Researchers' Concerns on Artificial Intelligence Ethics: Results from a Scenario-Based Survey

Conference Paper · January 2024

CITATIONS
0

READS
98

6 authors, including:

Marianna Jantunen
University of Jyväskylä
17 PUBLICATIONS **166** CITATIONS
SEE PROFILE

Richard Meyes
Bergische Universität Wuppertal
34 PUBLICATIONS **266** CITATIONS
SEE PROFILE

Veronika Kurchyna
Deutsches Forschungszentrum für Künstliche Intelligenz
10 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Tobias Meisen
Bergische Universität Wuppertal
267 PUBLICATIONS **2,591** CITATIONS
SEE PROFILE

Researchers' Concerns on Artificial Intelligence Ethics: Results from a Scenario-Based Survey

Marianna Jantunen
marianna.s.p.jantunen@jyu.fi
University of Jyväskylä
Jyväskylä, Finland

Richard Meyes
meyes@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Veronika Kurchyna
veronika.kurchyna@dfki.de
German Research Center for Artificial Intelligence
Trier, Germany

Tobias Meisen
meisen@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Pekka Abrahamsson
pekka.abrahamsson@tuni.fi
University of Tampere
Tampere, Finland

Rahul Mohanani
rahul.p.mohanani@jyu.fi
University of Jyväskylä
Jyväskylä, Finland

ABSTRACT

The ethical impacts of Artificial Intelligence (AI) are causing concern in many areas of AI research and development. The implementation of AI ethics is still, in many ways, a work in progress, but various initiatives are tackling the issues by creating guidelines and implementation methods. This study investigates concerns about the negative impacts of AI systems posed by researchers working with AI. The study was conducted as a scenario-based survey, in which participants answered the question, "*What could go wrong?*" regarding five scenarios depicting fictional AI systems. The study concludes with the results from 33 survey participants who gave 161 responses to the scenarios. The results suggest that researchers can identify threats posed by AI systems, particularly regarding their social and ethical consequences. This is even though half of the participants reported limited involvement with AI ethics in their work. The widespread understanding of ethics among researchers could positively impact AI software development due to increased capabilities to bring theoretical AI ethics to practice.

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **Software and its engineering** → *Requirements analysis*; • **Social and professional topics** → *Computing / technology policy*.

KEYWORDS

Artificial Intelligence, AI Ethics, AI impacts, Qualitative study, Survey

ACM Reference Format:
Marianna Jantunen, Richard Meyes, Veronika Kurchyna, Tobias Meisen, Pekka Abrahamsson, and Rahul Mohanani. 2018. Researchers' Concerns on Artificial Intelligence Ethics: Results from a Scenario-Based Survey. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As artificial intelligence (AI) technologies develop rapidly and are being adopted into many processes, more people are affected by AI. Among them, not all will be experienced and well-informed on how AI affects them, increasing the importance of design decisions that consider these issues when developing AI systems. Questions on the impact of AI are studied in *AI ethics*, a research field with growing popularity, particularly in the last decade, mirroring the increased impact and adoption rate of AI technologies.

The theoretical nature of ethics remains a major obstacle in its application to AI development. However, awareness of ethical issues has increased along with the potential uses of AI [19]. Several stakeholders in AI research and development have expressed concerns about AI, especially in fields dealing with sensitive or impactful issues, such as healthcare [16] and military technology [18]. Despite the efforts of multiple stakeholders to bring conceptual AI ethics closer to practice, the endeavors do not always lead to results [13, 19]. Furthermore, in the business sector, there are varying levels of commitment to ethical thinking [2, 26].

