, in the frame of this thesis, these two modes became necessary to recruit enough participants, especially for the AIEE condition, as a small section of the overall AI student population.

Limitations in Measures. Firstly, one limitation concerns the measure of attitudes towards AI. Since previous research often investigated users of AI regarding their knowledge or familiarity with AI, this study used these established measurements to investigate attitudes on AI. Considering that AI students in their lives are users of AI and developers, it might be possible that they read and answered these questions on attitudes towards AI from a user's perspective. If true, the results of this part of the study rather reflect the perspective of AI students as users than their professional perspective. Thus, the two parts of the study (AI in general and EAIPs specifically) would have measured different perspectives on AI. It is up to future research to investigate this possibility, as well as to verify that the measures used only allow for one interpretation if a person can take over multiple roles within the same context.

Additionally, there are three limitations concerning the measurements on the ABG. Firstly, the behavioral measure of this study might rather measure intention than strictly measure behavior. This is a common limitation in behavioral ethics research (Sadeghi et al., 2022; Gino et al., 2009). The researcher attempted to balance between a study design that was easy to administer and analyze while gaining valid insights into behavior and the ABG, which are harder to operationalize. In addition, the binary approach to measuring the ABG may have oversimplified the nuanced interactions between attitudes and behaviors (Blake, 1999; Ajzen, 1991). Further, differences in the framing of questions between the attitudinal ’principle importance’ and the behavioral ’challenge importance’ may have affected the validity of comparisons. While there is a nuanced difference, asking for the most important challenges in the working was one way to acquire insights into practical consideration of principles, without asking for these principles. Thus, this study may not fully capture the complexity of ethical attitudes and ethical behavior. Further research should develop and validate measures that operationalize the ABG practically.

Overall, this thesis aimed to investigate many aspects of ethical AI development under AI ethics education within one research design. While this allowed for a broad overview on multiple topics, from general algorithm appreciation over attitude and behavior to the gap between both regarding a large set of ethical principles within an environment to simulate a working environment, the internal validity of the study suffered under these ambitions. Still, future work can possibly build on the results of this study.

5.3 Future Work

Future research has multiple directions to investigate. First, the results of this study offer more aspects to be analyzed, s.a. the technical arguments used by the participants on how to develop AI in tax fraud detection or a more in-depth investigation of studies regarding the measures. Additionally, the qualitative data might provide further insights into how, in detail, future developers approach the individual EAIPs, their interplays and how they view the EAIPs to be related with the concept of trust, being mentioned multiple times during the mini focus groups

Beyond this study itself, even more aspects are crucial to investigate. First, it is worthwhile to explore a sample that goes beyond a single university and takes a national or international view on the topic. Contrarily, future samples could also be dedicated to individual studies to find out whether the effects in this study are solely based on one ethics course or whether they rather stem from a large effect coming from an individual study program. In this study, the study program of *Computational Social Systems* has shown to take up a large section of participants from the AIEE condition. It could be worthwhile investigating differences between this subsample alone and a sample from other studies in a deeper manner to validate the program and understanding of the topic.

The topic also asks for more analyses of curriculum design and pedagogies (Fiesler et al., 2020; Hess and Fore, 2018). This study did not give a result on especially good or bad pedagogical approaches that lead to differences between AI students with and without AIEE. Further studies could include comparative designs of different pedagogical approaches, such as case studies, hands-on projects, or interdisciplinary collaborations, to determine which strategies are most effective in eliciting changes or even for bridging the attitude-behavior gap.

Also, it is important for future research to take a more in-depth look into the intersection of AI expertise and ethical decision-making and courage (Hess and Fore, 2018). Especially for AI ethics, validated assessment tools are still missing, but they yield great benefits for the research community. This includes measures for AI ethics literacy and ethical AI decision-making. Additionally, ethical AI courage might be harder to operationalize, but insights might be even more relevant to follow up on demands from EAIGs (Prem, 2023; Hagendorff, 2020) when their bare existence has shown not to be effective for ethical behavior (McNamara et al., 2018).

Further, the actual implementation surrounds AI developers. While this study showed differences in AIEE for AI students, it may be relevant to understand the current conditions for teams that current students will soon be a part of. These new social contexts might come with new social norms (Ajzen, 2014) and thus different possible outcomes for the ABG.

Additionally, it is important to consider other scenarios aside from AI in tax fraud investigations. Different contexts of AI have been shown to generate different responses for ethical measures (Reimenschneider et al., 2011). Thus, AI students' attitudes towards AI, attitudes and behavior regarding EAIPs might be severely different for other application domains. This, in fact, might be in a close interplay with the educational background of students focusing on a certain domain or being influenced by the culture at their university.

Taken together, this study opens up numerous directions for future research in AI ethics education. These future research directions will contribute to a more comprehensive understanding of how to effectively instill ethical considerations in AI education and practice, ultimately leading to more responsible AI development and deployment.

5.4 Conclusion

The development of AI systems has significant ethical implications, with the potential for both beneficial and harmful impacts on society (Hagendorff, 2020; Prem, 2023). Despite the existence of numerous ethical AI guidelines and principles, AI students and developers often lack high levels of ethical literacy from their education, to consider them in their work (Schiff et al., 2020; Harding et al., 2013; Sanderson et al., 2022). Simultaneously, some students are receiving education on technology ethics (Fiesler et al., 2020). Griffin et al. (2024) even framed this as a growing amount of ethical wisdom in the community. Still, there is little known about the differences between future AI developer with and without ethics education in their general attitudes toward AI and their attitudes and behavior regarding ethical AI.

To fill this gap, 98 AI students participated in a two-part study. While the whole sample answered a questionnaire on their perception of AI and EAIPs, a subsample worked on how to approach an AI case study within a focus group. Students either previously had taken courses on technology ethics or not.

The results of this quasi-experimental study indicate that AI students have positive attitudes toward *whether* to apply AI systems, not differing between students who underwent AIEE and those who did not. Still, students who did have AIEE, approach questions of the *how* in AI development differently to those who did not undergo AIEE. Also, they show to have greater knowledge and interest in AI ethics. This shows itself differently in their attitudes and behavior regarding the nine investigated EAIPs. While attitudinally, students with AIEE view EAIPs with greater importance, this is different for behavior.

Overall, the AI students showed great consideration of the principle of *Reliability, Robustness, Security*, going along with other studies (Sanderson et al., 2022), which indicates that there is a focus of AI students on questions around reliable results of AI systems (Sanderson et al., 2022). Students without ethics education even show significantly greater behavioral consideration of *Reliability, Robustness, Security* in AI. This effect occurred both in group work and individual behavior. In contrary, students who had AIEE included a wider range of principles in their behavior than those without AIEE. Generally, this study indicated the prevalence of complex ABGs among AI students: While typical tendencies of ABGs were prevalent for e.g., *Interpretability (Explainability, transparency, provability) Accountability*, other EAIPs have been considered fairly less attitudinally, than behaviorally. AI students showed fewer ABGs for *Lawfulness and compliance* as well as *Fairness*. While there are limitations on the investigation of the ABGs due to selection of measurements, and sample size, the general tendency could be an indicator of the benefits of AI ethics education on how AI students approach ethical AI development.

In conclusion, this study suggests that AIEE may be an important factor in better incorporating EAIPs into practice, but the complex interactions of EAIPs within guidelines are difficult to address with AIEE alone. This is particularly worrying as the guideline for this study (Rao et al., 2021) came from a section of approaches explicitly designed to support ethical AI development in practice. However, this thesis only provides a glimpse into a more holistic investigation of the impact of ethics education on the development of ethical AI and also points to gaps that still need to be filled before we can move even closer to ethical AI.

Appendix A

Recruitment Material

This appendix includes the recruitment material for the study. Those include a flyer for the mini focus group study, a flyer for the questionnaire study and a web post and image for the study in general. All links and QR codes lead to the sign up procedure of the study in Appendix B.

AI STUDENT?

SOLVE A COMPLEX CASE STUDY TOGETHER WITH YOUR PEERS

EARN 20€ AS A THANK YOU FOR PARTICIPATION

+ a chance to win prices worth 200€

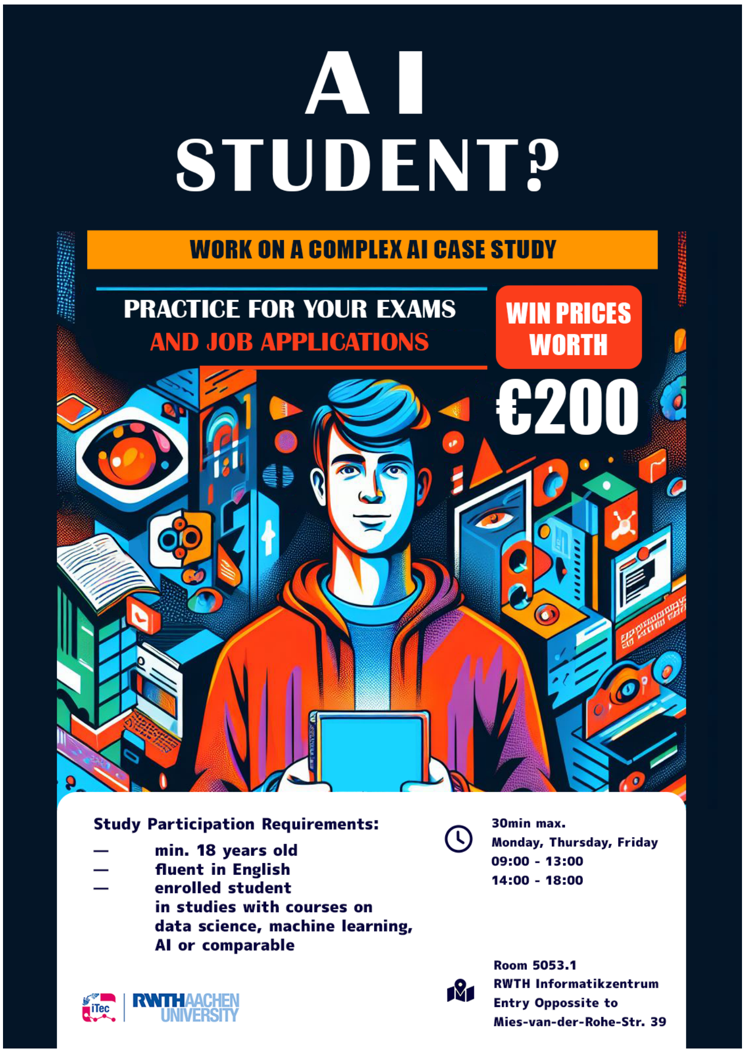
Study Participation Requirements:

- min. 18 years old
- fluent in English
- enrolled student in studies with courses on data science, machine learning, AI or comparable

RWTH Informatikzentrum
Entry Opposite to
Mies-van-der-Rohe-Str. 39

soscisurvey.de/perceptionofai/?q=signup

Figure A.1: Mini Focus Group Flyer



AI STUDENT?

WORK ON A COMPLEX AI CASE STUDY

PRACTICE FOR YOUR EXAMS
AND JOB APPLICATIONS

WIN PRICES WORTH
€200

Study Participation Requirements:

— min. 18 years old
— fluent in English
— enrolled student
in studies with courses on
data science, machine learning,
AI or comparable

30min max.
Monday, Thursday, Friday
09:00 - 13:00
14:00 - 18:00

Room 5053.1
RWTH Informatikzentrum
Entry Opposite to
Mies-van-der-Rohe-Str. 39

Figure A.2: Questionnaire Flyer

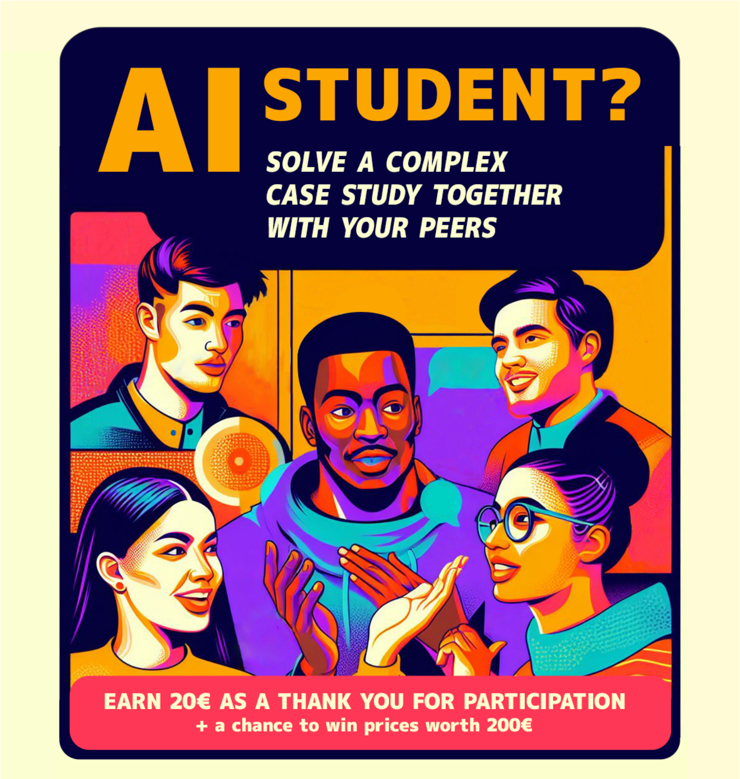

AI STUDENT?
SOLVE A COMPLEX CASE STUDY TOGETHER WITH YOUR PEERS
EARN 20€ AS A THANK YOU FOR PARTICIPATION
+ a chance to win prices worth 200€

Figure A.3: Online Recruitment Image

Online Recruitment Text. In this study, you'll work on a complex AI Case Study, to practice your case study skills together with fellow students - both for your upcoming exams and for future job applications. You'll learn about an industry standard AI Design Workflow to work on the matter. Additionally you answer questions on your views on AI.
For compensation, you receive either 20€ cash or 2 Participation hours for your studies in informatics at RWTH University. Additionally, you have the chance to win 200€ in cash prizes with extra chances to win if you refer the study to friends.
Sign-up via www.soscisurvey.de/perceptionofai/?q=signup or using the QR code on the flyer. Students with less time or outside Aachen can participate in a shorter online version via https://www.soscisurvey.de/perceptionofai/?q=ga01.

Appendix B

Signup

B.1 Signup-Procedure

This appendix shows the SoSci Survey (Leiner, 2024) web pages for selecting how to participate in the study. The participants could either sign up for the in person mini focus group, the online mini focus group, the in person questionnaire or the online questionnaire. The procedure for this and for splitting up the participants in AIEE and nAIEE groups can be viewed in the first section. The second section of this appendix includes the signup-page from meetergo.com (meetergo, 2024) that is embedded within the signup process or was to be entered via an individual link above.

SU17

Sign-Up for Study on a Complex AI Case Study

While the public is discussing AI without much knowledge, you are on your way to potentially joining the AI development work force.
In this study, you will work on a complex AI Case Study, to practice your Case Study skills and give important insights to research on AI.

You must fulfill the following requirements:

- 18+ years old
- enrolled as a student with courses in data science, machine learning or other AI topics.
- proficient in English (B2 or above)

The study is part of a research project at the Chair Individual and Technology at RWTH supervised by Prof. Rosenthal-von der Pütten.

Previously, this study was run in person, so I was able to give out compensation in cash. Within the online version, you enter the lottery on three money-prizes worth 200€.

SU05

1. There are three different versions to take part.

How do you want to participate?

I like to participate in

○

- **a focus group study with fellow students** on a complex AI case study based on an industry AI development lifecycle model
- over **90min**
- at **Informatikzentrum RWTH Aachen University**

  As a thank you, I'll receive
  - **20€** cash OR **2 participation hours**
  - an entry in the **lottery** on cash prizes worth 200€

○

- **a focus group study with fellow students** on a complex AI case study based on an industry AI development lifecycle model
- over **90min**
- **Online via zoom in Central-European-Time**

  As a thank you, I'll receive
  - **20€** cash OR **2 participation hours**
  - an entry in the **lottery** on cash prizes worth 200€

○

- **an individual questionnaire study** on a complex AI case study
- over **45min**
- at **Informatikzentrum RWTH Aachen University**

  As a thank you, I'll receive
  - **10€** cash OR **1 participation hour**
  - an entry in the **lottery** on cash prizes worth 200€

○

- **a questionnaire study** on a complex AI case study
- over **30min**
- **Online**

  As a thank you, I'll receive
  - an entry in the **lottery** on cash prizes worth 200€
  - **0.5 participation hours if needed**

SU05 What study?

1 = a focus group study with fellow students on a complex AI case study based on an industry AI development lifecycle model over 90min at Informatikzentrum RWTH Aachen University As a thank you, I'll receive 20€ cash OR 2 participation hours an entry in the lottery on cash prizes worth 200€
4 = a focus group study with fellow students on a complex AI case study based on an industry AI development lifecycle model over 90minOnline via zoom in Central-European-Time As a thank you, I'll receive 20€ cash OR 2 participation hours an entry in the lottery on cash prizes worth 200€
2 = an individual questionnaire study on a complex AI case study over 45min at Informatikzentrum RWTH Aachen University As a thank you, I'll receive 10€ cash OR 1 participation hour an entry in the lottery on cash prizes worth 200€
3 = a questionnaire study on a complex AI case study over 30min Online As a thank you, I'll receive an entry in the lottery on cash prizes worth 200€0.5 participation hours if needed
-9 = Not answered

6 Active Filter(s)

Filter SU05/F1
If any of the following options is selected: 1
Then display the questionnaire page(s) whatcourses-fgsignup (otherwise hide them)

Filter SU05/F2
If any of the following options is selected: 2, 3
Then hide the questionnaire page(s) whatcourses-fgsignup (otherwise display them)

Filter SU05/F3
If any of the following options is selected: 2
Then display the questionnaire page(s) gaoffline (otherwise hide them)

Filter SU05/F4
If any of the following options is selected: 1, 3
Then hide the questionnaire page(s) gaoffline (otherwise display them)

Filter SU05/F5
If any of the following options is selected: 3
Then display the questionnaire page(s) gaonline (otherwise hide them)

Filter SU05/F6
If any of the following options is selected: 1, 2
Then hide the questionnaire page(s) gaonline (otherwise display them)

2. What courses have you finished so far?
Select the one's that are closest to what you visited so far.

We will order you into groups based on your broad previous knowledge.