Research and business are heavily intertwined in AI development [13], and researchers play an important role in the ethics and governance of AI [31]. Thus, it is essential to examine researchers as one of the groups impacted by the rapid development of AI. This study addresses the popularly used term *concern* as experienced by researchers following the development of the field of AI ethics. The concern is, as defined by the Merriam-Webster dictionary, "a matter that causes feelings of unease, uncertainty, or apprehension"[1]. Through concerns, this study looks into the current state of AI researchers' capabilities and views about the impacts of AI. The results provide an overview of the scope of ethics-related understanding among this important group of stakeholders in the AI-related software engineering (SE) atmosphere.

Several studies have been conducted to discover the views on AI ethics of target groups such as AI developers and practitioners [21, 26, 27], with the findings commonly pointing to a mismatch

[1] https://www.merriam-webster.com/dictionary/concern

of theory and practice in ethics considerations. The views of researchers regarding AI have been studied as well, e.g., from the perspectives of ethics and governance [31] and their beliefs regarding the progress of AI [12]. To accompany such large-scale studies, there is also in-depth qualitative research with small sample sizes, such as Rousi et al. [23], which inspired this work as a slightly larger, detailed qualitative investigation into the concerns of AI researchers, who are working in different research fields and with varying levels of familiarity with AI ethics. This study offers a perspective in the middle of the existing studies, investigating the types of threats most commonly detected by AI researchers.

This paper is part of a research process investigating the views of AI development stakeholders regarding the ethical impacts of AI to map the existing capabilities of ethical considerations of the people who are developing AI systems, continuing the research branch from an earlier paper about AI prototype developers [27]. The results of this research endeavor will enable the discovery of areas needing improvement regarding the ethical awareness of AI stakeholders. In the academic setting, it may also form a basis for improving the introduction of AI ethics in higher education.

Through the lens of AI ethics, we ask, how do AI researchers assess the impacts of AI systems? To find an answer, two central questions guide the study:

(1) *What are the concerns of researchers from diverse backgrounds regarding the impacts of AI systems?*
(2) *How do the concerns expressed by researchers relate to AI ethics?*

The results aim to pave the way to discover what to improve in the less-known areas of AI ethics. They could use more familiarization among AI researchers and what has already been accomplished. In light of the increasing adoption rate of AI into different commercial and private use cases, traditional SE and design issues, such as requirements engineering, are now extended with the need to consider ethical concerns when developing a product or service that utilizes AI. As such, awareness of common themes of AI ethics is an important first step towards implementing ethically responsible AI-powered systems.

2 BACKGROUND: AI ETHICS

Along with the advancement of AI technologies, the field of *AI ethics* (from here on shortened as *AIE*) has emerged both in research and in the software industry, examining the ethical implications of AI. Numerous calls to pay attention to how researchers, developers and regulators should deal with the ethical impacts of increasingly intelligent AI have arisen for years now (e.g., [6, 7, 24]). At the same time, AIE as science has emerged, dealing with questions such as the analysis of the AIE needs (e.g. [9, 13, 15]) and, in its application, suggestions on how to tackle the challenges of keeping AI aligned to human values (e.g., [1, 8, 19]).

The young field has faced challenges in reaching consensus on numerous questions. These challenges understandably begin with the elusive definition of Artificial Intelligence itself [29], as the term contains concepts and capabilities of artificial entities on a wide range of what is considered "autonomy" or "intelligence". As pointed out by Wang [29], many scientific concepts mature over time - and while AI as a concept is not new, the high speed at which new AI inventions are being adopted is rapidly shaping expectations even outside the traditional research field of AI. After all, the now ubiquitous ChatGPT[2] is still a relatively recent product, yet its use became rapidly widespread, prompting questions both exciting and uncomfortable in terms of issues such as how AI affects employment[3] and intellectual property[4], and regarding its possible exploitation and negative impacts[5]. This development also highlights a major issue researchers across disciplines face: AI advancements often outpace academic research and scientific publishing. The level of adoption of ethically oriented practices has been found to vary to a high degree between companies [26] and ethics are often under-addressed, possibly due to the competitive reality of AI development business, among other things [2]. Although avoiding reputational damage is not an insignificant issue due to the presence of business collaboration in AI research [13], stories of sacrificing ethical considerations in favor of fast technological advancement occasionally surface[6]. As such, guidelines on the ethical development and usage of AI are even more important to guide developers and users alike.