- □ **Actions and Planning in AI**
- □ **Advanced Methods in Automatic Speech Recognition**
- □ **Algorithmic Foundations of Datascience**
- □ **Algorithmische Lerntheorie**
- □ **Algorithms for Politics**
- □ **Applied Data Analysis**
- □ **Automatische Spracherkennung**
- □ **Business Process Intelligence**
- □ **Combinatorial Optimization**
- □ **Computer Vision 1 / 2**
- □ **Data Analysis and Visualization**
- □ **Data Driven Medicine**
- □ **Einführung in Algorithmisches Differenzieren**
- □ **Einführung in die Ethik (Technikphilosophie) / Introduction to the Philosophy of Science and Technology**
- □ **Ethics, Technology, and Data / Ethics of Artificial Intelligence and Robotics**
- □ **Explainable AI and Applications**
- □ **Exploratory Data Analysis**
- □ **Fundamentals of Business Process Management**
- □ **High-dimensional Probability for Mathematicians and Data Scientists**
- □ **Introduction to Data Science**
- □ **Konzepte und Modelle der parallelen und datenzentrischen Programmierung**
- □ **Künstliche Intelligenz**
- □ **Machine Learning**
- □ **Mathematical Foundations of Machine Learning**
- □ **Mathematical Methods of Signal and Image Processing**

□ **Mathematics of Data Science**

□ **Moral Reasoning / Ethics of Economics**

□ **Nonlinear Optimization**

□ **Privacy and Big Data**

□ **Privacy Enhancing Technologies for Data Science**

□ **Probabilistic Programming**

□ **Project: Analytics and Optimization**

□ **Reinforcement Learning and Learning-based Control**

□ **Semantic Web**

□ **Social and Technological Change**

□ **Statistische Klassifikation und Maschinelles Lernen**

□ **Statistische Methoden zur Verarbeitung natürlicher Sprache**

□ **Others and I still have some background in the field of AI**

SU02 What Courses: Residual option (negative) or number of selected options

Integer

SU02_37 Actions and Planning in AI

SU02_01 Advanced Methods in Automatic Speech Recognition

SU02_02 Algorithmic Foundations of Datascience

SU02_03 Algorithmische Lerntheorie

SU02_04 Algorithms for Politics

SU02_05 Applied Data Analysis

SU02_06 Automatische Spracherkennung

SU02_31 Business Process Intelligence

SU02_07 Combinatorial Optimization

SU02_38 Computer Vision 1 / 2

SU02_08 Data Analysis and Visualization

SU02_09 Data Driven Medicine

SU02_10 Einführung in Algorithmisches Differenzieren

SU02_34 Einführung in die Ethik (Technikphilosophie) / Introduction to the Philosophy of Science and Technology

SU02_11 Ethics, Technology, and Data / Ethics of Artificial Intelligence and Robotics

SU02_36 Explainable AI and Applications

SU02_12 Exploratory Data Analysis

SU02_13 Fundamentals of Business Process Management

SU02_14 High-dimensional Probability for Mathematicians and Data Scientists

SU02_15 Introduction to Data Science

SU02_16 Konzepte und Modelle der parallelen und datenzentrischen Programmierung

SU02_17 Künstliche Intelligenz

SU02_18 Machine Learning

SU02_19 Mathematical Foundations of Machine Learning

SU02_20 Mathematical Methods of Signal and Image Processing

SU02_21 Mathematics of Data Science

SU02_33 Moral Reasoning / Ethics of Economics

SU02_22 Nonlinear Optimization

SU02_35 Privacy and Big Data

SU02_23 Privacy Enhancing Technologies for Data Science

SU02_24 Probabilistic Programming

SU02_25 Project: Analytics and Optimization

SU02_26 Reinforcement Learning and Learning-based Control

SU02_27 Semantic Web

SU02_28 Social and Technological Change

SU02_29 Statistische Klassifikation und Maschinelles Lernen

SU02_30 Statistische Methoden zur Verarbeitung natürlicher Sprache

SU02_32 Others and I still have some background in the field of AI

1 = Not checked
2 = Checked

SU02_32a Others and I still have some background in the field of AI (free text)

Free text

5 Active Filter(s)

Filter SU02/F1
If any of the following options is selected: 11, 23, 28, 33, 34, 35, 36
Then display question/text SU03 placed later in the questionnaire (otherwise hide)

Filter SU02/F2
If any of the following options is selected: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29, 30, 31, 32, 37, 38
Then display question/text SU04 placed later in the questionnaire (otherwise hide)

Filter SU02/F3
If any of the following options is selected: 11, 23, 28, 33, 34, 35, 36
Then display question/text SU03 placed later in the questionnaire (otherwise hide)

Filter SU02/F4
If any of the following options is selected: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29, 30, 31, 32, 37, 38
Then display question/text SU04 placed later in the questionnaire (otherwise hide)

Filter SU02/F5
If any of the following options is selected: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38
Then hide question/text SU12 placed later in the questionnaire

Page 03

fgsignup

SU03

Thanks for coming so far!

Please sign up for sessions that have other participants already, to create full groups. Thank you!

-- Please wait for the calendar integration to load. If it does not work, please go to https://my.meetergo.com/itec-study/aiep-eail-fg --

If you refer this link to friends, be reminded that they should also select which courses they took already, for the study-relevant balance of group members.
Thanks!

SU04

Thanks for coming so far!

Please sign up for sessions that have other participants already, to create full groups. Thank you!

-- Please wait for the calendar integration to load. If it does not work, please go to https://my.meetergo.com/itec-study/aiep-neail-fg --

More timeslots will be opened up in the upcoming days. Sign up for the waitlist: https://forms.office.com/e/NC7hP29TAf

If you refer this link to friends, be reminded that they should also select which courses they took already, for the study-relevant balance of group members.
Thanks!

You can close this tab after completing this page. Thanks!

Page 04

gaoffline

SU07

Drop by spontaneously or book a slot below!

If you come around spontaneously, please keep these slots in mind. Sometimes other organizations block the room with their events.

-- Please wait for the calendar integration to load. If it does not work, please go to https://my.meetergo.com/itec-study/aiep-ga--

On the next page you find the location of your session.
Looking forward!

You can close this tab after completing this page. Thanks!

Page 05

gaonline

SU08

Okay, nice.

You selected to take the online questionnaire.

***[Start right here](#)*.**

You can close this tab after using the link. Thanks!

Last Page

Thank you for supporting science!

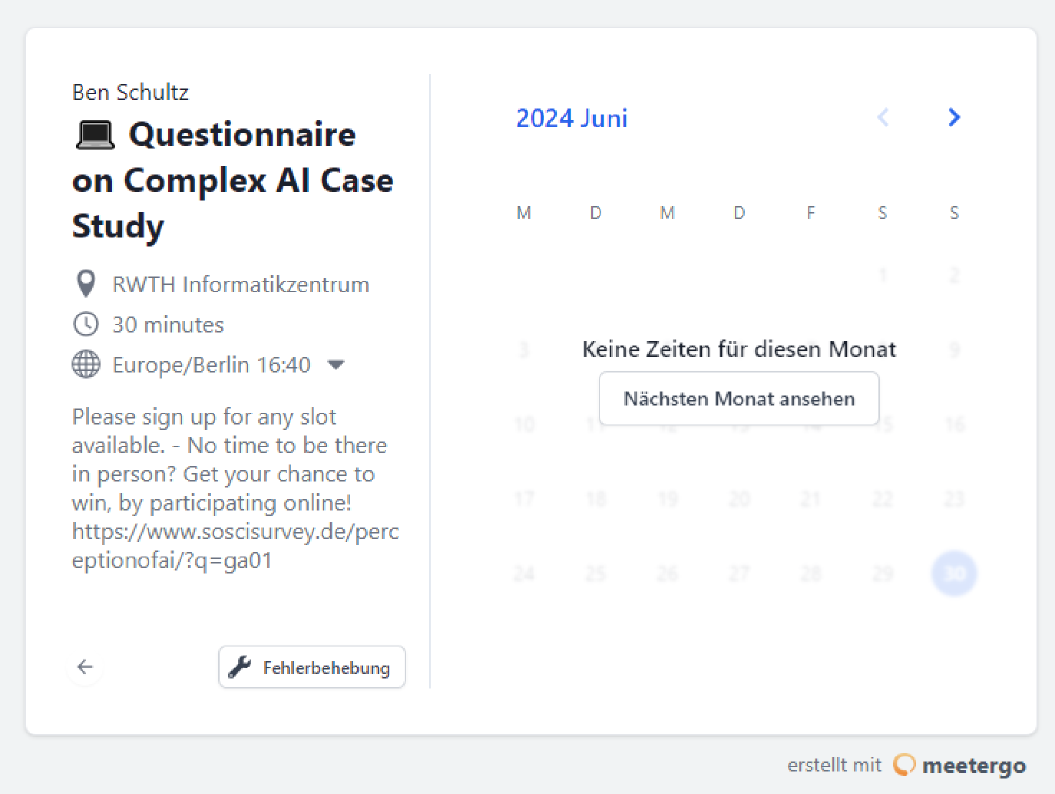
Take care

B.Sc. Ben Schultz, RWTH Aachen – 2023

B.2 meetergo.com signup page

This appendix includes the focus group signup page in the meetergo (2024) online tool for both the questionnaire study alone as well as the mini focus group study, combining both questionnaire and mini focus group participation.

Signup for a slot to take the questionnaire study



Ben Schultz

💻 Questionnaire on Complex AI Case Study

RWTH Informatikzentrum

30 minutes

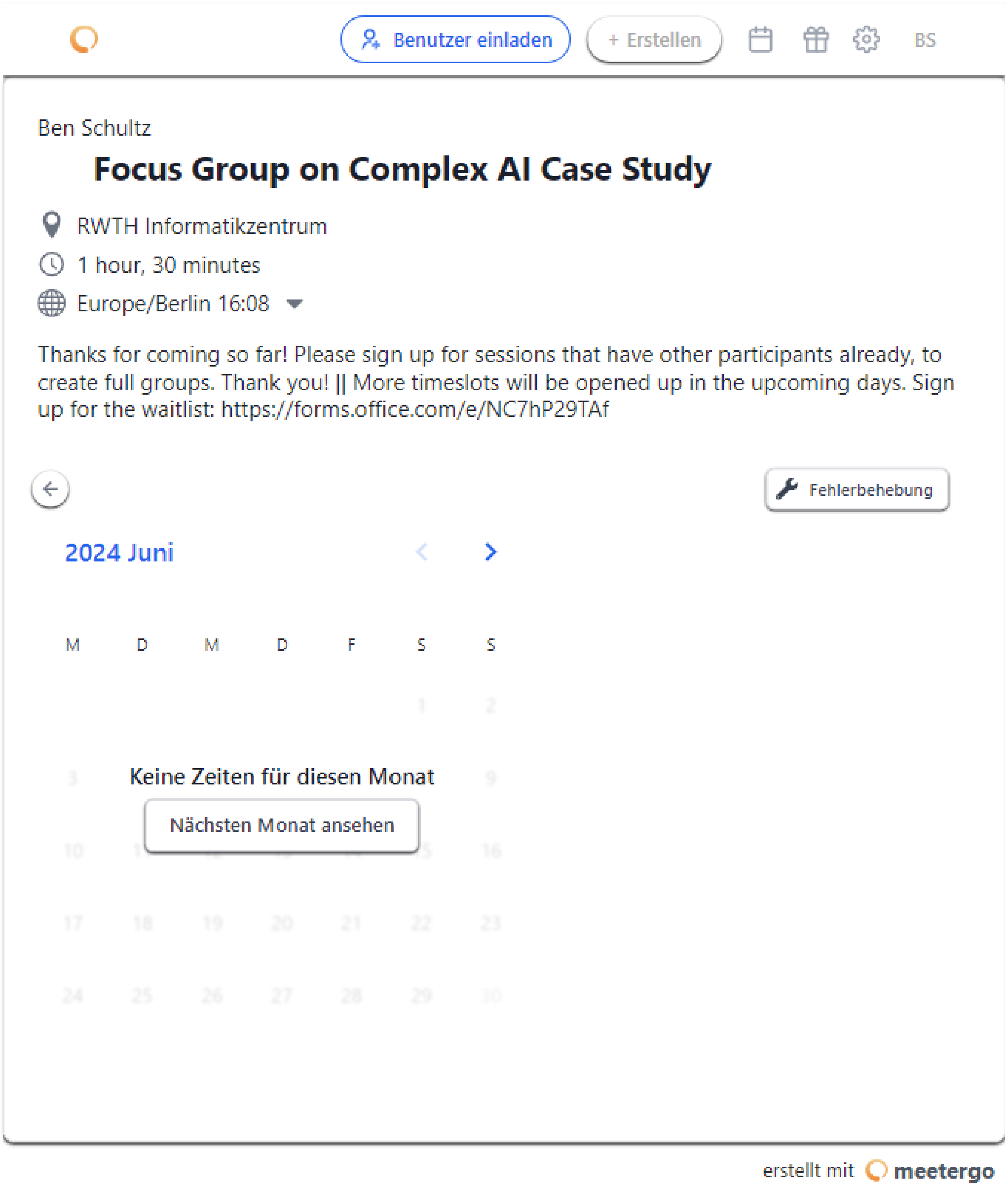
Europe/Berlin 16:40

Please sign up for any slot available. - No time to be there in person? Get your chance to win, by participating online! https://www.soscisurvey.de/perceptionofai/?q=ga01

2024 Juni

Keine Zeiten für diesen Monat

Signup for a slot to take the focus group study



Ben Schultz

Focus Group on Complex AI Case Study

RWTH Informatikzentrum

1 hour, 30 minutes

Europe/Berlin 16:08

Thanks for coming so far! Please sign up for sessions that have other participants already, to create full groups. Thank you! || More timeslots will be opened up in the upcoming days. Sign up for the waitlist: https://forms.office.com/e/NC7hP29TAf

2024 Juni

Keine Zeiten für diesen Monat

Appendix C

Email Communication with Participants

This appendix includes meetergo scripts used to automatically send an email to future participants after signup and as a reminder one day before the mini focus group.

C.1 After SignUp

Title:

Confirmed: !M.Meeting.Name on !M.Meeting.StartDate

Text:

Hi !M.Guest.Fullname ,

Your !M.Meeting.Name on !M.Meeting.StartDate at !M.Meeting.StartTime o clock has been confirmed.

Full groups are crucial for the success of our study. If you need to reschedule, please let me know well in advance. Your cooperation is greatly appreciated, and we hope to see you at the event as scheduled.

I'm looking forward to seeing you. Best Regards Ben Schultz +49 15771300374

Reschedule: !M.Meeting.Reschedule

Cancel: !M.Meeting.Cancel

C.2 Reminder Email

Title:

Reminder: !M.Meeting.Name at !M.Meeting.StartTime on !M.Meeting.StartDate

Text:

Hey !M.Guest.Fullname ,

this is a friendly reminder for your !M.Meeting.Name on !M.Meeting.StartDate at !M.Meeting.StartTime .

The session will take place at the RWTH Informatikzentrum in Seminar-Room b-it 5053.1. Enter the Informatikzentrum opposite to Mies-van-der-Rohe-Str. 39. Go left in the entrance hall. Take the second glass door on the right (opposite to lecture hall AH06.) to the learning facility of b-it 5053. Within, you find the room in the left corner. If you do not find the location, please do not hesitate to call me. I'm looking forward to seeing you. Best Regards Ben Schultz +49 15771300374

!M.Meeting.Link

!M.Meeting.Description

!M.Guest.QuestionsAndAnswers

Appendix D

Questionnaire

This appendix includes the full version of the questionnaire in variable view for participants that, in person, participated both in the mini focus group and answered the questions afterwards on the PC. Thus, this is the longest version of the questionnaire, including explanation of both mini focus group and questionnaire procedure. This version also includes variable IDs and code for automatic adjustments based on previous answers. The inclusion of these influences the layout within the appendix, but allows for replication of the study. Both a version as shown to participants (Gallery view) as well as the other questionnaire versions (Mini focus group + questionnaire online; questionnaire in person, questionnaire online) are uploaded to the OSF. Also, the legal documents for participants can be found uploaded to the OSF.

A001

Study: Complex AI Case Study

We are thrilled that you are willing to participate in an exciting research project focused on understanding future developers' perspectives on emerging technologies and their applications, focusing on artificial intelligence (AI).

We are investigating the opinions and decisions of future AI system developers. As a student in computer science, data science, or comparable studies, you will be in the position of shaping the technological landscape. Thereby, your insights are valuable to our understanding of the future landscape of artificial intelligence.

Your answers cannot be right or wrong. It is about your perceptions and opinions. If you are not completely confident in your answers, there is no harm in that. Most questions can and should be answered "from the gut".

Please read all the information thoroughly.
If you delete the cache of the study during the questionnaire, your intermediate progress will be lost.

Procedure of This Study: Complex AI Case Study

In this study, we aim to explore your experiences, knowledge, and opinions regarding the exciting world of emerging technologies.

We are particularly interested in understanding how you'd personally proceed to **tackle a development case** within a **focus group setting**. You'll learn about an **industry-standard development cycle for AI** to solve the a use case within the group. Additionally, you'll go through a questionnaire on your interest in AI, its potential applications in different fields, and your attitude to the advantages and disadvantages of these technologies.

The session is split into two sections:

1. Focus Group (approx. 60 min)
    1. *case study presentation*
    2. *input on a industry-standard development lifecycle*
    3. *discussion*
    4. *individual summary of discussion results*

--- if needed: Short Break---

2. Questionnaires (approx. 30 min)

Including buffers and break, the study takes max. 90 minutes to finish.

Compensation

After participation in this study, you receive a ticket for **a lottery for 100€/50€/50€**. Additionally, you can refer the study to colleagues and friends to **increase your chances of winning**. More information on the lottery at the end of the study.

Additionally, you can choose between two compensations for a thank you to being here in person:

- **20€ in cash**
- **2 participation hours** for your studies

Terms of participation

- Prerequisites for participation are that you are at least 18 years old, have a minimum level B2 proficiency in English (independent use of language) and are currently enrolled as a student in a programme or degree that includes courses on data science, machine learning or artificial intelligence (AI) in general.
- Participation in this study will take approx. 90 minutes.
- No personal data will be collected. Your data will be stored completely anonymously and used exclusively for scientific purposes.
- For anonymization, your recordings, and answers are stored under a random code word that you yourself will create based on a certain rule. It results in an individual code word that no one but you can know. This means that it is not possible for anyone to associate your data with your name.
  You can subsequently use this code to request the deletion of your data.
- The retention period for the fully anonymized data is at least 10 years after data evaluation, or at least 10 years after the publication of a paper on this study.
- Participation in the study is voluntary. You may discontinue participation in this study at any time, without giving any reason, and without any disadvantage.
- A computer or tablet should be used for optimal display. If you are using a smartphone, please use it in landscape format.
- You are only entitled to compensation as long as you complete the study honestly.

You can download a detailed version of the conditions of participation as a PDF file here:
Download file
If you have any questions, please contact
itec@humtec.rwth-aachen.de

Declaration of Consent

A006

As this survey is conducted online and the data is collected anonymously, your consent to participate is not confirmed with a signature. Rather, the selection below constitutes your consent.

(If you agree to participate, you will begin the study after clicking the “Continue” button.)

○ I declare that I have been fully informed about the terms of participation and that I understand that my participation in the study is voluntary.
I also understand that I can withdraw from participation at any time without any disadvantages and that I can have the data subsequently deleted with reference to the random code assigned to me.

○ No, I do not agree and do not wish to take part in the survey.

A006 Declaration of Consent QU

1 = I declare that I have been fully informed about the terms of participation and that I understand that my participation in the study is voluntary. I also understand that I can withdraw from participation at any time without any disadvantages and that I can have the data subsequently deleted with reference to the random code assigned to me.
2 = No, I do not agree and do not wish to take part in the survey.
-9 = Not answered

2 Active Filter(s)

Filter A006/F1
If any of the following options is selected: **2**
Then display the text **Z910** and finish the interview, after the next button was clicked

Filter A006/F2
If any of the following options is selected: **1**
Then hide the questionnaire page(s) **NoAgr** (otherwise display them)

Page 03
NoAgr

Z910

Thank you for considering taking part. We understand your decision, although we would like to express our regret:

Should you change your mind at a later date or if you have misclicked, you are welcome to participate in the study later on.

Still, I'd be happy if you refer this study to others, that match the criteria.
Thank you very much!

Page 04
INCL

A013

1. To participate in this study, you must fulfill all four criteria.

□ I am older than 18 years.

□ I speak English fluently on a level similar or above B2.

□ I am currently enrolled/have been enrolled lately as a student in a program or degree that includes courses on data science, machine learning or artificial intelligence (AI) in general.

□ I did not take part in the other version of this study before (Focus Group/Online Questionnaire).