An important topic in the landscape of AIE is *trustworthy AI*. The need to establish trust in autonomous systems - generally referring to AI technologies - is a crucial issue brought up by the European Commission in the *Ethics Guidelines for Trustworthy AI* [8]. As stated by Floridi [10], the nature of AI technologies requires a new kind of ethical balance between human and artificial autonomy, as AI systems have the potential to have an enormous impact on our society.

In the discussion about AIE, at least one commonly recognized form of communicating ethical commitments has emerged, commonly titled *AI Ethics Guidelines* or *principles*. Guidelines are important, for ethical uncertainty breeds both "reckless risk-taking and excessive caution", as suggested by Floridi [10]. A common language helps pave the way for forming consensus and setting direction, as explored by the study of Jobin, Ienca and Vayena[15] on guidelines appearing across the AIE manifests of various organizations and institutions. They found that the prevalent AIE topics that appeared across studied material were transparency, justice, fairness and equity, non-maleficence, responsibility and accountability, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity [15].

While the principles exist, recognizing their limitations [30] and applying them to practice is still work in progress. Various proposed measures have been established to tackle the challenge [20] such as 'Ethics as a Service' [19], and the ECCOLA method for developers [28], which some of the authors of this study have been involved in testing and developing.

3 RESEARCH DESIGN

To address our research question – to examine researchers' views on the impacts of AI – this study was performed via distributing

[2]https://openai.com/blog/chatgpt
[3]https://www.cbsnews.com/news/chatgpt-chatbot-artificial-intelligence-job-replacement/
[4]https://www.forbes.com/sites/joemckendrick/2022/12/21/who-ultimately-owns-content-generated-by-chatgpt-and-other-ai-platforms/?sh=15b089125423
[5]https://cybernews.com/security/dark-side-of-chatgtp/
[6]https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html

a scenario-based survey. This section details the research process, including the description of data collection and analysis. An external repository with navigation instructions at the end of the paper contains further materials used in the study.

3.1 Research Process

The process begins with descriptive AI-related scenarios that participating AI researchers respond to in the survey. The survey collects the participants' thoughts on the scenarios and concerns associated with such possible outlooks as a free-form survey. These concerns are processed through content analysis, a qualitative analysis method. The data analysis detailed in section 3.3 operates on these topics utilizing further categorization, which lends potential to further aggregation.

First, scenarios were prepared for the survey. Once the survey was finished, it was distributed online for participants to fill. The survey was distributed in two iterations, which will be explained in more detail in section 3.2. After comparing the versions and the responses, responses to both were included in the data pool, resulting in 37 response sheets. Response sheets from non-researchers or researchers without affiliation with AI research were disqualified as invalid. After filtering out invalid responses, 33 remained, with 161 individual answers (response units). Content analysis [3] was used to categorize the results using content codes to obtain qualitative findings.

3.2 Data Collection

The data from this scenario-based survey consists of responses from 33 participants from 10 different universities in Germany and Finland. In this section, the term *response* refers to the entire submitted answer sheet of a single participant, including their answers to all questions, whereas *response unit* will refer to the single unit of answer a participant gave to a question. Often, a response unit contains one or more (sometimes enumerated) *concerns* of the participant regarding negative outcomes of the scenario. A total of 33 responses were included in the study. The sample size of 33 participants makes it a sufficient sample size for a qualitative study, according to Boddy [4].

With nearly all participants answering each question in the survey, the total number of valid response units received was 161. A response unit was excluded if the participant had left it empty or given an answer irrelevant to the analysis. The survey was distributed in two versions: Survey 1.0 and Survey 2.0. Survey 1.0 was distributed to one university only as an offline text document to serve as a pilot test for the final survey. After this first phase, the scenarios were reviewed and amended for better understandability and distributed as version 2.0. The changes were minor: they clarified the intent of each scenario without altering their content. Due to the data privacy statement issued in Survey 2.0, the responses will not be published. For study replication purposes, access to the research data can be requested from the authors.

The finished survey sheet 2.0 was distributed in Google Forms. The survey was advertised on LinkedIn, and invitations to the survey were sent to selected researchers in Finnish and German universities based on their expertise in AI research. The emailed researchers were discovered by their public researcher profiles on university websites. Due to the number of universities in Finland, the survey was sent to some researchers in all universities in Finland with a public agenda in AI research. For German universities, emails about the survey were sent to researchers in the authors' academic networks. All researchers who were emailed the survey were asked to redistribute it. Participants who came across the survey in any of the distribution channels were self-selected to participate in the survey. The participants' identities (in the second phase) are unknown to the researchers, but the responses are included based on trust in the participants' legitimacy.

Responses of both survey stages were included in the same data pool despite minor revisions, based on three considerations:

- the scenarios were modified with only minor revisions,
- the participants had included the same topics in their answers before the changes were implemented, indicating that they were already detectable in the original wording,
- the semantic content of the scenarios was unaltered.

First, demographic information was gathered. The survey introduced five fictional scenarios, written in a short story style, in which an AI system was used for a specific purpose. The scenarios did not offer technical details of exactly how the AI system works, allowing participants to make assumptions. The participant was not supposed to focus too much on whether the legislation of a specific country would currently allow such a system. However, they were encouraged to list reasons why the described AI system might not be legal. Some of the scenarios depict obviously legally dubious AI systems to provoke a variety of thoughts in the participants. After each scenario description, participants were asked *"what could go wrong?"*, inviting them to list all their concerns or negative outcomes as a result of the AI system described in each scenario. These concerns could be related to technical problems, societal impacts, or concerns about who and what could be negatively affected. The survey was deliberately constructed without a worded emphasis on AIE to avoid influencing the answers with biases such as the framing effect due to expectations introduced by terms such as 'ethics'.

The five scenarios of the survey depicted situations with different types of people and different variations of AI systems. The scenarios were inspired by different dominant topics in the field of AIE, particularly AI ethics guidelines by Jobin et al. [15], created with the intent to present realistic-sounding situations. The scenarios were designed with the aspiration to expand the participant's thought process to cover a variety of topics outside the constraints of pre-selected options. Despite the avoidance of a strict thematic framing, some AIE themes are still present in the scenarios. The five scenarios of the survey can be found in the external repository but are summarised below:

(1) A family is driving to a destination in an autonomous car that can only read the traffic from the same cues as a human. It was specified that there is a hybrid environment of both autonomous and human-steered vehicles. The scenario depicts a popular theme in AI research and thus serves as an easily approachable introduction to the topic, assuming that most participants will have heard of ethical questions associated with autonomous vehicles before. Major themes of this scenario include accountability for the car's actions,

data quality of its training, and philosophical dilemmas in situations where choices must be made autonomously.
(2) A recruitment algorithm selects the best candidates for an interview based on the company's hiring agenda. Candidates are selected autonomously for interviews which are then conducted by people in the company. The scenario includes questions related to the fairness and transparency of the AI system being used, as well as its usefulness to the company.
(3) A woman who's unfamiliar with technology attempts to contact a pharmacy's customer service, initially not understanding that she's talking to a ChatBot. The ChatBots presentation does not clearly state that it isn't human. The scenario deals with themes of transparency and human oversight over AI systems in terms of inexperienced user consideration, with an underlying question of how easy or difficult it is to opt out of using AI systems without explicit consent.
(4) An elderly man with Alzheimer's disease is put under a monitoring system utilizing AI for behavior analysis. The same AI is also implied to monitor other patients. While the man has signed a consent form, the system is used in his home by the decision of his family. The scenario is based on an example case used in tutorial workshops of the ECCOLA method [28]. The scenario deals with the questions of privacy and human dignity, freedom and autonomy. It also includes a factor of dubious human decision-making that is enabled or encouraged by excessive trust in an AI system.
(5) During a movie production, the actress of the main role falls into coma with only half of the intended scenes filmed. The director decides to use Deep Fake technology to create the likeness of her on screen. The movie is finished and published before the actress wakes up from the coma. The scenario deals with themes such as privacy, dignity and autonomy.