A013 Inclusion Criteria: Residual option (negative) or number of selected options
Integer
A013_01 I am older than 18 years.
A013_02 I speak English fluently on a level similar or above B2.
A013_03 I am currently enrolled/have been enrolled lately as a student in a program or degree that includes courses on data science, machine learning or artificial intelligence (AI) in general.
A013_04 I did not take part in the other version of this study before (Focus Group/Online Questionnaire).
1 = Not checked
2 = Checked

1 Active Filter(s)

Filter A013/F1
[inactive] No condition selected
Then display the text **Z905** and finish the interview, after the next button was clicked

Page 05
RC

PHP code

```
put('A009_01', random(10, 99));
put('A009_02', random(10, 99));
put('A009_03', random(10, 99));
put('A009_04', random(10, 99));


$code1 = value('A009_01', 'free');
$code2 = value('A009_02', 'free');
$code3 = value('A009_03', 'free');
$code4 = value('A009_04', 'free');

html('<p><b>Anonymous Random Code</b><br><br><br>You are assigned a random code to anonymously map your n
<br><br>Additionally, you can use this code to delete your data in hindsight. You can request the deletio
<b><br><br> Your code is: '.$code1.'-'.$code2.'-'.$code3.'-'.$code4.' </b></p>');
```

A004

2. Introduction to the study:

AI in Tax Fraud Detection Systems

Efficient and equitable taxation is essential to ensure the proper utilization of financial resources in a society. Regrettably, instances of tax fraud by individuals and businesses can compromise the integrity of the taxation system. In this context, technology, especially Artificial Intelligence (AI), can be given a new role in identifying tax fraud and safeguarding the interests of law-abiding taxpayers.

The Case of Germany

Tax fraud involves deliberate attempts to mislead authorities by employing dishonest means: concealing income, inflating expenses, money laundering, falsifying documents, offshore tax evasion, VAT fraud, payroll tax evasion. In 2020 alone, only 1.25 billion euros of the estimated 50 billion euros in illegal tax fraud were legally confirmed by final judgements in 7153 cases (Süddeutsche Zeitung, 2021; Frankfurter Allgemeine Zeitung). Compared to the German tax revenues of around 831 billion euros in 2020 (Bundesfinanzministerium, 2021), there is an estimated potential of up to +5% in tax revenues by better prevention and identification of tax fraud.

The Methods

In Germany, the tax authorities rely on various methods for identifying tax fraud. These include regular tax audits, cross-checking, and a risk-based approach to identify cases for further investigation. Additionally, Germany has a whistleblower program that allows individuals to report confidential information about potential tax fraud. On an international level, the country cooperates with other nations to combat cross-border tax evasion and financial crimes.

Tax fraud investigations involve evidence collection like financial documents and bank records. After sufficient evidence, a preliminary assessment identifies the parties involved and the amount of evaded taxes. Once a strong case is established, the accused is formally notified and allowed to respond. When someone denies wrongdoing, legal action ensues. A court evaluates evidence and arguments before giving a verdict based on the law. If found guilty, the accused may face fines, tax repayment, or even imprisonment. Both parties can challenge the decision, leading to a higher court review.

AI in tax fraud detection

For a long time, tax authorities have relied on manual techniques, which can be both time-intensive and inefficient, to detect fraudulent activities. This is where the application of AI could become significant. Through the utilization of algorithms and data analysis, AI could scrutinize vast volumes of financial and personal data. It could discern patterns and deviations that might signify potential instances of fraud. By continuously learning from past tax fraud cases and adapting to evolving fraudulent tactics, AI systems can improve their ability to identify suspicious behavior and alert tax authorities.

Still, algorithms used in tax fraud detection can differ widely: From rule-based systems through general anomaly detection to deep learning or behavioral analytics; This diversity of AI-powered tax fraud detection offers a broad spectrum of tools for identifying fraudulent activities. The choice of approach depends on factors like the system's complexity, available data, and desired level of accuracy. Striking a balance between reducing false positives, detecting emerging fraud tactics, and maintaining interpretability remains a challenge across these approaches.

Potential of AI

Tax experts see AI as an important tool for achieving tax justice, claiming that it will become increasingly impossible to evade one's tax obligations (Handelsblatt, 2021). For example, the Financial Crimes Investigation Office of the State North Rhine-Westphalia (Germany), which was set up in March 2023, aims to make greater use of digital investigation methods and artificial intelligence in the fight against financial crime and tax evasion in the future.

Your Role

Imagine, that you are one of the developers working in this team at the Financial Crimes Investigation Office.

By integrating the principles of German tax enforcement and utilizing advanced AI technologies, your team aims to effectively and efficiently strengthen the integrity of the tax system to ensure just treatment for all taxpayers.

HTML code

```
<p>---------<br>Please read until the end of the case study as well as the 9 step development lifecycle model. <br>Afterward, join back at the
</p>
```

Please read until the end of the case study as well as the 9 step development lifecycle model.
Afterward, join back at the main table.

Page 07

FG

HTML code

```
<h2>Discussion</h2>

<p>Please read until the end of the case study as well as the 9-step development lifecycle model printed on the paper. <br>Afterwards, join ba
</p>


<br><br>Please wait for each other to start into the collective discussion.</p>
<br>
<p> Continue here, after the discussion.</p>
```

Discussion

Please read until the end of the case study as well as the 9-step development lifecycle model printed on the paper.
Afterwards, join back at the main table.

Please wait for each other to start into the collective discussion.

Continue here, after the discussion.

Page 08

EndFG

FG01

Final Questionnaire

Thank you for participating in the focus group.

Now, for deeper understanding, you'll answer a follow-up questionnaire (30min) on your decision-making process, your interest in AI, its potential applications in different fields and your attitude to the advantages and disadvantages of these technologies.

A008

In a development process for an AI system like the tax fraud detection system as used by the Financial Crimes Investigation Office of North Rhine-Westphalia, developers go through many steps and have many diverse decisions to make. Every time, they have to prioritize between these different aspects, also called design principles, since time and resources are scarce and usually, you cannot optimize all of these aspects towards perfection.

Please read through the list of different design principles to consider when developing such a system.
These design principles will be included in the next set of questions, where they will also be displayed.

A007

Design Principle	**Description**
Interpretability (Explainability, Transparency, Provability)	*Explanation of the decision:* Each/any person concerned is explained in a generally understandable way why the system has classified him/her as a potential tax fraudster.
Reliability, Robustness, Security	*Reliable and secure tax fraud detection:* The automated identification of tax fraud by the computer system works consistently almost without errors, even in edge cases. It utilizes advanced security measures against hacker attacks and is always kept up to date with the latest security technology.
Accountability	*Full responsibility with the tax authority:* Should the automated tax investigation system lead to false accusations, the responsible tax authority bears full responsibility for any damage incurred.
Data privacy	*Use of data for a specific purpose only:* Only the necessary data is used by the automated tax investigation system. Any other use of the considered data is excluded.
Lawfulness and compliance	*Adherence to legal and regulatory requirements:* All stakeholders involved in the design and implementation of the algorithmic tax fraud detection system strictly comply with the law and relevant regulatory regimes. They ensure that the tax fraud detection systems' procedures and decisions are lawful and in accordance with established guidelines.
Beneficial AI	*Promoting the common good:* Both the process of development and the tax fraud detection system itself consider the common good of the society for safeguarding economic resources in an open, cooperative and sustainable way.
Safety	*Preservation of human well-being:* The algorithmic tax fraud detection system prioritizes (physical and psychological) safety of accused tax fraudsters throughout its operational lifespan, ensuring that it does not compromise their well-being until they are found guilty.
Human agency	*Appropriate human intervention:* The degree of human intervention required in the identification of tax fraud is dictated by the seriousness of ethical risks associated with the individual accusation.
Fairness	*No systematic discrimination:* No individuals (or groups) are systematically disadvantaged by the automated tax investigation.

HTML code

```
<h2>Questions?</h2>
<p>Please ask if there are any quetions regarding the principles. Clear understanding of these is important for the next tasks</p>
<br>
<p> Continue here, afterwards.</p>
```

Questions?

Please ask if there are any quetions regarding the principles. Clear understanding of these is important for the next tasks

Continue here, afterwards.

EI01

3. How important do you evaluate following design principles to be when you develop (parts of) an AI system?

In a development process for an AI system, you go through many steps and have many diverse decisions to make. Every time, you have to prioritize between different goals.
Imagine, that you are going through such a process.
Please indicate how important you find the described design principles to be if you yourself develop or are involved in developing an AI system.

Below the task you find the list of principles. Note that the order of questions is randomized and thereby not in the same order as the list of principles.

	not important at all						very important
Lawfulness and compliance	○	○	○	○	○	○	○
Interpretability (Explainability, Transparency, Provability)	○	○	○	○	○	○	○
Reliability, Robustness, Security	○	○	○	○	○	○	○
Data privacy	○	○	○	○	○	○	○
Fairness	○	○	○	○	○	○	○
Beneficial AI	○	○	○	○	○	○	○
Accountability	○	○	○	○	○	○	○
Human agency	○	○	○	○	○	○	○
Safety	○	○	○	○	○	○	○

EI01_01 Interpretability (Explainability, Transparency, Provability)
EI01_02 Reliability, Robustness, Security
EI01_03 Accountability
EI01_04 Data privacy
EI01_05 Lawfulness and compliance
EI01_06 Beneficial AI
EI01_07 Safety
EI01_08 Human agency
EI01_09 Fairness

1 = not important at all
7 = very important
-9 = Not answered

A015

As a reminder, below you find a list of the AI design principles:

A007

Design Principle	**Description**
Interpretability (Explainability, Transparency, Provability)	*Explanation of the decision:* Each/any person concerned is explained in a generally understandable way why the system has classified him/her as a potential tax fraudster.
Reliability, Robustness, Security	*Reliable and secure tax fraud detection:* The automated identification of tax fraud by the computer system works consistently almost without errors, even in edge cases. It utilizes advanced security measures against hacker attacks and is always kept up to date with the latest security technology.
Accountability	*Full responsibility with the tax authority:* Should the automated tax investigation system lead to false accusations, the responsible tax authority bears full responsibility for any damage incurred.
Data privacy	*Use of data for a specific purpose only:* Only the necessary data is used by the automated tax investigation system. Any other use of the considered data is excluded.
Lawfulness and compliance	*Adherence to legal and regulatory requirements:* All stakeholders involved in the design and implementation of the algorithmic tax fraud detection system strictly comply with the law and relevant regulatory regimes. They ensure that the tax fraud detection systems' procedures and decisions are lawful and in accordance with established guidelines.
Beneficial AI	*Promoting the common good:* Both the process of development and the tax fraud detection system itself consider the common good of the society for safeguarding economic resources in an open, cooperative and sustainable way.
Safety	*Preservation of human well-being:* The algorithmic tax fraud detection system prioritizes (physical and psychological) safety of accused tax fraudsters throughout its operational lifespan, ensuring that it does not compromise their well-being until they are found guilty.
Human agency	*Appropriate human intervention:* The degree of human intervention required in the identification of tax fraud is dictated by the seriousness of ethical risks associated with the individual accusation.
Fairness	*No systematic discrimination:* No individuals (or groups) are systematically disadvantaged by the automated tax investigation.

IP03

4. Order the following aspects after importance when you develop (parts of) an AI system.

In a development process for an AI system, you go through many steps and have many diverse decisions to make. Every time, you have to prioritize between different goals. Imagine, that you are going through such a process.

Please indicate how important you find the described ethical principles to be if you yourself develop or are involved in developing an AI system.

1: most important
9: least important

Tips & Tricks

- Double-clicking to include a principle in the ranking.
- Double-click on an item to remove it from the ranking.
- Drag & Drop does work but is not recommended.

Below the task you find the explanation of principles. Note that the order of questions is randomized and thereby not in the same order as the list of principles.

1 = Most Important
2
3
4
5
6
7
8
9 = Least Important

Human agency

Reliability, Robustness, Security

Data privacy

Fairness

Interpretability (Explainability, Transparency, Provability)

Lawfulness and compliance

Beneficial AI

Safety

Accountability

IP03_01 Interpretability (Explainability, Transparency, Provability)
IP03_02 Reliability, Robustness, Security
IP03_03 Accountability
IP03_04 Data privacy
IP03_05 Lawfulness and compliance
IP03_06 Beneficial AI
IP03_07 Safety
IP03_08 Human agency
IP03_09 Fairness

1 = Rank 1
2 = Rank 2
3 = Rank 3
4 = Rank 4
5 = Rank 5
6 = Rank 6
7 = Rank 7
8 = Rank 8
9 = Rank 9
-9 = Not ranked

A015

As a reminder, below you find a list of the AI design principles:

A007

Design Principle	**Description**
Interpretability (Explainability, Transparency, Provability)	*Explanation of the decision:* Each/any person concerned is explained in a generally understandable way why the system has classified him/her as a potential tax fraudster.
Reliability, Robustness, Security	*Reliable and secure tax fraud detection:* The automated identification of tax fraud by the computer system works consistently almost without errors, even in edge cases. It utilizes advanced security measures against hacker attacks and is always kept up to date with the latest security technology.
Accountability	*Full responsibility with the tax authority:* Should the automated tax investigation system lead to false accusations, the responsible tax authority bears full responsibility for any damage incurred.
Data privacy	*Use of data for a specific purpose only:* Only the necessary data is used by the automated tax investigation system. Any other use of the considered data is excluded.
Lawfulness and compliance	*Adherence to legal and regulatory requirements:* All stakeholders involved in the design and implementation of the algorithmic tax fraud detection system strictly comply with the law and relevant regulatory regimes. They ensure that the tax fraud detection systems' procedures and decisions are lawful and in accordance with established guidelines.
Beneficial AI	*Promoting the common good:* Both the process of development and the tax fraud detection system itself consider the common good of the society for safeguarding economic resources in an open, cooperative and sustainable way.
Safety	*Preservation of human well-being:* The algorithmic tax fraud detection system prioritizes (physical and psychological) safety of accused tax fraudsters throughout its operational lifespan, ensuring that it does not compromise their well-being until they are found guilty.
Human agency	*Appropriate human intervention:* The degree of human intervention required in the identification of tax fraud is dictated by the seriousness of ethical risks associated with the individual accusation.
Fairness	*No systematic discrimination:* No individuals (or groups) are systematically disadvantaged by the automated tax investigation.

GO02

There are different views in society on the use of artificial intelligence in different areas. Some people are more in favor, some against. Below you can see a list of different areas in which artificial intelligence could be used in the future.

Are you rather for or against the use of artificial intelligence...	completely against				completely in favor
			No answer		
...in industrial production?	○	○	○	○	○
...in secret and intelligence services?	○	○	○	○	○
...in personal everyday life?	○	○	○	○	○
...in schools and universities?	○	○	○	○	○
...in transportation?	○	○	○	○	○
...with police and security authorities?	○	○	○	○	○
...in political decisions?	○	○	○	○	○
...in land forces, air force and navy?	○	○	○	○	○
...choose “completely against”.	○	○	○	○	○
...in health care?	○	○	○	○	○
...in court?	○	○	○	○	○
...in financial institutions?	○	○	○	○	○
...in public administration?	○	○	○	○	○
			○		

GO02_01 ...in financial institutions?
GO02_02 ...in health care?
GO02_03 ...in industrial production?
GO02_04 ...in transportation?
GO02_05 ...in personal everyday life?
GO02_06 ...in schools and universities?
GO02_07 ...in public administration?
GO02_08 ...in political decisions?
GO02_09 ...in court?
GO02_10 ...with police and security authorities?
GO02_11 ...in land forces, air force and navy?
GO02_12 ...in secret and intelligence services?
GO02_13 ...choose "completely against".

1 = completely against
5 = completely in favor
-1 = No answer
-9 = Not answered

Page 14
RiskAw

GO03

You can associate both advantages and disadvantages with artificial intelligence.
Completely independent of how big you think a possible benefit is.

	no risk at all									very high risk
How great do you think the <u>risk</u> posed by artificial intelligence is for...										
...yourself?	○	○	○	○	○	○	○	○	○	○
...your friends and family?	○	○	○	○	○	○	○	○	○	○
...the whole society?	○	○	○	○	○	○	○	○	○	○

GO03_02 ...yourself?
GO03_03 ...your friends and family?
GO03_01 ...the whole society?

1 = no risk at all
10 = very high risk
-9 = Not answered

Page 15
OppAw

GO04

You can associate both advantages and disadvantages with artificial intelligence.
Completely independent of how big you think possible risks are.

	no benefit at all									very high benefit
How great do you think is the <u>benefit</u> to be gained from artificial intelligence for...										
...yourself?	○	○	○	○	○	○	○	○	○	○
...your friends and family?	○	○	○	○	○	○	○	○	○	○
...the whole society?	○	○	○	○	○	○	○	○	○	○

GO04_03 ...yourself?
GO04_02 ...your friends and family?
GO04_01 ...the whole society?

1 = no benefit at all
10 = very high benefit
-9 = Not answered

GO08

The development of artificial intelligence will continue to advance in the near future and will have an impact on various areas of life. Below you will find some possible reactions on how to deal with this situation.

To what extent do the following statements apply to you?	does not apply at all				applies completely	No Answer
I will integrate Artificial Intelligence into as many areas of my life as possible.	○	○	○	○	○	○
I will stay away from artificial intelligence wherever possible.	○	○	○	○	○	○
I will express my opinion in public discussions on Artificial Intelligence.	○	○	○	○	○	○
I will support parties or organizations that pursue the development of Artificial Intelligence as a central issue.	○	○	○	○	○	○
I will always try to use the advantages of Artificial Intelligence.	○	○	○	○	○	○
I am willing to give up advantages in order not to have to use artificial intelligence.	○	○	○	○	○	○
I will consider parties' positions on Artificial Intelligence in my future voting decisions.	○	○	○	○	○	○

GO08_01 I will stay away from artificial intelligence wherever possible. (reversed)

GO08_02 I am willing to give up advantages in order not to have to use artificial intelligence. (reversed)

1 = applies completely
5 = does not apply at all
-1 = No Answer
-9 = Not answered

GO08_03 I will always try to use the advantages of Artificial Intelligence.

GO08_04 I will integrate Artificial Intelligence into as many areas of my life as possible.

GO08_05 I will consider parties' positions on Artificial Intelligence in my future voting decisions.

GO08_06 I will support parties or organizations that pursue the development of Artificial Intelligence as a central issue.

GO08_07 I will express my opinion in public discussions on Artificial Intelligence.

1 = does not apply at all
5 = applies completely
-1 = No Answer
-9 = Not answered

GO07

5. How much do you agree with the following statements?

Below is a list of statements for evaluating trust between people and Artifical Intelligence. There are several scales for you to rate intensity of your feeling of trust, or your impression of Artificial Intelligence.

Please mark the point which best describes your feeling or your impression.	not at all						extremely
AI’s actions will have a harmful or injurious outcome.	○	○	○	○	○	○	○
I am confident in AI.	○	○	○	○	○	○	○
AI is deceptive.	○	○	○	○	○	○	○
I am wary of AI.	○	○	○	○	○	○	○
AI is dependable.	○	○	○	○	○	○	○
AI is reliable.	○	○	○	○	○	○	○
I am suspicious of AI’s intent, action or outputs.	○	○	○	○	○	○	○
I can trust AI.	○	○	○	○	○	○	○
AI provides security.	○	○	○	○	○	○	○
AI behaves in an underhanded manner.	○	○	○	○	○	○	○
I am familiar with AI.	○	○	○	○	○	○	○
AI has integrity.	○	○	○	○	○	○	○

GO07_01 AI is deceptive. (reversed)
GO07_02 AI behaves in an underhanded manner. (reversed)
GO07_03 I am suspicious of AI’s intent, action or outputs. (reversed)
GO07_04 I am wary of AI. (reversed)
GO07_05 AI’s actions will have a harmful or injurious outcome. (reversed)

1 = extremely
7 = not at all
-9 = Not answered

GO07_06 I am confident in AI.
GO07_07 AI provides security.
GO07_08 AI has integrity.
GO07_09 AI is dependable.
GO07_10 AI is reliable.
GO07_11 I can trust AI.
GO07_12 I am familiar with AI.

1 = not at all
7 = extremely
-9 = Not answered

GO09

6. How much do you agree with each statement?

Who do you think should be responsible for the actions of an AI?

The responsibility for AI is held by ...	not at all						extremely	Don't specify
the individuals or organizations that use AI systems in their projects.	○	○	○	○	○	○	○	○
the AI itself, once it has reached a certain level of autonomy and intelligence.	○	○	○	○	○	○	○	○
the government and regulators who regulate AI systems.	○	○	○	○	○	○	○	○
the companies that produce and distribute AI systems.	○	○	○	○	○	○	○	○
the developers and programmers of AI.	○	○	○	○	○	○	○	○

GO09_01 the companies that produce and distribute AI systems.
GO09_02 the government and regulators who regulate AI systems.
GO09_03 the individuals or organizations that use AI systems in their projects.
GO09_04 the developers and programmers of AI.
GO09_05 the AI itself, once it has reached a certain level of autonomy and intelligence.

1 = not at all
7 = extremely
-1 = Don't specify
-9 = Not answered

GO01

One can engage with Artificial Intelligence more or less intensively. What about you:

How much do the following statements apply to you?	does not apply at all				applies completely	No answer
Choose “does not apply at all”.	○	○	○	○	○	○
In my studies, I am more interested in classes on artificial intelligence than in classes on other domains.	○	○	○	○	○	○
I read articles about artificial intelligence with great attention.	○	○	○	○	○	○
I follow processes related to artificial intelligence with great curiosity.	○	○	○	○	○	○
In general, I am very interested in artificial intelligence.	○	○	○	○	○	○
I watch or listen to publications and contributions about artificial intelligence with great interest.	○	○	○	○	○	○
In my studies, I take great interest in courses on artificial intelligence.	○	○	○	○	○	○

GO01_01 I follow processes related to artificial intelligence with great curiosity.
GO01_02 In general, I am very interested in artificial intelligence.
GO01_03 I read articles about artificial intelligence with great attention.
GO01_04 I watch or listen to publications and contributions about artificial intelligence with great interest.
GO01_05 In my studies, I take great interest in courses on artificial intelligence.
GO01_06 In my studies, I am more interested in classes on artificial intelligence than in classes on other domains.
GO01_07 Choose "does not apply at all".

1 = does not apply at all
5 = applies completely
-1 = No answer
-9 = Not answered

DE15

7. For how long do you already engage with topics of AI in your (self-)educational journey?

[Please choose] ⌄

DE15 Length of study with AI

1 = <1 year
2 = 1-2 years
3 = 2-3 years
4 = 3-4 years
5 = 4-5 years
6 = 5-6 years
7 = 6-7 years
8 = 7-8 years
9 = 8-9 years
10 = 9-10 years
11 = 10-11 years
12 = 11-12 years
13 = >12 years
-9 = Not answered

AL03

8. I have knowledge of...

	strongly disagree						strongly agree
the input data requirements for AI.	○	○	○	○	○	○	○
AI processing methods and models.	○	○	○	○	○	○	○
using AI output and interpreting it.	○	○	○	○	○	○	○

AL03_01 the input data requirements for AI.
AL03_02 AI processing methods and models.
AL03_03 using AI output and interpreting it.

1 = strongly disagree
7 = strongly agree
-9 = Not answered

AL01

9. Thinking back over the last six months, how often have you dealt with the development of (parts of) Artificial Intelligence...?

...for the university?	□ Not at all	□ Once a month	□ About two to three times a month	□ About once a week	□ Several times a week / Daily
...in work projects?	□ Not at all	□ Once a month	□ About two to three times a month	□ About once a week	□ Several times a week / Daily
...for private purposes?	□ Not at all	□ Once a month	□ About two to three times a month	□ About once a week	□ Several times a week / Daily

AL01_03 ...for the university?
AL01_02 ...in work projects?
AL01_01 ...for private purposes?

1 = Not at all
2 = Once a month
3 = About two to three times a month
4 = About once a week
5 = Several times a week / Daily
-9 = Not answered

10. How much do you intend to work with AI in the future?

DE05

	strongly disagree				strongly agree
I intend to work on AI in the future.	○	○	○	○	○
I predict that I would work on AI.	○	○	○	○	○
Working on AI is something I would do in the future.	○	○	○	○	○

DE05_01 I intend to work on AI in the future.
DE05_02 I predict that I would work on AI.
DE05_03 Working on AI is something I would do in the future.

1 = strongly disagree
5 = strongly agree
-9 = Not answered

EL07

11. Please think about how knowledgeable you were before participating in this study:

Please assess your knowledge from before you participated in this study on the following AI design principles to consider in technology development.

	not at all				very knowledgeable
Interpretability (Explainability, Transparency, Provability)	○	○	○	○	○
Reliability, Robustness, Security	○	○	○	○	○
Accountability	○	○	○	○	○
Data Privacy	○	○	○	○	○
Lawfulness and Compliance	○	○	○	○	○
Beneficial AI	○	○	○	○	○
Safety	○	○	○	○	○
Human Agency	○	○	○	○	○
Fairness	○	○	○	○	○

EL07_01 Interpretability (Explainability, Transparency, Provability)
EL07_02 Reliability, Robustness, Security
EL07_03 Accountability
EL07_04 Data Privacy
EL07_05 Lawfulness and Compliance
EL07_06 Beneficial AI
EL07_07 Safety
EL07_08 Human Agency
EL07_09 Fairness

1 = not at all
5 = very knowledgeable
-9 = Not answered

A015

As a reminder, below you find a list of the AI design principles:

A007

Design Principle	Description
Interpretability (Explainability, Transparency, Provability)	*Explanation of the decision:* Each/any person concerned is explained in a generally understandable way why the system has classified him/her as a potential tax fraudster.
Reliability, Robustness, Security	*Reliable and secure tax fraud detection:* The automated identification of tax fraud by the computer system works consistently almost without errors, even in edge cases. It utilizes advanced security measures against hacker attacks and is always kept up to date with the latest security technology.
Accountability	*Full responsibility with the tax authority:* Should the automated tax investigation system lead to false accusations, the responsible tax authority bears full responsibility for any damage incurred.
Data privacy	*Use of data for a specific purpose only:* Only the necessary data is used by the automated tax investigation system. Any other use of the considered data is excluded.
Lawfulness and compliance	*Adherence to legal and regulatory requirements:* All stakeholders involved in the design and implementation of the algorithmic tax fraud detection system strictly comply with the law and relevant regulatory regimes. They ensure that the tax fraud detection systems' procedures and decisions are lawful and in accordance with established guidelines.
Beneficial AI	*Promoting the common good:* Both the process of development and the tax fraud detection system itself consider the common good of the society for safeguarding economic resources in an open, cooperative and sustainable way.
Safety	*Preservation of human well-being:* The algorithmic tax fraud detection system prioritizes (physical and psychological) safety of accused tax fraudsters throughout its operational lifespan, ensuring that it does not compromise their well-being until they are found guilty.
Human agency	*Appropriate human intervention:* The degree of human intervention required in the identification of tax fraud is dictated by the seriousness of ethical risks associated with the individual accusation.
Fairness	*No systematic discrimination:* No individuals (or groups) are systematically disadvantaged by the automated tax investigation.

12. Please select the highest educational qualification you have achieved so far already. DE03

○ School graduation

○ Apprenticeship degree

○ Master of crafts, specialist degree

○ Bachelor

○ Master

○ Doctorate

○ Other ________

○ do not want to answer

DE03 Level of Education

1 = School graduation
5 = Apprenticeship degree
6 = Master of crafts, specialist degree
2 = Bachelor
3 = Master
4 = Doctorate
7 = Other
8 = do not want to answer
-9 = Not answered

DE03_07 Other

Free text

13. In which field of study are you currently pursuing a degree? DE04

○ Computer Science / Computer Engineering / Computational Engineering

○ Data Science

○ Data Analystics and Decision Science / Computational Social Systems

○ Economics / Management & Engineering in Technology

○ Automatisierungstechnik

○ Automotive Engineering

○ Elektrotechnik

○ Simulation Science

○ Software Systems Engineering

○ Other with a minor in Computer Science ________

○ Other ________

DE04 Field of Studies

1 = Computer Science / Computer Engineering / Computational Engineering
6 = Data Science
12 = Data Analystics and Decision Science / Computational Social Systems
9 = Economics / Management & Engineering in Technology
13 = Automatisierungstechnik
15 = Automotive Engineering
14 = Elektrotechnik
7 = Simulation Science
8 = Software Systems Engineering
3 = Other with a minor in Computer Science
11 = Other
-9 = Not answered

DE04_03 Other with a minor in Computer Science

DE04_11 Other

Free text

14. What university are you studying at?

DE13

○ RWTH Aachen University

○ Universität zu Köln

○ Other ____________________

DE13 University

1 = RWTH Aachen University
3 = Universität zu Köln
2 = Other
-9 = Not answered

DE13_02 Other

Free text

Page 23

15. Have you taken any dedicated full-length training/course/seminar on ethics in technology, data, AI, autonomous systems or comparable?

EL08

Example courses are

- *“Ethics, Technology and Data” by Prof. Nagel*
- *“Lecture Introduction to the Philosophy of Science and Technology” by Prof. Pantsar*
- *“Seminar Moral Reasoning” by Dr. Colombo*
- *“Social and Technological Chance” by Prof. Geffner*
- *“Ethik im Zeitalter der Digitalisierung” by Dr. Teille*
- *“Digitale Menschenrechte” by Projekt Leonardo*
- *...*

(This does not include courses, that have a short section (<70%) on ethics, like “Introduction to Data Science by Prof. van der Aalst”. In cases like these, please select the last option.)

○ Yes, I finished a dedicated full-length training/course on ethics in technology, data, AI, autonomous systems or comparable

○ No, I finished a short (1-3h) dedicated training/course on ethics in technology, data, AI, autonomous systems or comparable

○ No, I began a dedicated training/course but did not finish it.

○ No, I just finished *a general training/course on AI* with a short section on ethics in technology, data, AI, autonomous systems or comparable

○ No.

EL08 AIEL?

1 = Yes, I finished a dedicated full-length training/course on ethics in technology, data, AI, autonomous systems or comparable
3 = No, I finished a short (1-3h) dedicated training/course on ethics in technology, data, AI, autonomous systems or comparable
4 = No, I began a dedicated training/course but did not finish it.
5 = No, I just finished a general training/course on AI with a short section on ethics in technology, data, AI, autonomous systems or comparable
2 = No.
-9 = Not answered

1 Active Filter(s)

Filter EL08/F1
If any of the following options is selected: **2, 5**
Then jump to page **EAIL interest** after the next button was clicked

EL03

16. <u>What dedicated training/course on ethics</u> in technology, data, AI or autonomous systems did you visit?

- "Ethics, Technology and Data" by Prof. Nagel
- "Lecture Introduction to the Philosophy of Science and Technology" by Prof. Pantsar
- "Seminar Moral Reasoning" by Dr. Colombo
- "Social and Technological Chance" by Prof. Geffner

○ "Ethics, Technology and Data" by Prof. Nagel at RWTH Aachen University

○ "Lecture Introduction to the Philosophy of Science and Technology" by Prof. Pantsar at RWTH Aachen University

○ "Seminar Moral Reasoning" by Dr. Colombo at RWTH Aachen University

○ "Social and Technological Change" by Prof. Geffner at RWTH Aachen University

○ "Ethik im Zeitalter der Digitalisierung" by Dr. Teille at Otto von Guericke Universität Magdeburg

○ Other
"course name" by "course host" at "institute"

EL03 What EAIL?

1 = "Ethics, Technology and Data" by Prof. Nagel at RWTH Aachen University
3 = "Lecture Introduction to the Philosophy of Science and Technology" by Prof. Pantsar at RWTH Aachen University
4 = "Seminar Moral Reasoning" by Dr. Colombo at RWTH Aachen University
5 = "Social and Technological Change" by Prof. Geffner at RWTH Aachen University
6 = "Ethik im Zeitalter der Digitalisierung" by Dr. Teille at Otto von Guericke Universität Magdeburg
2 = Other
-9 = Not answered

EL03_02 Other

Free text

EL05

17. <u>When did you last participate</u> in a dedicated course/trainings on ethics in technology, data, AI or autonomous systems?

○ 2024

○ 2023

○ 2022

○ 2021

○ 2020

○ 2019

○ 2018

○ 2017

○ Before 2017:

EL05 Last Time AIEL

9 = 2024
1 = 2023
2 = 2022
3 = 2021
4 = 2020
5 = 2019
6 = 2018
7 = 2017
8 = Before 2017:
-9 = Not answered

EL05_08 Before 2017

Free text

EL09

18. How much do you agree with the following statement?

	not at all						completely
I enjoy to discuss how technology affects society.	○	○	○	○	○	○	○
I enjoy to read about implications that technology has on society.	○	○	○	○	○	○	○
I like to inform myself about how technology impacts the life of individuals.	○	○	○	○	○	○	○
I am interested in ethics of technology, data, algorithmic systems or autonomous systems.	○	○	○	○	○	○	○

EL09_01 I am interested in ethics of technology, data, algorithmic systems or autonomous systems.
EL09_02 I enjoy to read about implications that technology has on society.
EL09_03 I enjoy to discuss how technology affects society.
EL09_04 I like to inform myself about how technology impacts the life of individuals.

1 = not at all
7 = completely
-9 = Not answered

DE01

19. How old are you?

○ 18-21
○ 22-25
○ 26-30
○ 30+
○ do not want to answer

DE01 Age
1 = 18-21
2 = 22-25
3 = 26-30
5 = 30+
6 = do not want to answer
-9 = Not answered

DE02

20. What gender do you identify yourself with?

○ female
○ male
○ diverse
○ do not want to answer

DE02 Gender
1 = female
2 = male
3 = diverse
4 = do not want to answer
-9 = Not answered

DE14

21. Where do you position yourself on the political spectrum?

I identify myself as...	liberal/ libertarian	□	□	□	□	□	conservative/ authoritarian

DE14_01 I identify myself as...
1 = liberal/ libertarian
7 = conservative/ authoritarian
-9 = Not answered

DE16

The power should lie with the...	government	□	□	□	□	□	market

DE16_01 The power should lie with the...
1 = government
7 = market
-9 = Not answered

Z904

Debriefing

Thank you for participating in this study investigating the decisions and attitudes of future developers of Artificial Intelligence.

This research project investigates the views of future developers on emerging technologies and their applications, with a focus on Artificial Intelligence (AI). We want to find out whether students with experience in ethics/philosophy courses show different behavior in their rating and ranking ethical AI Principles, as well as different attitudes towards AI compared to students without. Do students, with and without the educational experiences, show different behavior and attitudes? If yes, how do these look like?

The results of this study will help to determine the relevance of different study curricula and (un)intended side effects.

If you decide that you want to delete your data partially or completely, contact the experimental supervisor by naming your code for anonymization.

If you have any further questions about our study, you may contact me by E-Mail:
ben.schultz@rwth-aachen.de

Z912

22. Do you want to add anything?

Z912_01 [01]
Free text

Z906

23. Study Results – Long term Follow Up

Publication of this Study
This study will be published in collaboration with the Chaos Computer Club Germany.
If you are interested in the results, you can sign up to receive information about the study.
A list of email addresses of interested participants is stored separately from the collected data of the questionnaire.

Long-Term Follow-Up Study & other Studies
Since, attitudes and behavioral patterns can change over time, we'd love to stay in touch with you. In an irregular manner, we'll ask you whether you want to participate in further follow-up studies to this topic, as well as on studies in the related fields from human-technology interaction in general as well as human-robot interaction especially. You will be able to opt out at any point, by sending a mail to itec@humtec.rwth-aachen.de

Your E-Mail address will be stored separately from your information collected during the study.

□ I'm interested in being informed about **further studies at the iTec Institute at RWTH Aachen University**.

I agree that my email address will be saved for study information purposes.

My study data will continue to be anonymous, and my personal data will not be passed on to third parties.

□ I am interested in the **results of this study**. Please send me the open-access paper by e-mail.

I agree that my email address will be saved until the results are sent.

My study data will continue to be anonymous, and my personal data will not be passed on to third parties.

Z802

24. How do want to be compensated?

participation hours

If you like to receive participation hours as a student of RWTH Aachen University, be informed that this is only possible for Informatics students with a minor in psychology.

monetary compensation
For participation in this study, you receive 20€ cash.

In each case you have the chance to sign up for the lottery.

○ I like to receive 2 participation hours.

○ I like to receive 20€ in cash.

Z802 Which Compensation? FGxGA

2 = I like to receive 2 participation hours.
1 = I like to receive 20€ in cash.
-9 = Not answered

2 Active Filter(s)

Filter Z802/F1
If any of the following options is selected: **1**
Then hide question/text **Z801** placed later in the questionnaire

Filter Z802/F2
If any of the following options is selected: **2**
Then display question/text **Z801** placed later in the questionnaire (otherwise hide)

25. Participation Hours

Z801

For compensation, you can receive participation hours for your studies.
Eligible are, at RWTH Aachen University: Computer science students with a minor in psychology.

For RWTH Aachen students:
After participation in this study, the institute will be informed about your additional participation hours.
Please reach out to ben.schultz@rwth-aachen.de to receive your verification document by referring to your participation in this study “Future Developers’ Perspective on AI”.

□ I would like to receive participation hours.

I agree that my name and matriculation number will be saved to grant me participation hours for this study.

My study data will continue to be anonymous, and my personal information will not be passed on to third parties.

I accept all terms and conditions of the process to receive my participation hours as stated above.

26. Lottery for over 200€

Z803

You can take part in a lottery to win one of three prices:

- 100€
- 50€
- 50€

On the next page, you receive your personal referral code to increase your chances even further. If your fellow students use this code and finish this questionnaire, both you and them receive an extra ticket for this lottery.

If you yourself already received a referral code from somebody else, note it down at this point, so that you and your referrer receive one additional lottery ticket.

You can only take part in the lottery once.

The lottery will be drawn after data collection finished. If you win, you will receive an email to indicate your bank details for direct bank transfer.

□ I would like to participate in the lottery.

I agree that my name, email address and referral code will be saved until the winner is drawn.

My study data will continue to be anonymous, and my personal data will not be passed on to third parties.

I accept all terms and conditions of the lottery as named above.