3.3 Data Analysis

The responses of the participants were analysed qualitatively, applying an adaptation of the *content analysis* method according to the stages of Bengtsson [3] as follows:

(1) **Decontextualisation.** First, meaningful information in the data, called 'meaning units' [3], were condensed into shorter descriptions for easier identification in the next analysis step. These *condensed meaning units* are noted as a short, textual description.
(2) **Recontextualisation.** Previously identified meaning units were condensed into shorter, single-concept codes. The codes describe the essence of a meaning unit in its shortest possible form. Examples of these codes are, e.g. "bias", "accountability" and "unfavorable transfer of human qualities". This stage is repeated iteratively: the first iteration used various codes which were condensed into fewer, broader descriptive codes. Finally, while a single concern may provide more than one code, each code was only used once per response unit to prevent redundancy. This decision was made based on the occurrence of some response units containing references to the other concerns within the same response unit, or stating a similar concern twice. The analysis thus measures *what kinds of concerns does a single participant have to a single scenario.* The analysis considered the deeper meaning behind the wording of the concerns, making it a type of *latent analysis* as opposed to manifest analysis (a broad surface structure analysis) [3].
(3) **Categorisation.** The final single-concept codes were organized through further categorization and considered for their implications. The categories that were found for the coded data lead to a distinction of the codes into *causes* and *effects*, as well as three categories of practical implications.
(4) **Compilation.** The final processing of the results aims to find the "essence" of the studied phenomenon, understanding the results and presenting them in an informative way [3]. Here, this stage also includes deducting the overall implications of the results, as presented in section 4 as results.

To ensure inter-coder reliability, the data was split between two researchers. The validation yielded moderately similar results between the two analysts. In the categorization phase, uniform consent was achieved. The detailed validation data of the qualitative coding phase can be found in the external repository.

4 RESULTS

This section provides an overview of the demographic information of the participants and the results of the qualitative analysis.

4.1 Demographic Information

Participants were asked to answer questions related to their research field and their experience with AI systems research. Gender and other identity-related issues were deliberately excluded from the demographic questions, and instead the questions focused on the research experience of the participants.

Table 1 lists the research fields of the participants. Due to almost half of the participants having a background in Computer Science and the overall representation of other research fields was left low in numbers, significant differences between researchers in different fields could not be distinguished. The multidisciplinary fields the participants reported were: Computer Science/Information Systems (3), Computer Science, Information Systems, Business/Economics (1), Physics, Chemistry, Nanotechnology, Materials Science and Computer Science (1), Software Engineering, Information Systems (1), Psychology, Computer Science (1) , Computer Science and Biology (1) and Artificial Intelligence, Data Analysis (1).

Table 1: Research fields of the participants

Computer Science	Information Systems	Multidisciplinary	Other
15	2	9	7

Table 2 presents the time the participants report to have allocated to considering the societal impacts of AI in their work, a question that measures the practical experience the participant has with AI ethics. The results indicate that the biggest portion of participants, 16, had allocated a little time in their work to consider these issues. The next largest portion of them, nine, had not considered these themes at all. The options for a lot of consideration and having it as a central theme of their research both received four responses.

Table 2: "Have you dedicated time in your research to considering the societal impacts of AI?"