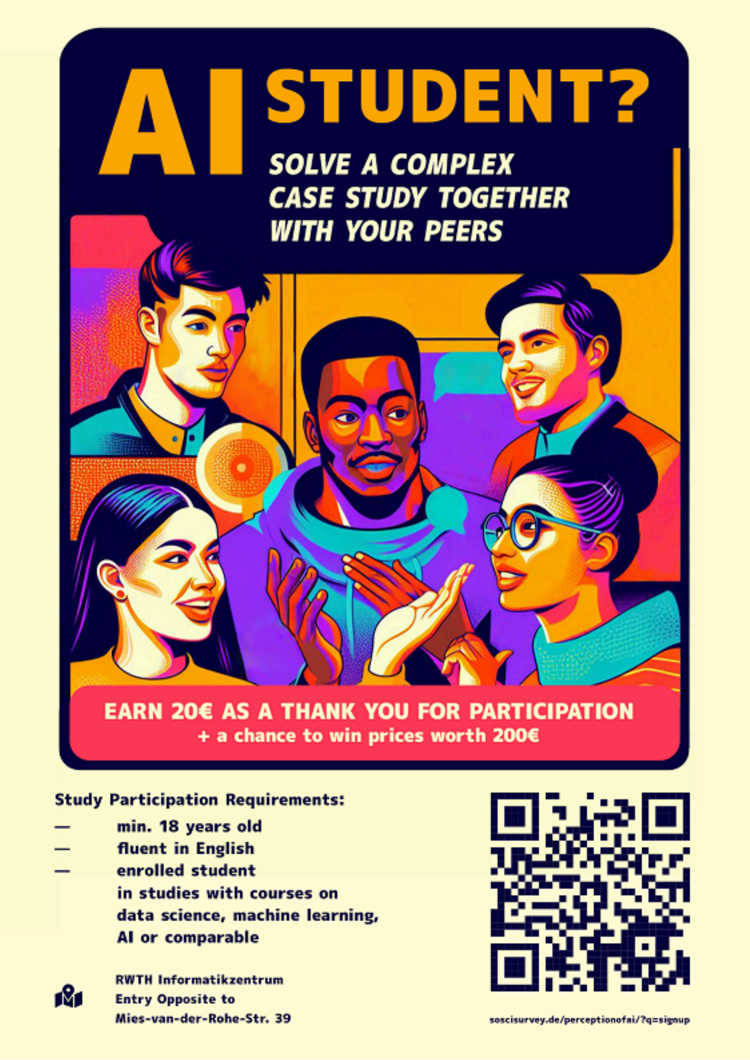
PHP code

```
put('Z804_01', random(100000, 999999));



$code1 = value('Z804_01', 'free');


html('<p><b>Your Referral Code to increase your chances of winning</b>
<br><br><br>If you want to participate in the lottery, refer this code to others. If they complete the st
<br>
<b><br><br> Your code is: '.$code1.' </b></p>
<br><br>
<i>Take a picture or note it down. The study is finished afterwards.</i>
');
```

HTML code

```
<br> For referral, you can use the following flyer:
<br> <br>
<div style="text-align: center;">

<img src="FocusGroupFlyer_small.png">
</div>
```

For referral, you can use the following flyer:


AI STUDENT?
SOLVE A COMPLEX CASE STUDY TOGETHER WITH YOUR PEERS
EARN 20€ AS A THANK YOU FOR PARTICIPATION
+ a chance to win prices worth 200€
Study Participation Requirements:
— min. 18 years old
— fluent in English
— enrolled student in studies with courses on data science, machine learning, AI or comparable
RWTH Informatikzentrum
Entry Opposite to
Mies-van-der-Rohe-Str. 39
soscisurvey.de/perceptionofai/?q=signup

Thank you very much for taking part in this study!

I would like to thank you very much for helping me.

Your answers were transmitted, you may tell the experimenter and swap back to the main table.

Appendix E

A Priori G*Power Analysis

E.1 Study 1 Power Analysis

F-test – ANOVA: Fixed effects, omnibus, one-way

Analysis: A priori: Compure required sample size given α, power, and effect size

Input:	Effect size f	= .4
	α err prob	= .05
	Power (1-β err prob)	= .95
	Number of groups	= 2
Output:	Noncentrality parameter λ	13.44
	Critial F	3.9573883
	Numerator *df*	1
	Denominator *df*	82
	Total sample size	84
	Actual power	.9518269

E.2 Study 2 Power Analysis

Input	Effect size w	= .5
	α err prob	= .05
	Power (1-β err prob	= .8
	Df	= 2
Output:	Noncentrality parameter λ	= 9.75
	Critial χ^2	= 5.9914645
	Total sample size	= 39
	Actual power	= .8049793

Appendix F

Mini Focus Group Protocol Template

This appendix includes the mini focus group protocol template used within the focus groups to guide the semi-structured discussion by the researcher. The mini focus group guidebook as a similar variant with less information for participants can be found uploaded to the OSF.

AI in Tax Fraud Detection Systems

Efficient and equitable taxation is essential to ensure the proper utilization of financial resources in a society. Regrettably, instances of tax fraud by individuals and businesses can compromise the integrity of the taxation system. In this context, technology, especially Artificial Intelligence (AI), can be given a new role in identifying tax fraud and safeguarding the interests of law-abiding taxpayers.

The Case of Germany

Tax fraud involves deliberate attempts to mislead authorities by employing dishonest means: concealing income, inflating expenses, money laundering, falsifying documents, offshore tax evasion, VAT fraud, payroll tax evasion. In 2020 alone, only 1.25 billion euros of the estimated 50 billion euros in illegal tax fraud were legally confirmed by final judgements in 7153 cases (Süddeutsche Zeitung, 2021; Frankfurter Allgemeine Zeitung). Compared to the German tax revenues of around 831 billion euros in 2020 (Bundesfinanzministerium, 2021), there is an estimated potential of up to +5% in tax revenues by better prevention and identification of tax fraud.

The Methods

In Germany, the tax authorities rely on various methods for identifying tax fraud. These include regular tax audits, cross-checking, and a risk-based approach to identify cases for further investigation. Additionally, Germany has a whistleblower program that allows individuals to report confidential information about potential tax fraud. On an international level, the country cooperates with other nations to combat cross-border tax evasion and financial crimes.

Tax fraud investigations involve evidence collection like financial documents and bank records. After sufficient evidence, a preliminary assessment identifies the parties involved and the amount of evaded taxes. Once a strong case is established, the accused is formally notified and allowed to respond. When someone denies wrongdoing, legal action ensues. A court evaluates evidence and arguments before giving a verdict based on the law. If found guilty, the accused may face fines, tax repayment, or even imprisonment. Both parties can challenge the decision, leading to a higher court review.

AI in tax fraud detection

For a long time, tax authorities have relied on manual techniques, which can be both time-intensive and inefficient, to detect fraudulent activities. This is where the application of AI could become significant. Through the utilization of algorithms and data analysis, AI could scrutinize vast volumes of financial and personal data. It could discern patterns and deviations that might signify potential instances of fraud. By continuously learning from past tax fraud cases and adapting to evolving fraudulent tactics, AI systems can improve their ability to identify suspicious behavior and alert tax authorities.

Still, algorithms used in tax fraud detection can differ widely: From rule-based systems through general anomaly detection to deep learning or behavioral analytics; This diversity of AI-powered tax fraud detection offers a broad spectrum of tools for identifying fraudulent activities. The choice of approach depends on factors like the system's complexity, available data, and desired level of accuracy. Striking a balance between reducing false positives, detecting emerging fraud tactics, and maintaining interpretability remains a challenge across these approaches.

Potential of AI

Tax experts see AI as an important tool for achieving tax justice, claiming that it will become increasingly impossible to evade one's tax obligations (Handelsblatt, 2021). For example, the Financial Crimes Investigation Office of the State North Rhine-Westphalia (Germany), which was set up in March 2023, aims to make greater use of digital investigation methods and artificial intelligence in the fight against financial crime and tax evasion in the future.

Your Role

Imagine, that you are one of the developers working in this team at the Financial Crimes Investigation Office. By integrating the principles of German tax enforcement and utilizing advanced AI technologies, your team aims to effectively and efficiently strengthen the integrity of the tax system to ensure just treatment for all taxpayers.

Self Intro: I'm Ben, having a Bachelor in Business Information Technology and now I'm working on my masters thesis on the work of developers in the field. I set up this discussion to understand more about how you as students interested in AI; possibly future developers of AI; approach complex scenarios for AI development. That's why I made you read the Use Case early on. And I want to find out how Development Frameworks go in line with this. That's why I introduced you to the development lifecycle model.

First of all, I would like to ask if everyone could briefly introduce themselves.

Please include only your first or acronym for anonymity, what you study and most importantly your motivation for being involved in AI.

ask all – 1min

Familiarity with Tax Fraud Detection*:*

I now like to hear from you What is your understanding of tax fraud detection systems and their technical complexities?

ask all – 5min

Data Collection Strategies:
Think back to the last time you worked on data collection; either in your studies or for work:

What are your preferred methods for collecting and managing the data? How would you address a system for fraud detection?

ask 2 – 5min

Algorithm Selection:
For algorithms I'd like to discuss the selection with a similar approach. Just from your mind, how would you usually decide on the algorithms and models to be used in developing a tax fraud detection system?

ask 2 – 5min

Imagine coming together with the development team after you ran enough test runs and tweaked the system:

What factors do you consider to integrate the AI tax fraud detection models into previous the given world? How do you integrate it into administrative workflows and juridical processes?

ask 2

Performance Metrics*:*
Just from your experience and what you learned in your studies and career;
How do you approach the task to measure the performance as well as the overall impact of the tax fraud detection system the team develops?

ask 4

Case Study – Future Developers' Perspective on AI

Write Down:

Key Challenges*:*

What do you consider as the three most significant challenges in creating a good tax fraud detection system?

… let them write…

You can share those insights with us in the group or we'll only analyze them afterwards anonymously. So, does somebody want to talk about it?

ask up to 2

Summarize:

As an ending question; We'll go around the table.

From all the things that we talked about today. If you could talk to the manager overseeing this project; What most important advice would you give them?

ask all

Appendix G

Mini Focus Group Transcripts

This appendix includes two exemplary mini focus group transcripts giving insights into two mini focus groups on AIEE and nAIEE: The first transcript in this appendix is on the in-person group #3 with four nAIEE students. The recording has been 33 minutes and 51 seconds long. The second transcript shown here is on the online mini focus group #17 with two AIEE students. The recording is 34 minutes and 7 seconds long.

All 15 transcripts and audio data are uploaded under XXXXX in the OSF of this thesis. Due to anonymization of participants in the mini focus group, the people involved are stated as "Speaker" and have been numbered by the order of entering the discussion. Participants ID's refer to the ID of the questionnaire dataset.

G.1 Transcript Mini Focus Group #3

(*Mini Focus Group 3 - in person*; *Condition: No AI Ethics Education*; *Participants: #598, #600, #601, #602*; *Moderator: Speaker 1*)

00:00:00 Speaker 1: OK. So if we think about this case and the first step maybe from the model to to think about like what from your understanding are the complexities of the situation? Yeah, maybe.

00:00:25 Speaker 2: So I can. Probably like so the data will be from various sources and in different countries. Also because people set up their bank accounts in different countries. So just collecting that data for a particular person like I feel like that will be a pretty challenging task in itself. OK. Yeah. So I would say data collection in.

00:00:47 Speaker 1: General. Yeah. Yeah. And also having the full view on one person already.

00:00:57 Speaker 3: I can add like I also think my first thought about it was more on the a legal and cooperation aspect of data collection site because like we have a name like on our passport and if that is like the same throughout all comes, I don't think it's technically that difficult I think but more the collaboration. So getting a whole picture definitely I would agree on and maybe also now technically also on the model side. So the data and the model. On the model side, I would say depending on what we will discuss, maybe like the once training it with the white labels and also let's see what how many like. I don't know if there's many false positives and so on. Maybe. I don't know. Yeah.

00:01:43 Speaker 2: Yeah, I agree with that. I think like for having false positives like you should have additional frameworks in place along with this AI system to sort of minimize the false positives. And as you do the continuous testing, learning and stuff, you try to minimize the false positive rate of the model.

00:02:02 Speaker 1: So what would you think would happen if you would not be able to minimize it.

00:02:10 Speaker 2: So, like almost in, I'm not very sure about Germany, but in almost every country the legal system is already very burdened and trying to like investigate more people who are innocent and stuff like that? It's kind of not a very good use of resources, I would say. And apart from that, it's also, yeah, like I think from a personal like a liberty sort of point of view. Like, yeah, like if if we only trust the AI system, like prosecuting a person who is innocent, it's yeah. Like, not a good idea. I don't know how to frame it. Yeah.

00:02:53 Speaker 1: Do you also have any ideas on it? Like just from the general view on the thing?

00:02:58 Speaker 4: Yeah, I'm currently asking myself if it's even possible to get all the data, but could it also be the case that some country doesn't want to give the information? I know if that's even a possibility, but that for example, you don't even know if there aren't any more bank accounts in some other region that we don't know about. So I don't know if it's even able. If we are able to get all the data that we need. I don't know.

00:03:29 Speaker 1: Any thoughts from your side?

00:03:30 Speaker 5: ohh yeah, I was just thinking also like. I think the taxation laws kind of change every year, so I how how does like that impact the data collection and you know how do you interpret the data then if the like the law the law related to taxes like change every. I was thinking like.

00:03:51 Speaker 1: It's an important point, true, and especially if you want to use the old data as. Well, yeah.

00:03:59 Speaker 4: But in the end, do we want to have an fully autonomous AI which is just working on itself, or do we also have other employers which are also looking over the data in parallel? So I think it's also something to consider.

00:04:17 Speaker 1: That's a point, yeah.

00:04:18 Speaker 5: Is it like is it just a tool for a prosecutor or it's it's like a partner to a prosecutor or a tool for a prosecutor like, you know, like, who's probably looking through the cases.

00:04:30 Speaker 1: What would you recommend?

00:04:32 Speaker 2: My personal recommendation would be a tool.

00:04:40 Speaker 1: Yeah. So maybe we can jump to the next question on the data collection, I mean, we talked about it already a little bit how what are your preferred methods in regard to this? Like what would you, how would you tackle this big problem of both in within the country nationally, but also internationally trying to collect data.

00:05:08 Speaker 2: So like I would also like to ask here that what kind of previous old data we have like you mentioned that we have that in a digital format, so we won't have to go through paper records. I hope so. Like trying like, just gathering more information on what previous data the the Crime Office already has.

00:05:31 Speaker 3: Which attributes, yeah. And building on top of that data model. And then what I have imagined is like the most data sources have, like the banks, the financial institutions And then I think you must manually depending which database they have, but it's called like merging all those those data table, sources and a common one. I would think of it at the moment.

00:06:01 Speaker 1: Any ideas on like how to solve the problem of having so many different banks and different sources?

00:06:10 Speaker 5: I think probably you would need like very skilled data engineers who can query like different type of databases. Probably a bit more technical but yeah.

00:06:19 Speaker 1: That's that's fine.

00:06:22 Speaker 3: Yes. Yeah.

00:06:22 Speaker 2: Yeah, probably like more domain specialists as well because they would have some experience working in this field and like they are at least like subject matter experts or. Trying to like work on this problem itself, something like that will.

00:06:38 Speaker 1: OK.

00:06:40 Speaker 2: And I think as she mentioned, like having the like jumping through the legal burdens of, like asking another country for the, like, the bank records of that particular individual, like it will be a very difficult task considering some countries are like they have built their economy around being a tax haven themselves. So yeah.

00:07:02 Speaker 1: But I mean, we also have to differentiate between training data and the data which is checked. So I think for both cases you need to have the contract with the other countries. But focusing on the training data and the old data, any other things you would like try? To work on, I mean, you've been talking about the variables already. What would you try to recognize there?

00:07:35 Speaker 3: Yeah, yeah, there was maybe later. I can think of. What do you mean by? I'm thinking. But I was thinking more about the common attributes. Like for example, in Germany, there's like the tax number, let's say that should be always be accounted for either you must write in the financial transactions like between companies right to have that number of other countries have and other variables. I currently I don't. I'm thinking I don't know.

00:08:09 Speaker 2: So you're asking like the features in the training data, OK. I mean identification of like different data sources will have different sort of tax IDs sort of and like trying to merge. So all that, let's assuming we have data in a tabular format like merging them all together, I think that will be the first step and a very challenging task, yeah.

00:08:39 Speaker 3: Because it needs also like manual control of a human to see if it did right. Not like automate that I would say because if there's already an error that's that that's like... Like the whole I model depends on that.

00:08:52 Speaker 3: That's fine.

00:08:54 Speaker 1: So for maybe the next if we imagine we had good data. And how would you proceed for? Yeah, the models and what would you try to work on there? Which approach do you think might be helpful to use based on the old data to predict and or find people that might have, yeah, dealt in the area of tax fraud and are active in there. Even without prediction, but maybe just seeing where we have this data, but usually it might be similar or not. How would you approach that?

00:09:47 Speaker 3: Like I'm thinking the the the main question now. For this but like. Is it now already about the model or are we a step?

00:09:58 Speaker 1: What are you thinking about?

00:10:06 Speaker 3: One thought is like I wanted. I was thinking about the model selection, let's say on which technical foundation or concept or model itself like from decision trees to I don't know, a neural network that's so on. I was thinking about those models, but that's like also more on the master level. That's what I think about how to ask questions. And if that's already one step too much and if I'm missing something before that, that's what.

00:10:37 Speaker 2: So I think like here. So I'm assuming we have like supervised learning data like and it also depends on how we are sort of like framing the problem like are we going for a anomaly detection or like risk based approach like or something like that. So yeah like I think framing the problem itself at this point will be a pretty good idea.

00:11:04 Speaker 1: Again. So what would you imagine to to frame it?

00:11:08 Speaker 2: Yeah. So I would like I don't have previous experience in tax fraud detection, but I would probably go for something like anomaly detection problem.

00:11:20 Speaker 5: I think like taxes also like if it's some organizations like you know e-commerce related stuff might be different and probably like retail. I don't know like the data itself might be different, so I feel like you'll have to approach it probably a little bit domain specific industry specific because the laws apply to each industry also kind of differs. Probably so yeah, like No two organizations or people might be taxed the same way like because yeah. So I mean, when you collect the data, like if you're detecting that somebody is a fraud in the retail industry, you would probably need like a lot of data points from that industry.

00:12:06 Speaker 2: So like a different sort of modeling approach for each industry?

00:12:09 Speaker 5: Industry. Yeah. I don't think there'll be, like, one model that fits everything. Like you might have to run, like, multiple different variants.

00:12:17 Speaker 1: It's an interesting idea to bring those together in the system under one hood, but still with different models. OK, any other thoughts on this?

00:12:29 Speaker 4: I think I would just agree that those are pretty good points and I think also anomaly detection would be the best model. I could also think of.

00:12:40 Speaker 1: OK, so if you had run enough tests on like setting up the system and it would work from your side like what factors would you consider to integrate the system into the previous given processes? So coming back from the idea, To the idea of, well, it's it was quite manually before. How should it look like now or what do you have to consider when bringing that into practice and why?

00:13:15 Speaker 2: So I don't know how relevant this point is, but I think like. Especially in the countries in the EU, like having the like the privacy agreement and everything, and like the models they have previous data of individuals. So like navigating that hurdle itself is like a primary challenge. I would say one of the primary challenges.

00:13:41 Speaker 1: OK.

00:13:44 Speaker 5: I would, I would say, like people who are like entering the data, like at that level, they need to be given like the like proper instructions on how and how they should enter data so that no error creeps into it. Like. I mean, there's no human error or less human error like yeah.

00:14:04 Speaker 3: Which step?

00:14:05 Speaker 5: Like right at the data collection, like somebody's entering like some records of some financials like they should have a protocol that they should follow strictly. And yeah. So for for future purpose also when you combine the old and new data like you have less effort in harmonizing them together and so you probably have to start at that. That level the latest.

00:14:30 Speaker 4: I wanted to say something else. I don't know. Is the privacy point, really an issue? Because doesn't the government has all the data anyways so they know who you are and they know. Your your ID and what I know because they have to check. Everything anyways, so I don't know if that's really a point, but don't don't.

00:14:53 Speaker 3: Yeah, yeah, to that question. I would say also like so it's about the transactions we make, I would say and. Yeah. I would also agree to the anonymity, the detection, but also like maybe there's some kind of pattern of, let's say, suspicious behaviour. Let's say I would look out for that at the training step. Like I would say, best case is that it's like that step, the most automatic one. I hope that like we just build the pipelines and the data, the training data will be as quick as possible. And then like?

00:15:36 Speaker 3: The labeling of that or whatever model we will choose like. How that is handled? and more also I was thinking about how specific the training data looks like so. The best thing would be if we had like an already solved data source of confirmed tax evasion and then it's labeled. And then like more of a pattern. But because my first thought through also keep the false positives and the Yeah, yeah the false negatives...

00:16:20 Speaker 1: That's a good point. Yeah, especially like to to have a real thing to train it on, yeah. Even though that. Even in, yeah, those processes, it's quite interesting to see and work with that. Umm. So if we think about bringing that into the processes, we also got to think about how do we check whether it works out in the end as it is intended and what the, how Pricewaterhouse Coopers, framing it, the value of the system is and whether that value is met and it's really bringing it into practice. If you think that on that for this system, how would you try to? Yeah, check for the Impact and the value?

00:17:18 Speaker 2: Do you mean like how will we roll out the model and like? Something like that.

00:17:23 Speaker 5: The deployment of the.

00:17:24 Speaker 1: Model we talked about deployment already, but. The effects of the model.

00:17:29 Speaker 4: For the evaluation, how good it performs?

00:17:33 Speaker 3: On one, yeah, but also how how we practically measure it is also the main question. Yeah. OK.

00:17:41 Speaker 4: Maybe I would try to use it on past cases and see if it can detect those past cases.

00:17:51 Speaker 2: I mean like that will be like I think before the deployment itself that step should be taken care of during the training, validation and testing. So like I would probably like implement it in a like a rollout phase sort of like for a particular demographic or let's say a particular industry, I would roll the model out. I would also like at the same time have the manual tax fraud detecting people who are there and like sort of, like, collaborate, the findings of both the AI model and the people to see how good it is performing and yeah, like it should be slow for each industry or each whatever cluster we decide on previously like it should be rolled out in steps. It probably will take at least. A year or more for this to happen in my yeah.

00:18:43 Speaker 5: OK, I feel like the stakeholders like, yeah, like the people like, you know, the the legal people like in it, they have to like lawyers or prosecutors and maybe the legal teams of organizations themselves like, they probably have to give, like, continuous feedback on the results and because I feel like a lot of times, like people are more open to AI nowadays, but at the end of the day, they probably still think their decision is right and you know, like humans have a final say in stuff. Even though AI might be predicting something. So a lot of people will rather go with their experience than like trust a model. So to build that trust, we got to like, work closely with stakeholders and, you know, make them involved in it. So they also feel like they're part of it. Part of the decision making and just be relying on the AI, etc.

00:19:38 Speaker 3: Yeah. I would also agree also with this. So with the say, so in my opinion also AI should like on one side it can like automate us let's say, but that shouldn't be like it should be like added value in my view. So stuff that's definitely the most value there is also, the patterns it can see and that we human like can really difficult see. So with those recommendation of the model how I would measure the effect would be like from the recommendation like how I imagined this the like, what process is then initiated. So from the recommendation, if it's a fraud or maybe if we have maybe we have more than two levels. So not just like positive and negative. Let's say what happens to that and then on the business site like KPI would be and of course how much money was, let's say collected and was positive. But also like the length aspect of that recommendations for me, in my opinion, also important so.

00:20:47 Speaker 1: What do you mean by the last point?

00:20:48 Speaker 3: The last point, yeah. So let's for specific cases. I would say that there are more difficult ones and maybe if you also try to minimize that. So it's about like low effort, high result let's say. Focus on that. But also there are also really difficult ones with high reward, let's say also to balance that out. But definitely like the ones that did take a long time. But in the end it's not that much money that's being summed up and yeah, working that out with the stakeholders. Also may be an important part I would say also maybe with other departments of course of the business like the controlling side, yeah.

00:21:33 Speaker 1: And super point. So if you bring all those steps together and think about what we just discussed, do you have like maybe 3 aspects in your mind which are crucial to be recognized but or consider when setting up such a system, I would like you to just maybe write it down on last. I think on the last page this is a slot for it.

00:22:04 Speaker 2: OK, so like oh. Or in the entirety of our discussion till now, Sir.

00:22:08 Speaker 1: Yeah, right. Because I think we don't have time to discuss. All of those. Aspects, but just to to have short here in it and so that I can read through it later on in detail. So are you done and happy with with working your return? Thanks for that. Maybe as a short wrap up around if you had the possibility to speak to the project manager and the team leader.

00:30:17 Speaker 5: So like the project manager of the technical team, who is going to implement the solution?

00:30:21 Speaker 3: Or in general, with all stakeholders involved, like project manager and which sense?

00:30:24 Speaker 1: I mean, maybe also all stakeholders, why not?

00:30:29 Speaker 2: In that case, probably the like the integration of this new system with the existing teams and how they like they develop it. And they also like build the trust among the existing team and the system. So I think that's one of the probably the most important part.

00:30:50 Speaker 1: OK.

00:30:51 Speaker 5: Yeah, I think mine would be something on similar lines, yeah.

00:30:52 Speaker 1: OK.

00:30:56 Speaker 3: Yeah, maybe. Yeah, before that. So I would say so there's lots of money impact, let's say. So I would say go far and beyond, I would say to him personally, I would try to do kind of a complex model, so doing in with this kind of impact, more difficult things rather than going for an easier solution in my opinion I would say. And also like it's an important aspect like with already existing team not focusing also focusing kind of on them and the use of the tool with them and not just like the two. And then just give it to them like I think the effect can be that I would think definitely observe like the productivity productivity side and so on and also even the ideas from the existing team getting input and inspiration for that but then also like if you do it complex and the interface doing it easy and the values and recommendations comes out of it. Doing that also like in a way that they easily understand and like everyone should be able to understand it, not like some kind of zero point some value and they are like what? Yeah.

00:32:16 Speaker 1: Thank you.

00:32:17 Speaker 4: Yeah, for me something similar. So that they also look at how to integrate it into the workflow so that everyone knows what this is and how and that they can also build some kind of trust to this. New method and that they all can kind of see how this works and have some slow start into this.

00:32:43 Speaker 5: Also like a good team that sustains the project after it deploys like you know this, this team like should be technically sound as well as like you know, legally. Also I would say they would need people in it because at some point your machine might be going out of I mean your model that you have created might be going out of date. no longer be good enough. So to detect this it would require technical people as well as subject experts.

00:33:12 Speaker 1: What do you mean by subject experts?

00:33:13 Speaker 5: So something for example like people have done, say international taxation laws or you know like who have done their taxes like Chartered Accountants. Right. Yeah, something like that. So I feel like they also have to review it as well as the technical person associated with it.

00:33:31 Speaker 1: OK, perfect. Thanks for the wrap up of this round. So if you don't have any other ideas to to bring in, I just say that you can swap over to the laptops again to

G.2 Transcript Mini Focus Group #17

(*Mini Focus Group 17 - online*; *Condition: AI Ethics Education*; *Participants: #1515, #1516*; *Moderator: Speaker 1*)

00:00:00 Speaker 1: What do you understand as the most important things to consider when setting up such a system?

00:02:11 Speaker 2: OK. First of all, I would see what is, what is the kind of the data you can access because I know it's not really ideally you would get, you would just use all the data you can use about the person, but I think there could be illegal according to some EU laws. Like I know they're like, even the state needs respect on privacy concerns when, when designing something like this. I know that because like in Italy they're talking about this. And I think even at the European Union level some days ago they talked about whether this could be done or not, and I feel like it's they concluded that you could do it. But there are like many concerns because, like the European rights chart that is called the European the chart of human rights to the European Union, something like that ensures that, like you have to, you have a right to privacy that is very strong. So it's unlikely they, the state could just use every information they can have on you to do this like technically, the ideal thing was even to like buy data from advertisers like to buy data from Google and Facebook to, to do that like there is a the, the Minister of like digital transition or something like that in Italy some days ago said that he wanted to scrape Instagram profiles and Facebook profiles to, to gather data to do AI tax fraud detection. But that's like that is not doable like the European laws and do not allow that. So, the first thing you have to do to make sure which data you can use which data you cannot use from an ethical, but especially in a legal point of view 00:03:59 Speaker 1: That's an important point. Do you have, yeah, I mean, additional comments on that?

00:04:05 Speaker 3: I guess something else I think we could consider is like the patterns that we're trying to recognize in order to identify specifically how someone is committing tax fraud. That needs to be very accurate and so I'm just wondering how, how do you define, like how do you define that pattern? Because I can just like already said, because there's some, there's so much information that's not legal that would be legal for us to derive because of privacy concerns. So what the passions that would be yielded through this, how accurate would they be since you don't get the full picture?

00:04:45 Speaker 1: So how would you proceed maybe for the data collection to tackle those things and then to collect the right data we have any like things that you would consider there?

00:05:00 Speaker 3: I guess maybe overlap of income, things like that. I mean, I, I don't know that much about tax fraud, like the, I don't know what exactly what are the identifiers we're using.

00:05:20 Speaker 1: So especially with the income, that's how that's mostly for the private people, right? Could you apply the same approach also for businesses?

00:05:37 Speaker 3: I guess so. I mean that probably depends as well, right on the type of business, what are the requirements? I guess that probably varies within, that probably varies from person to person as well and from household to household. So I guess there are a lot of factors that you need to consider in the model.

00:05:52 Speaker 1: Yeah. Do you see the same complexities for businesses as for private people?

00:06:01 Speaker 3: I guess it's possible but not too sure.

00:06:13 Speaker 2: I think maybe businesses are less legally protected from the point of view of like privacy other thing because like, I think privacy is thought to be like an individual right. So, I don't think it applies to businesses or it does not apply strongly as much as to businesses. Maybe it's like, it's you can, you can, like surveil them more. Like you can do, you can do more of a gathering around them and something you could, you should not be able to do about individuals. And also like, also the, about like responsibility of things in, if you, if you screw up with the tax fraud detection, you could ruin a personal life.fI you do, if you do that in with a business, you could also ruin people lives, but it's different. It's like more, not that direct because you're not dealing with black people lives individually, you're dealing with them in a different way. Because of course, like companies make people survive. But it's, it's a different thing than just going somebody 's life, because maybe you committed an error in tax fraud detection.

00:07:17 Speaker 3: Yeah, that's right. But I mean, even as an employee, you probably still have rights as to how much information the company is allowed to see anywhere. So would that maybe perhaps be the same even if it is an individual case, or if it's within a business, because those rights still apply even if you're just, even if you work at a company, right? They can't just have access to everything.

00:07:39 Speaker 1: So also the information about the people working there.

00:07:44 Speaker 3: Yeah, right. Yeah.

00:07:46 Speaker 1: It's also quite important point, right. Especially since individual behavior is hard, heavily linked to the act of tax fraud itself.

00:07:59 Speaker 3: Yeah, exactly.

00:08:06 Speaker 1: It's an interesting point, especially like how to use those links together to set up a model. Do you have any ideas and thoughts about how to do that, like which kind of models you would see is quite usable in this case?

00:08:37 Speaker 2: In what sense?

00:08:39 Speaker 1: Umm, so we could start like from the basic thing of differentiating between supervised and unsupervised. For example, what would you rather select for this kind of thing and where do you see the complexities when setting that up?

00:09:02 Speaker 3: We'll be trying.

00:09:02 Speaker 2: OK. Like.

00:09:02 Speaker 3: With abnormalities essentially. Right, I think that's where you start.

00:09:04 Speaker 1: Again?

00:09:07 Speaker 3: You will be trying to detect abnormalities within the pattern, I'm sorry, within the data.

00:09:14 Speaker 1: So that's the question. What would you rather prefer to have as an outcome?

00:09:22 Speaker 3: I'm not sure.

00:09:34 Speaker 2: I, I like, I think that the, the simplest method is the most the, the least likely, like that rights are, are being violated. Uh and I, I think like the most viable approach and also the simplest one to implement would be just some kind of like linear regression that like score, like that predicts people tax people like tax rate or people how much people should pay taxes based on other indicators like where do they live and other information than the state has access to and the people that have a wide, that have like very, very different tax being paid compared to what the model says, those people should be flagged like suspicious and they should manually. Anything that's the, the most essential model you can have. You can even implement that using a neural network, but still in this like basic way that you, you take a set of indicators and you try to predict how much taxes that person or that company should pay and you try to, if there is some discrepancies between what the model says and what the company is actually paying in taxes, but anything that is more complex than that, like some model that is like, like some feedback learning or some like modeling is like a lamps or model that uses like other kinds of more advanced or like maybe even like some clustering algorithm like those, those could have way more implications into violating people rights. Like I feel like the simpler, the simpler you make it the most likely you are to like just be in, inside ethical and legal boundaries.

00:11:26 Speaker 1: Why do you think so?

00:11:28 Speaker 2: Because like the, if you use more complex models you could be doing, you could be incurring like some kind of discrimination. I'm thinking, for example if you use some like clustering method and you do something like or like

AI cluster people based on like, where do they, where they live and, and some other things and if somebody should be in a cluster, but it's not paying taxes like people in in this cluster, Uh, that person is probably doing tax fraud, but that could lead to discrimination, that could easily lead to discrimination. Because like then you could end up using. OK, I'm pretty sure that in the US, they probably would do something that, you could do something like your ethnic group as a, as a way to like do clustering and things like that and that and that could lead to a lot of like biases or, or maybe, or maybe something, something easy you could do something like? Or if you do like clustering to do tax fraud detection, maybe fish like fishermen like people that fish like to, like to say something like that, that could, could work like people that fish in your truck are likely to be, likely to commit tax fraud and, and, and this means that the clustering algorithm would classify other people that do that job in new tracks are likely to the tax fraud and that could be a discrimination road. So that's like, yes, the point.

00:13:00 Speaker 1: It's a tricky thing.

00:13:02 Speaker 2: I feel like if you say it's on, or just a linear regression and we are like trying to account for like biases like gender and other protected categories I think it's easier to control.

00:13:17 Speaker 3: But there also maybe wouldn't there be more generalization if the model is very simple? Like, like you said with the fisherman example? Like if the algorithm picked it up once, it's going to detect the second person to supposedly do the same thing. So because people are complex, maybe we do need something more complex, but the model needs to be a bit more complex as well within 2 different cases, I don't know.

00:13:39 Speaker 2: But because like in the, because in the clustering you're trying to like profile people by their group, like you're trying to group people together and use that as an information to predict things. But if you just do a supervised linear regression the, there could be biased because for example like some, some one hot encoded trades could lead to biases like if you, if you put as a, as a parameter being a feature menu track that could lead to biases, but I think it's easier to control for biases into linear, linear regression model. Because it's like, it's a fairly simple model, so you can, you can, if like groups of people are being discriminating in a way and also I, I think it's easier to defend yourself by saying, oh, it's a very simple model, we are not grouping people together. The only thing that is changing it is how you train the model, so the training, like only, the only thing would be the training data and in this case the training data would be like past tax fraud, like people, like people that have been detected committing tax fraud and people they have not committed tax fraud and or also just technically, you could train the model even without considering tax fraud because you can train the model in real tax data, so like you should do this job and you live in this area and you have this income, you should pay this taxes. And you can just train the model of, on data that doesn't have tax fraud in it, and then you just use the model to see where to, to find discrepancies. And in this way you don't even have to use data from people that have committed tax fraud to, to train the model, and I feel like that's slightly more ethical because maybe like people that they've committed tax fraud would not be happy to have their data be used to detect other tax fraud people. And like, it's not, it's not a very good argument, but I think it's also a legal argument that it's about it. Like, that's like tax fraud people could be like: No, I don't want, I don't want my data to be used to, to

00:15:50 Speaker 3: Of course.

00:15:51 Speaker 2: train a model that could like you know, just need more people.

00:15:54 Speaker 1: Yeah. Yeah, that's an interesting point.

00:15:56 Speaker 2: But you just do like a very simple model you don't need the.

00:16:00 Speaker 3: Yeah, that's right.

00:16:04 Speaker 1: Well, if we assume still that we like, we're able to overcome those strikes and have a model that is working. How would you proceed for the rollout phase, like how do you, what factors would you consider when implementing this system both in the work flows at the Financial Investigation Office, but also into the judicial processes?

00:16:41 Speaker 3: Well, specific laws as well of that area or country, region that we're looking at in the traditional process, Business schools of the country.

00:17:01 Speaker 1: What do you mean by that?

00:17:02 Speaker 3: Like specific to the business, whatever the goals are, I guess the data would depend on that, like how much information, what type of information they're trying to get.

00:17:07 Speaker 1: OK. Mhm, true.

00:17:21 Speaker 2: I would like propose to like sample random people, no, we don't, you don't know the name of and and run the model on them to like have some test cases to see how the model performs. But I would also use that to establish the practice of how you would use the model, and I feel like the practice of how you should use it, they should, there should be something like that the model analyzes people with tax returns based on their other indicators, and if the model thinks that the tax returns are not the ones that should like, they don't line up with how much people should be paying taxes and then they should flag those people as suspicious and somebody should manually review them, especially in the testing phase. And uh, and if the manual reviewers think that the model is OK you can, and like the model could actually be right, you could proceed in, you could proceed in putting the model like in implementing the model in a higher scale and and the mean, in the meanwhile the people in the test have been found to commit to a, likely committed tax fraud should be referred to the judicial side and you can present the model as some evidence. I wouldn't present the, what the model says as like definitely that evidence I would just use it as like some, as a one hint of tax fraud, but I still think the, the manually review, the manually review people should be the damning evidence of the tax code, not the model itself. But it should, it could be one indicator. The commit, like judges or the like the legal side, proceeding as a person and on the, on the wider scale if the model like like survived this test run and, and people decide to apply it on a wider scale, I would still, I would hire more manual reviewers to, that should be trained on, on checking what the model says in order to check the people that, that have been flagged the suspicious by the model because I don't think that the model should be taking decision. I think the model should flag people for manual review, but they should not be the ultimate, like you should not be, you should not be sent a letter by the state that you have commit tax fraud just because the model said. There should always be one human that reviews what the model says and and decides whether there is, there are grounds to proceed or not.

00:19:54 Speaker 3: I agree with that as well. I don't think the model alone should be the only thing working in the process and making the decisions for sure, yeah.

00:20:03 Speaker 1: And that makes it, yeah, maybe a little easier to, to outbalance the complexities that we talked about already. Yeah. So, how would you proceed to measure of the impact, like the overall impact of such a system afterwards, and whether you designed a good system or not?

00:20:32 Speaker 3: You could compare it maybe with other system that's already in use, right. That is working well, or maybe the one that uses AI or not, but we can just compare that to one that is efficient and is giving us, is reliable basically. Maybe you can compare that with that, but maybe a specific area we can just use both systems to see what information they give out and compare.

00:21:08 Speaker 2: I agree on that.

00:21:13 Speaker 3: So basically we can only measure it by if it actually got the right people. Otherwise, we won't.

00:21:20 Speaker 1: That's a good point. OK. So if we like, look back on all the things that we just talked about you could, I don't know whether you have opened the questionnaire right now on the page with the discussion.

00:21:42 Speaker 2: OK.

00:21:44 Speaker 3: I'm like at the page where it's like this, with those the cartoon.

00:21:48 Speaker 2: The page, the page that like titled discussion?

00:21:51 Speaker 1: Yeah, the page that is titled discussion. There's a link to, uh, a pad sitting here and.

00:21:58 Speaker 2: Yeah, yeah. OK.

00:22:03 Speaker 1: So, I like you to shortly think about the 3 most significant complexities that you see or the 3 most significant challenges that you see when setting up such a system. You can just write it down for yourself, and afterwards we can shortly discuss it. Just cause your thoughts. Does it work for you too?

00:24:06 Speaker 2: Yeah, I think I'm, I think I, I have them.

00:24:28 Speaker 3: I think I'm just gonna write the privacy issues, but.

00:24:32 Speaker 1: OK, I think you could also just put it in there probably. So you both have 3 things. So who wants to shortly go over their aspects first?

00:25:15 Speaker 3: I don't mind going. Firstly, I guess for any model to be, to be efficient, it needs to be, it needs to yield accurate and reliable results. So, for that basically the system needs to detect the right people that are committing tax fraud. So there needs to be a way. If we don't get that then the system is not working. So that can be done by measuring that with the system that is already in place and is working, and maybe we can use that to compare results are reliable and our results are valid. Uh. Secondly, just using an AI system to make the whole decision process, I think poses some extra concerns because it could lead to generalize results prices. So, I don't think that, I think it should be a more balanced approach where there's also humans present in the decision making process. I think that's very important. And yeah, thirdly, just that in order to get all this information, it's possible that some peoples rights might get violated. So, I think regional individual rights based on the places need to be considered, heavily considered.

00:26:32 Speaker 1: OK. Do you want to add your points as well and shortly talk about it?

00:26:43 Speaker 2: Yeah, but it was mostly the things we talked about before like. First of all, the, the thing is you have to make sure that the data you're using doesn't cross boundaries and the legal framework of like what are individual privacy rights and also escalate to make sure that also you're not doing bad things. In, from a point of view, like violating people's privacy and then I would make sure that a responsibility gap is avoided by ensuring that decision making is always made by humans and manually review what the model says instead of letting the model decide by itself, because you cannot blame a model, but you can blame people. So, it's, it's important to have the people taking responsibility and making sure that the model is being used just as a tool and not as a decision maker and then there is the last point that is like a more it's a more pragmatic point related to the second one. You have to make sure that the model is affecting is worth implementing it so the, the state is gaining money from it, like the state is recouping tax fraud money from it even if like that in the ideal case you just implement the model, you don't have to hire more people. You can fire most of your tax Fraud Department office. So the model makes a lot of people unemployed because the the AI model is more efficient and cost less than people. But that I don't. I hope that doesn't happen because I, I think that people should be responsible for this. So I think that the state should not be hiring less people because of this model. You have to make a model that is even more efficient and more productive because you have to make the model be worth it. Despite the model is not making you save money on other things. Like the, on the human side of the, of the process. So, the model should be good enough it's worth implementing it even if you're not paying less people. You could actually be paying more people because you have to pay AI engineers to make it more.

00:28:48 Speaker 1: Yeah. True. It's an interesting point that you also brought this to the table so yeah. Actually good to consider that as well if you would though like after all those complexities and probably issues if you had the chance to speak to one of the managers of such a project, uhm. What you would you give them as advice to follow up and maybe end this discussion on the positive note?

00:29:21 Speaker 2: I would, I would tell them something that it was about the other day. There are some like data sets that are kinda unexpected that exist, but they exist. And not many people know about it. And because they're like some things are like according to some of those, some things need to be public like mandatorily and I don't know if the same case in Germany, but for example I found that, that in Italy when you sell a building like some real estate thing, the contract of the real estate like of selling that piece of real estate must be public. And I found out that there is a public data set. There is a public data set which buildings are being sold for what? And it is constantly refreshed because it's, it is constantly like every time you sell a building that should go into the data set and it is well done. And I think that, that would be the easiest. Like if that data set like that exist. If there are other devices like this, there are the easiest starting point to, to detect tax fraud. Because the, for selling real estate, you can just see which with like if somebody tells real estate for a certain price in one area and the next week is sold for like half the price. That, that building it was probably sold for a lower price just to pay less taxes than. And that's you. You can just very, very easily see which, which houses were sold for way too, way too little in a certain area and that's tax fraud detection.

00:31:00 Speaker 1: That's a good point, good starting point.

00:31:02 Speaker 2: Like this actually like there are many data sets that like people like, that need to be legally public, but very few people know about them and I would use the, I would search for those that are good starting point for a tax fraud detection thing. And it is also a bit outrageous and nobody has done this before also but because for, for some of those that's, it's, it's like anybody knows anything about machine learning would think that this is like a way to reduce tax fraud.

00:31:40 Speaker 1: What would your advice be?

00:31:47 Speaker 3: I think mostly just elaborating on the last point that I made, which is that I think there needs to be a good balance of human and both human aspect and tech aspects in this process, I think. Personally, at least I think that, I think that the process, the model needs to be trained a lot of times so that we make sure that we are getting the right information and not to just rely on data completely. I think that it needs to go through some cycles for us to know that it's reliable. And not just be like, oh, well, the model said that this person committed tax fraud so that's just the end of. I think that's a very dangerous spot. Yeah, we need people that are maybe, uh, both professional on the ethical side as well as AI and both working on this, I think that's something the manager needs to consider and not just have tech people working in this.

00:32:43 Speaker 1: OK. I think that's an really interesting point by the end.

00:32:47 Speaker 3: We're just promoting our degree basically.

00:32:50 Speaker 1: Right. Yeah. I mean somehow how we also need a job.

00:32:54 Speaker 3: Yeah, we need a job.

00:32:57 Speaker 1: Yeah, I really like that as a, as an ending point for the discussion, do you have anything else to add?

00:33:06 Speaker 3: No, not for now.

00:33:12 Speaker 1: Or is there anything by you?

00:33:14 Speaker 2: No, no, I think, uh, we said basically anything.

00:33:19 Speaker 1: OK. Yeah. So, thanks for the good ideas. And yeah, diving deep on this. To understand a little more how you tackle this the questionnaire on the following page is is first focused on the case study and afterwards shifts a little more to this more general perspective and your personal perception of AI. So, if you want you can swap back. And, yeah, just tell me if you have, if any questions come up in between. Thank you. I think I can stop the recording here as well.

Abbreviations

ABG Attitude-Behavior Gap. 1, 19, 30, 31, 42, 45, 46, 60–64, 73–81

AI Artificial Intelligence. v–vii, 18, 19, 24, 47–80

AIEE AI Ethics Education. 2, 33–36, 41–51, 58–60, 65–73, 75–81

CSS Computational Social Systems. 36, 43, 78

DADS Data Analytics and Decision Science. 36, 43, 78

EAIG Ethical AI Guideline. 7, 8, 12, 21, 76

EAIP Ethical AI Principle. v, vi, 1, 7, 14, 19–21, 24, 31, 32, 34, 37–40, 44, 45, 48, 50–55, 57–62, 64, 65, 68, 70–81

FAT Fairness, Accountability, and Transparency. 7, 8, 16

nAIEE no AI Ethics Education. 2, 33, 35, 36, 41–46, 48–51, 58–60, 65, 68–74, 76

OSF Open Science Framework. 31

RWTH Aachen University Rheinisch-Westfälische Technische Aachen University. 32, 33, 35, 36, 42, 43, 45, 78

STEM Science, technology, engineering, and mathematics. 13, 14, 24

TPB Theory of Planned Behavior. 19, 20, 70, 72, 75, 76

XAI Explainable AI. 7, 8, 10, 15

Bibliography

(2020a). Apa style jars: Jars–mixed | table 1. https://apastyle.apa.org/jars. Last accessed on 26.06.2024. American Psychological Association.

(2020b). Apa style jars: Jars–quant | table 1. https://apastyle.apa.org/jars. Last accessed on 26.06.2024. American Psychological Association.

(2020c). *Publication manual of the American Psychological Association*. American Psychological Association, Washington, 7 edition.

AI HLEG (2019). High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, 6.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

Ajzen, I. (2014). Attitude structure and behavior. In *Attitude structure and function*, pages 241–274. Psychology Press.

Alhazmi, A. and Arachchilage, N. A. G. (2021). I'm all ears! Listening to software developers on putting GDPR principles into software development practice. *Personal and Ubiquitous Computing*, 25(5):879–892.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805.

Almasri, N. and Tahat, L. (2018). Ethics vs IT ethics: A comparative study between the USA and the Middle East. *Journal of academic ethics*, 16(4):329–358.

Andrade, C. (2019). The p value and statistical significance: misunderstandings, explanations, challenges, and alternatives. *Indian journal of psychological medicine*, 41(3):210–215.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Applied Ethics (2023). General courses. https://www.ethics.rwth-aachen.de/cms/ethics/studium/~spyb/allgemeine-lehrveranstaltungen/?lidx=1. Last accessed on 27.06.2024.

Araujo, T., Helberger, N., Kruikemeier, S., and De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society*, 35:611–623.

Arizon-Peretz, R., Hadar, I., Luria, G., and Sherman, S. (2021). Understanding developers' privacy and security mindsets via climate theory. *Empirical Software Engineering*, 26:1–43.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115.

Asaro, P. (2019). What is an artificial intelligence arms race anyway. *ISJLP*, 15:45.

Aula, V. and Bowles, J. (2023). Stepping back from Data and AI for Good–current trends and ways forward. *Big Data & Society*, 10(1):20539517231173901.

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

Bada, M., Sasse, A. M., and Nurse, J. R. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672*.

Balebako, R., Marsh, A., Lin, J., Hong, J., and Cranor, L. F. (2014). The privacy and security behaviors of smartphone app developers. In *Workshop on Usable Security*, pages 1–10. Citeseer.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and social psychology review*, 3(3):193–209.

Barreno, M., Nelson, B., Joseph, A. D., and Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, 81:121–148.

Barretto, D., LaChance, J., Burton, E., and Liao, S. N. (2021). Exploring why underrepresented students are less likely to study machine learning and artificial intelligence. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, pages 457–463.

Bell, A., Nov, O., and Stoyanovich, J. (2023). Think about the stakeholders first! toward an algorithmic transparency playbook for regulatory compliance. *Data & Policy*, 5:e12.

Bell, A., Solano-Kamaiko, I., Nov, O., and Stoyanovich, J. (2022). It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 248–266.

Ben Schultz (2024). Supplementary material: I learned i ought, but will i? assessing future developers' perception of ethical ai principles in the context of teaching for ai ethics literacy. *Master Thesis Research Report*. Last accessed on 01.07.2024.

Bernnat, R., Mükusch, C., Schlemmer, M., Dahlen, E., and Tschiedel, M. (2023). AI Hype oder Game Changer? https://www.strategyand.pwc.com/de/de/industrie-teams/oeffentlicher-sektor/ki-hype-oder-game-changer.html.

Binns, R. and Gallo, V. (2019). Data minimisation and privacy-preserving techniques in ai systems. *Retrieved May*, 26:2020.

Blake, J. (1999). Overcoming the 'value-action gap' in environmental policy: Tensions between national policy and local experience. *Local environment*, 4(3):257–278.

Borenstein, J., Drake, M. J., Kirkman, R., and Swann, J. L. (2010). The engineering and science issues test (ESIT): A discipline-specific approach to assessing moral judgment. *Science and Engineering Ethics*, 16:387–407.

Brey, P. and Dainow, B. (2023). Ethics by design for artificial intelligence. *AI and Ethics*, pages 1–13.

Brezina, T. and Piquero, A. R. (2007). Moral beliefs, isolation from peers, and abstention from delinquency. *Deviant Behavior*, 28(5):433–465.

Brierley, C., Arief, B., Barnes, D., and Hernandez-Castro, J. (2021). Industrialising blackmail: Privacy invasion based IoT ransomware. In *Nordic Conference on Secure IT Systems*, pages 72–92. Springer.

Brittain, B. (2024). OpenAI hit with new lawsuits from news outlets over AI training. https://www.reuters.com/legal/litigation/openai-hit-with-new-lawsuits-news-outlets-over-ai-training-2024-02-28/. Last accessed on 20.05.2024.

Brown, N., Xie, B., Sarder, E., Fiesler, C., and Wiese, E. S. (2024). Teaching ethics in computing: A systematic literature review of acm computer science education publications. *ACM Transactions on Computing Education*, 24(1):1–36.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Bullock, B. B., Nascimento, F. L., and Doore, S. A. (2021). Computing ethics narratives: Teaching computing ethics and the impact of predictive algorithms. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 1020–1026.

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making*, 33(2):220–239.

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public administration review*, 81(5):825–836.

Byrne, G. J. and Staehr, L. J. (2004). The evaluation of a computer ethics program. *Journal of Issues in Informing Science and Information Technology*, 1:935–939.

Campolo, A., Sanfilippo, M. R., Whittaker, M., and Crawford, K. (2017). AI Now 2017 Report. Annual report, AI Now Institute at New York University, New York.

Cech, E. A. (2014). Culture of disengagement in engineering education? *Science, Technology, & Human Values*, 39(1):42–72.

Center for Advanced Internet Studies (2024). Meinungsmonitor Künstliche Intelligenz – Bevölkerungsbefragung: Dashboard zur Bevölkerungsbefragung. `https://www.cais-research.de/forschung/memoki/memoki-bevoelkerungsbefragung/`. Last accessed on 03.06.2024.

Center for Artificial Intelligence (2022). Lectures. `https://www.ai.rwth-aachen.de/cms/ki/ausbildung/~ggomk/lehrveranstaltungen/?lidx=1`. Last accessed on 27.06.2024.

Chair of Data and Business Analytics (2023). Modules. `https://www.analytics.rwth-aachen.de/cms/analytics/studium/~zniho/lehrveranstaltungen/?lidx=1`. Last accessed on 27.06.2024.

Chang, M. K. (1998). Predicting unethical behavior: a comparison of the theory of reasoned action and the theory of planned behavior. *Journal of business ethics*, 17(16):1825–1834.

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., and Baum, S. D. (2021). AI certification: Advancing ethical practice by reducing information asymmetries. *IEEE Transactions on Technology and Society*, 2(4):200–209.

Cornell University (2024). arxiv.org e-print archive. `https://arxiv.org`. Last accessed on 01.07.2024.

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., et al. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10).

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., and Chaintreau, A. (2020). Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 679–681.

Cramer, J. and Toll, B. (2012). Beyond competency: a context-driven CSO course. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education*, pages 469–474.

Crawford, K., Whittaker, M., Elish, M. C., Barocas, S., Plasek, A., and Ferryman, K. (2016). The AI now report. *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*, 2.

Delacre, M., Lakens, D., and Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, 30(1):92–101.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62.

Diggelmann, O. and Cleis, M. N. (2014). How the right to privacy became a human right. *Human Rights Law Review*, 14(3):441–458.

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Haim, G., Kließ, M. S., Lopez-Sanchez, M., et al. (2018). Ethics by design: Necessity or curse? In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 60–66.

Dilhac, M.-A., Abrassart, C., and Voarino, N. (2018). Montreal Declaration for a Responsible Development of Artificial Intelligence. Report, Université de Montréal, Montreal, Canada. Accessed: 21.06.2024.

Discord Inc. (2024). discord. `https://discord.com/`. Last accessed on 01.07.2024.

Domingo, R. M. (2024). *Reimagining intersectional AI fairness: A decolonial feminist approach to mitigating auto-essentialization in facial recognition technologies*. Phd thesis, De La Salle University, Manila, Philippines. Accessed: 21.06.2024.

Domínguez Figaredo, D. and Stoyanovich, J. (2023). Responsible AI literacy: A stakeholder-first approach. *Big Data & Society*, 10(2):20539517231219958.

Došenović, P., Kieslich, K., and Marcinkowski, F. (2022). Methodensteckbrief: Monitor- und sonderbefragungen. Technical report, Universität Düsseldorf, Düsseldorf, Germany. Gefördert von der Stiftung Mercator.

Duffy, C. (06.04.2023). 'it's an especially bad time': Tech layoffs are hitting ethics and safety teams. `https://edition.cnn.com/2023/04/06/tech/tech-layoffs-platform-safety/index.html`. Last accessed on 20.05.2024.

Durov, P. (2024). telegram. `https://telegram.org/`. Last accessed on 01.07.2024.

Dutta, D., Mishra, S. K., and Budhwar, P. (2022). Ethics in competency models: A framework towards developing ethical behaviour in organisations. *IIMB Management Review*, 34(3):208–227.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human factors*, 44(1):79–94.

D'Acquisto, G. (2020). On conflicts between ethical and logical principles in artificial intelligence. *AI & SOCIETY*, 35(4):895–900.

Ellemers, N., Van Der Toorn, J., Paunov, Y., and Van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, 23(4):332–366.

EU, E. (2012). Charter of fundamental rights of the european union. *The Review of International Affairs*, 63(1147):109–123.

Exter, M. E. and Ashby, I. (2019). Preparing today's educational software developers: voices from the field. *Journal of Computing in Higher Education*, 31(3):472–494.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2020). G*Power version 3.1.9.6. https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., and Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and engineering ethics*, 26(6):3333–3361.

Fenneman, A., Sickmann, J., Pitz, T., and Sanfey, A. G. (2021). Two distinct and separable processes underlie individual differences in algorithm adherence: Differences in predictions and differences in trust thresholds. *Plos one*, 16(2):e0247084.

Fidelangeli, A., Galli, F., et al. (2021). Artificial intelligence and tax law: Perspectives and challenges. *CERIDAP*, 4(Ottobre-Dicembre):24–58.

Field, H. and Vanian, J. (27.05.2023). Tech layoffs ravage the teams that fight online misinformation and hate speech. https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html. Last accessed on 20.05.2024.

Fiesler, C., Friske, M., Garrett, N., Muzny, F., Smith, J. J., and Zietz, J. (2021). Integrating ethics into introductory programming classes. In *Proceedings of the 52nd ACM technical symposium on computer science education*, pages 1027–1033.

Fiesler, C., Garrett, N., and Beard, N. (2020). What do we teach when we teach tech ethics? A syllabi analysis. In *Proceedings of the 51st ACM technical symposium on computer science education*, pages 289–295.

Finelli, C. J., Holsapple, M. A., Ra, E., Bielby, R. M., Burt, B. A., Carpenter, D. D., Harding, T. S., and Sutkus, J. A. (2012). An assessment of engineering students' curricular and co-curricular experiences and their ethical development. *Journal of Engineering Education*, 101(3):469–494.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707.

Freiesleben, T. and Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109.

Garrett, N., Beard, N., and Fiesler, C. (2020). More than" If Time Allows" the role of ethics in AI education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 272–278.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., and Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115:106607.

Gilligan, C. (1993). *In a different voice: Psychological theory and women's development*. Harvard university press.

Gillon, R. (1985). " Primum non nocere" and the principle of non-maleficence. *British medical journal (Clinical research ed.)*, 291(6488):130.

Gilpin, L. H., Paley, A. R., Alam, M. A., Spurlock, S., and Hammond, K. J. (2022). " Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI. *arXiv preprint arXiv:2207.00007*.

Gino, F., Ayal, S., and Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological science*, 20(3):393–398.

Graham, R. (2007). Theory of cognitive dissonance as it pertains to morality. *Journal of Scientific Psychology*, 29.

Griffin, T. A., Green, B. P., and Welie, J. V. (2024). The ethical wisdom of AI developers. *AI and Ethics*, pages 1–11.

Gröger, C. (2021). There is no AI without data. *Communications of the ACM*, 64(11):98–108.

Grosan, C., Abraham, A., Grosan, C., and Abraham, A. (2011). Rule-based expert systems. *Intelligent systems: A modern approach*, pages 149–185.

Grynbaum, M. M. and Mac, R. (27.12.2023). The times sues openai and microsoft over a.i. use of copyrighted work: Dec. 27, 2023. `https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html`. Last accessed on 20.05.2024.

Hadar, I., Hasson, T., Ayalon, O., Toch, E., Birnhack, M., Sherman, S., and Balissa, A. (2018). Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*, 23:259–289.

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and machines*, 30(1):99–120.

Harding, T. S., Carpenter, D. D., and Finelli, C. J. (2013). Two years later: A longitudinal look at the impact of engineering ethics education. In *2013 ASEE Annual Conference & Exposition*, pages 23–1272. Taylor Francis.

Harris, A. L. (2000). Is ethical attitudes among college students: A comparative study. In *Proceedings of ISECON*, volume 200, pages 801–807. Citeseer.

Hedayati-Mehdiabadi, A. (2022). How do computer science students make decisions in ethical situations? implications for teaching computing ethics based on a grounded theory study. *ACM Transactions on Computing Education (TOCE)*, 22(3):1–24.

Hess, J. L. and Fore, G. (2018). A systematic literature review of us engineering ethics interventions. *Science and engineering ethics*, 24:551–583.

Hinds, J., Williams, E. J., and Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. *International Journal of Human-Computer Studies*, 143:102498.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. (2019). 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.

Hornsey, M. J., Majkut, L., Terry, D. J., and McKimmie, B. M. (2003). On being loud and proud: Non-conformity and counter-conformity to group norms. *British journal of social psychology*, 42(3):319–335.

Horowitz, M. C., Kahn, L., Macdonald, J., and Schneider, J. (2023). Adopting AI: how familiarity breeds both trust and contempt. *AI & society*, pages 1–15.

Horton, D., McIlraith, S. A., Wang, N., Majedi, M., McClure, E., and Wald, B. (2022). Embedding ethics in computer science courses: Does it work? In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*, pages 481–487.

IBM Corp. (2021). IBM SPSS Statistics for Windows. version 28.0.1.0.

Individual and Technology (2023). Offered courses. https://www.itec.rwth-aachen.de/cms/itec/studium/~sjce/lehrveranstaltungen/?lidx=1. Last accessed on 27.06.2024.

Individualized Production (2022). M.Sc. Construction and Robotics. https://www.ip.rwth-aachen.de/teaching/. Last accessed on 27.06.2024.

Institut für Kraftfahrzeuge (2024). Lehrveranstaltungen. https://www.ika.rwth-aachen.de/de/studium/lehrveranstaltungen.html. Last accessed on 27.06.2024.

International Telecommunication Union (2017). AI for good global summit 2017. Technical report, International Telecommunication Union.

Isaac, E. R. and Reno, J. (2023). AI Product Security: A Primer for Developers. *arXiv preprint arXiv:2304.11087*.

Jain, P., Gyanchandani, M., and Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, 3:1–25.

Jay, J. A., D'Auria, R., Nordby, J. C., Rice, D. A., Cleveland, D. A., Friscia, A., Kissinger, S., Levis, M., Malan, H., Rajagopal, D., et al. (2019). Reduction of the carbon footprint of college freshman diets after a food-based environmental science course. *Climatic Change*, 154:547–564.

Jeffrey Dastin (11.10.2018). Insight - amazon scraps secret ai recruiting tool that showed bias against women. `https://www.reuters.com/article/idUSKCN1MK0AG/`. Last accessed on 30.05.2024.

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

John-Mathews, J.-M., Cardon, D., and Balagué, C. (2022). From reality to world. A critical perspective on AI fairness. *Journal of Business Ethics*, 178(4):945–959.

Jui, T. D. and Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, pages 1–31.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *Proceedings of the 2020 European Conference on Information Systems*. Association for Information Systems.

Kamberelis, G. and Dimitriadis, G. (2005). Focus groups: Strategic articulations of pedagogy, politics, and inquiry. In Denzin, N. K. and Lincoln, Y. S., editors, *The Sage handbook of qualitative research*, chapter 36, pages 887–907. Sage Publications Ltd, 3 edition.

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., and Huber, P. (2016). Artificial intelligence and computer science in education: From kindergarten to university. In *2016 IEEE Frontiers in Education Conference (FIE)*, pages 1–9.

Kang, C. (14.11.2022). Google agrees to $392 million privacy settlement with 40 states. `https://www.nytimes.com/2022/11/14/technology/google-privacy-settlement.html`. Last accessed on 20.05.2024.

Kasinidou, M., Kleanthous, S., Orphanou, K., and Otterbacher, J. (2021). Educating computer science students about algorithmic fairness, accountability, transparency and ethics. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, pages 484–490.

Kelly, A. (2021). A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, 47(3):725–741.

Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., and Abrahamsson, P. (2023). AI ethics: an empirical study on the views of practitioners and lawmakers. *IEEE Transactions on Computational Social Systems*.

Kieslich, K., Keller, B., and Starke, C. (2022). Artificial intelligence ethics by design. evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society*, 9(1):20539517221092956.

Kilkenny, M., Hovey, C. L., Robledo Yamamoto, F., Voida, A., and Barker, L. (2022). Why should computer and information science programs require service learning? In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*, pages 822–828.

Kiran, A. H., Oudshoorn, N., and Verbeek, P.-P. (2015). Beyond checklists: toward an ethical-constructive technology assessment. *Journal of responsible innovation*, 2(1):5–19.

Kleanthous, S., Kasinidou, M., Barlas, P., and Otterbacher, J. (2022). Perception of fairness in algorithmic decisions: future developers' perspective. *Patterns*, 3(1).

Klein, N. and O'Brien, E. (2016). The tipping point of moral change: When do good and bad acts make good and bad actors? *Social cognition*, 34(2):149–166.

Knoth, N., Decker, M., Laupichler, M. C., Pinski, M., Buchholtz, N., Bata, K., and Schultz, B. (2024). Developing a holistic AI literacy assessment matrix – Bridging generic, domain-specific, and ethical competencies. *Computers and Education Open*, 6:100177.

Köchling, A. and Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. *Business Research*, 13(3):795–848.

Kok, J. N., Boers, E. J., Kosters, W. A., Van der Putten, P., and Poel, M. (2009). Artificial intelligence: definition, trends, techniques, and cases. *Artificial intelligence*, 1(270-299):51.

Kolter, Z. and Madry, A. (2018). Adversarial robustness: Theory and practice. *Tutorial at NeurIPS*, page 3.

Kong, S.-C., Cheung, W. M.-Y., and Tsang, O. (2023a). Evaluating an artificial intelligence literacy programme for empowering and developing concepts, literacy and ethical awareness in senior secondary students. *Education and Information Technologies*, 28(4):4703–4724.

Kong, S.-C., Cheung, W. M.-Y., and Zhang, G. (2023b). Evaluating an artificial intelligence literacy programme for developing university students' conceptual understanding, literacy, empowerment and ethical awareness. *Educational Technology & Society*, 26(1):16–30.

Kreth, Q., Schiff, D., Lee, J., Zegura, E., and Borenstein, J. (2022). Social responsibility attitudes among undergraduate computer science students: an empirical analysis. In *2022 ASEE Annual Conference & Exposition*. National Science Foundation.

Kwasny, T., Dobernig, K., and Riefler, P. (2022). Towards reduced meat consumption: A systematic literature review of intervention effectiveness, 2001–2019. *Appetite*, 168:105739.

Laupichler, M. C., Aster, A., Schirch, J., and Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 3:100101.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

Leiner, D. J. (2024). Sosci survey. https://www.soscisurvey.de. version 3.5.07, Last accessed on 01.07.2024.

Leitgeb, H. (2009). On formal and informal provability. In *New waves in philosophy of mathematics*, pages 263–299. Springer.

LinkedIn Ireland Unlimited Company (2024). linkedin. https://www.linkedin.com/. Last accessed on 01.07.2024.

Liu, H. and Priest, S. (2009). Understanding public support for stem cell research: media communication, interpersonal communication and trust in key actors. *Public Understanding of science*, 18(6):704–718.

Liu, P., Du, Y., and Xu, Z. (2019). Machines versus humans: People's biased responses to traffic accidents involving self-driving vehicles. *Accident Analysis & Prevention*, 125:232–240.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Long, D. and Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA. Association for Computing Machinery.

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access*, 7:154096–154113.

Lu, Q., Zhu, L., Xu, X., Whittle, J., Douglas, D., and Sanderson, C. (2022). Software engineering for responsible ai: An empirical study and operationalised patterns. In *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice*, pages 241–242.

Lyell, D. and Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431.

Lyles, T. (11.03.2024). Nvidia Sued for AI Tech Copyright Infringement by Three Authors. https://www.ign.com/articles/nvidia-sued-for-ai-tech-copyright-infringement-by-three-authors. Last accessed on 20.05.2024.

Machanavajjhala, A., Korolova, A., and Sarma, A. D. (2011). Personalized social recommendations-accurate or private? *arXiv preprint arXiv:1105.4254*.

Mahmud, H., Islam, A. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390.

Marsh, S. and Dibben, M. R. (2005). Trust, untrust, distrust and mistrust–an exploration of the dark (er) side. In *International conference on trust management*, pages 17–33. Springer.

Mäses, S., Aitsam, H., and Randmann, L. (2019). A method for adding cyberethical behaviour measurements to computer science homework assignments. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, pages 1–5.

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., and Perrault, R. (2023). The AI Index 2023 Annual Report. *arXiv preprint arXiv:2310.03715*.

McNamara, A., Smith, J., and Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 729–733.

Meesters, M., Heck, P., and Serebrenik, A. (2022). What is an AI engineer? an empirical analysis of job ads in The Netherlands. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, CAIN '22, page 136–144, New York, NY, USA. Association for Computing Machinery.

meetergo (2024). meetergo: Turbo für deine meetings. https://meetergo.com. Last accessed 30.06.2024.

Merritt, S. M. and Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50(2):194–210.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Microsoft Corporation (2019). Microsoft AI principles. https://www.microsoft.com/en-us/ai/our-approach-to-ai. Last accessed on 01.02.2019.

Mittal, B. (1988). Achieving higher seat belt usage: The role of habit in bridging the attitude-behavior gap 1. *Journal of applied social psychology*, 18(12):993–1016.

Mohanani, R., Salman, I., Turhan, B., Rodríguez, P., and Ralph, P. (2018). Cognitive biases in software engineering: A systematic mapping study. *IEEE Transactions on Software Engineering*, 46(12):1318–1339.

Mökander, J., Axente, M., Casolari, F., and Floridi, L. (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed european ai regulation. *Minds and Machines*, 32(2):241–268.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Munoko, I., Brown-Liburd, H. L., and Vasarhelyi, M. (2020). The ethical implications of using artificial intelligence in auditing. *Journal of business ethics*, 167(2):209–234.

Ng, D. T. K., Lee, M., Tan, R. J. Y., Hu, X., Downie, J. S., and Chu, S. K. W. (2023). A review of AI teaching and learning from 2000 to 2020. *Education and Information Technologies*, 28(7):8445–8501.

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., and Qiao, M. S. (2021a). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, 58(1):504–509.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., and Qiao, M. S. (2021b). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2:100041.

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. In *2017 15th international conference on ICT and knowledge engineering (ICT&KE)*, pages 1–6. IEEE.

Önkal, D., Gönül, M. S., and De Baets, S. (2019). Trusting forecasts. *Futures & Foresight Science*, 1(3-4):e19.

Osswald, S., Greitemeyer, T., Fischer, P., and Frey, D. (2010). What is moral courage? Definition, explication, and classification of a complex construct. In Pury, C. L. S. and Lopez, S. J., editors, *The psychology of courage: Modern research on an ancient virtue*, pages 149–164. American Psychological Association.

Pal, K. K., Li, R., Piaget, K., Baller, S., and Zahidi, S. (2023). Global gender gap report 2023. Technical report, World Economic Forum, Geneva. Last accessed on 27.05.2024.

Panas, E. E. and Ninni, V. E. (2011). Ethical decision making in electronic piracy: An explanatory model based on the diffusion of innovation theory and theory of planned behavior. *International Journal of Cyber Criminology*, 5(2).

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., and Turhan, B. (2024a). Ethics in the Age of AI: An Analysis of AI Practitioners' Awareness and Challenges. *ACM Trans. Softw. Eng. Methodol.*, 33(3).

Pant, A., Hoda, R., Tantithamthavorn, C., and Turhan, B. (2024b). Ethics in AI through the practitioner's view: a grounded theory literature review. *Empirical Software Engineering*, 29(3):67.

Park, H. J. and Lin, L. M. (2020). Exploring attitude–behavior gap in sustainable consumption: Comparison of recycled and upcycled fashion products. *Journal of business research*, 117:623–628.

Pearse, N. (2019). An illustration of deductive analysis in qualitative research. In *18th European conference on research methodology for business and management studies*, page 264.

Peixoto, M., Ferreira, D., Cavalcanti, M., Silva, C., Vilela, J., Araújo, J., and Gorschek, T. (2020). On understanding how developers perceive and interpret privacy requirements research preview. In *Requirements Engineering: Foundation for Software Quality: 26th International Working Conference, REFSQ 2020, Pisa, Italy, March 24–27, 2020, Proceedings 26*, pages 116–123. Springer.

Pickens, J. (2005). Attitudes and perceptions. *Organizational behavior in health care*, 4(7):43–76.

Pierson, E. (2017). Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*.

Pinski, M. and Benlian, A. (2023). AI Literacy - Towards Measuring Human Competency in Artificial Intelligence. In *Proceedings of the 56th Hawaii International Conference on System Sciences*, page 165, Honolulu, HI. University of Hawai'i at Manoa.

Portus, R., Aarnio-Linnanvuori, E., Dillon, B., Fahy, F., Gopinath, D., Mansikka-Aho, A., Williams, S.-J., Reilly, K., and McEwen, L. (2024). Exploring environmental value action gap and education research: a semi-systematic literature review. *Environmental Education Research*, pages 1–31.

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

Prybylo, M., Haghighi, S., Peddinti, S. T., and Ghanavati, S. (2024). Evaluating privacy perceptions, experience, and behavior of software development teams. *arXiv preprint arXiv:2404.01283*.

Pulfrey, C. and Butera, F. (2016). When and why people don't accept cheating: Self-transcendence values, social responsibility, mastery goals and attitudes towards cheating. *Motivation and Emotion*, 40(3):438–454.

Randall, D. M. and Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of business ethics*, 10:805–817.

Rao, A., Veillet, A., Kuperholz, M., Labovich, M., Cameron, E., and Ghosh, S. (2021). Responsible AI — Maturing from theory to practice. https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/pwc-responsible-ai-maturing-from-theory-to-practice.pdf. Last accessed on 01.07.2024.

Reimenschneider, C. K., Leonard, L. N., and Manly, T. S. (2011). Students' ethical decision-making in an information technology context: A theory of planned behavior approach. *Journal of Information Systems Education*, 22(3):203.

Rest, J. R., Narvaez, D., Thoma, S. J., and Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of educational psychology*, 91(4):644.

Rubinstein, A. (1998). *Modeling bounded rationality*. MIT press.

Rudy-Hiller, F. (2018). The epistemic condition for moral responsibility. *Stanford Encyclopedia of Philosophy*. Accessed: 21.06.2024.

Rulifson, G. and Bielefeldt, A. (2018). Influence of internships on engineering students' attitudes about socially responsible engineering. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–6. IEEE.

Rulifson, G. and Bielefeldt, A. R. (2019). Evolution of students' varied conceptualizations about socially responsible engineering: A four year longitudinal study. *Science and engineering ethics*, 25:939–974.

RWTH Aachen University (2024a). Computational Social Systems: Curriculum. https://computationalsocialsystems.rwth-aachen.de/. Last accessed on 25.06.2024.

RWTH Aachen University (2024b). M.Sc. Data Analytics & Decision Science. https://www.business-school.rwth-aachen.de/master/data-analytics-decision-science/. Last accessed on 25.06.2024.

Ryan, M. (2020). In ai we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5):2749–2767.

Sadeghi, B., Richards, D., Formosa, P., McEwan, M., and Hitchens, M. (2022). How to increase ethical awareness in cybersecurity decision-making. In *ACIS 2022 Proceedings*, pages 1–11, Melbourne, Australia. AIS Electronic Library (AISeL). 33rd Australasian Conference on Information Systems: The Changing Face of IS, ACIS 2022, 4-7 December 2022.

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkowicz, S., Robinson, C., and Hansen, D. (2023). AI ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society*.

Sanderson, C., Lu, Q., Douglas, D., Xu, X., Zhu, L., and Whittle, J. (2022). Towards implementing responsible AI. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5076–5081. IEEE.

Santoni de Sio, F. and Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15.

Saranya, A. and Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, page 100230.

Sartre, J.-P., Richmond, S., and Moran, R. (2022). *Being and nothingness: An essay in phenomenological ontology*. Routledge.

Schiefele, U. (1992). Topic interest and levels of text comprehension. *The role of interest in learning and development*, 1991:151–182.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., and Lennon, M. (2020). Principles to practices for responsible ai: Closing the gap.

Schleiss, J., Bieber, M., Manukjan, A., Kellner, L., and Stober, S. (2022). An interdisciplinary competence profile for ai in engineering. In *Towards a new future in engineering education, new scenarios that european alliances of tech universities open up*, pages 1601–1609. Universitat Politècnica de Catalunya.

Seppälä, A., Birkstedt, T., and Mäntymäki, M. (2021). From Ethical AI Principles to Governed AI. In *ICIS*.

Shih, P.-K., Lin, C.-H., Wu, L. Y., and Yu, C.-C. (2021). Learning ethics in ai—teaching non-engineering undergraduates through situated learning. *Sustainability*, 13(7):3718.

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization science*, 2(1):125–134.

Skirpan, M., Beard, N., Bhaduri, S., Fiesler, C., and Yeh, T. (2018). Ethics education in context: A case study of novel ethics activities for the CS classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 940–945.

Skitka, L. J. and Mullen, E. (2002). Understanding judgments of fairness in a real-world political context: A test of the value protection model of justice reasoning. *Personality and Social Psychology Bulletin*, 28(10):1419–1429.

Stappenbelt, B. (2013). Ethics in engineering: student perceptions and their professional identity. *JOTSE: Journal of technology and science education*, 3(1):3–10.

Stavrakakis, I., Gordon, D., Tierney, B., Becevel, A., Murphy, E., Dodig-Crnkovic, G., Dobrin, R., Schiaffonati, V., Pereira, C., Tikhonenko, S., et al. (2021). The teaching of computer ethics on computer science and related degree programmes. a european survey. *International Journal of Ethics Education*, pages 1–29.

Stoeber, J. and Yang, H. (2016). Moral perfectionism and moral values, virtues, and judgments: Further investigations. *Personality and Individual Differences*, 88:6–11.

Studydrive GmbH (2024). studydrive. https://www.studydrive.net//. Last accessed on 01.07.2024.

Sutton, C. and Gong, L. (2017). Popularity of arXiv. org within Computer Science. *arXiv preprint arXiv:1710.05225*.

Taddeo, M. and Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404):751–752.

Thurman, N., Moeller, J., Helberger, N., and Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital journalism*, 7(4):447–469.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., and Van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 272–283.

Ulman, M., Harris, A. L., Marreiros, C. G., Quaresma, R., and Ganiyev, M. (2019). What do college students do and think about IT ethics? *Proceedings of Learning Innova*, 2:49–58.

Vakkuri, V., Kemell, K.-K., and Abrahamsson, P. (2019a). Ethically aligned design: an empirical evaluation of the resolvedd-strategy in software and systems development context. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 46–50. IEEE.

Vakkuri, V., Kemell, K.-K., Kultanen, J., and Abrahamsson, P. (2020). The current state of industrial practice in artificial intelligence ethics. *Ieee Software*, 37(4):50–57.

Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., and Abrahamsson, P. (2019b). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv preprint arXiv:1906.07946*.

Van de Poel, I. and Sand, M. (2021). Varieties of responsibility: two problems of responsible innovation. *Synthese*, 198(Suppl 19):4769–4787.

van Stuijvenberg, O. C., Broekman, M. L., Wolff, S. E., Bredenoord, A. L., and Jongsma, K. R. (2024). Developer perspectives on the ethics of ai-driven neural implants: a qualitative study. *Scientific Reports*, 14(1):7880.

Verma, A., Lamsal, K., and Verma, P. (2022). An investigation of skill requirements in artificial intelligence and machine learning job advertisements. *Industry and Higher Education*, 36(1):63–73.

Vieira, J., Castro, S. L., and Souza, A. S. (2023). Psychological barriers moderate the attitude-behavior gap for climate change. *PloS one*, 18(7):e0287404.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11).

WhatsApp LLC (2024). whatsapp. `https://whatsapp.com/`. Last accessed on 01.07.2024.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., Schwartz, O., et al. (2018). *AI now report 2018*. AI Now Institute at New York University New York.

Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. (2019). The role and limits of principles in ai ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200.

Wiessner, D. (21.02.2024). Workday accused of facilitating widespread bias in novel AI lawsuit. https://www.reuters.com/legal/transactional/workday-accused-facilitating-widespread-bias-novel-ai-lawsuit-2024-02-21/. Last accessed on 20.05.2024.

Wong, R. Y., Boyd, K., Metcalf, J., and Shilton, K. (2020). Beyond checklist approaches to ethics in design. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, pages 511–517.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE access*, 4:2751–2763.

Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., and Bozzon, A. (2023). Disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Zhong, C.-B. and Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, 313(5792):1451–1452.

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., and Savage, S. (2020). A survey on ethical principles of AI and implementations. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 3010–3017. IEEE.

'Zoom Video Communications, Inc.' (2024). zoom. https://zoom.us/. Last accessed on 01.07.2024.

Typeset July 2, 2024