Answers	Responses
No	9
Yes, a little	16
Yes, a lot	4
It is a central theme in my research	4

Table 3: "How IS AI present in your work?" (a multiple choice question)

Answers	Responses
I develop AI systems	18
I research AI systems	26
I use or apply AI systems	17

Table 4: Question: "How long have you worked with AI-related topics (including both industry and research) in years?"

Years worked	1 or less	2	3	4	5	6	7	8	9	10 or more
Responses	3	6	4	4	1	0	3	2	1	9

Table 3 presents the ways the participants work with AI systems. The question allowed multiple choices to be selected, resulting in more answers than the number of participants. As was the aim of the study, the largest portion of responses was that the participant researches AI system, indicating that the survey reached its target audience. In almost equal proportions, the participants additionally reported to develop, or to use or apply AI systems. Table 4 presents how long the participants report to having worked in AI research.

4.2 Qualitative Analysis

Table 5 presents the content analysis recontextualization stage results, depicting the single-concept qualitative codes and their total number of occurrences in the data. The codes describe the themes of concern found in the response units. The single-concept codes were further categorized as *causes and effects* and divided into three broader categories: ***social and ethical** issues;* ***technical and design** related issues*; and ***safety and security** risks*. Categorizing the codes as technical or human-oriented was under consideration, but in the case of many codes, these aspects are intertwined, and the codes are aligned both ways. The significance of the findings is discussed in this section.

The categorization of the codes to causes and effects reveals whether participants saw direct problems in the described AI system or the consequences of its use. Many response units included a consequence (effect) and the explicit reason it happens (cause). Overall, the number of causes was slightly higher (258) than the number of effects (232). This finding indicates that the AI researchers who participated in the survey identified the direct problems in the AI systems and their impacts. Further, it suggests that AI researchers do not only look at the immediate, surface-level challenges related to AI systems but consider their sphere of influence on a broader scale. For the context of this study, effects are considered more relevant than causes due to effects bearing closer resemblance to *impacts*, the emphasis of the survey. Thus, the presentation of the qualitative codes will focus on them.

The most commonly raised theme was the concern for **physical harm** caused to humans due to the AI system, with 47 occurrences. Scenarios 1, 3, and 4 all had responses coded as physical harm. Most prominently, direct and tangible harm emerged as a prevalent concern. This is an interesting finding, as the scenarios were not built with the intention to focus on physical harm. Scenario 1 is the only scenario that included a clear built-in emphasis on physical harm, yet it was present in scenarios 1, 3, and 4. **Psychological** and **social harm** are related types of direct specific harm to a person, but psychological harm was lower ranked with 30 occurrences as well as social harm with 21 occurrences. Interestingly, in scenario 3, where the emphasis was psychological rather than physical, physical harm (with 13 occurrences) was almost equal to the number of psychological harm codes (15 occurrences). **Financial harm**, interestingly, had only seven occurrences in the entire data, but this could be partly explained by the distinction between financial harm to business entities (negative business effects) and individuals. Particularly in scenario 3, financial harm to individuals could have been an expected result since the participants were concerned about the chatbot functioning in a faulty way. Instead, physical harm was a common concern.

Bias & unfairness was the fifth most commonly occurring code, consisting of any situation where the AI system would likely make a biased or discriminatory decision or the system itself would create or uphold unfairness in society. The third most common concern was human rights, which includes issues related to human's ability to govern themselves and maintain control of their own lives, including issues related to informed consent. Like bias and fairness, this code aligned with human-oriented goals and was present, particularly in scenarios 3, 4, and 5.

The scenarios in the survey depicted specific instances of an AI system without systematically including a specific palette of ethical themes, which means the interpretation of the results should also consider the context-specific distribution of themes in each case. The single-concept codes in the analysis reveal the specific concerns that the participants had, but the categories of those codes reveal the broader implications of the concerns. This part of the analysis aimed to distinguish what *types* of concerns researchers are likely to have about AI systems. The broadly categorized codes are presented in Table 6. The detailed descriptions of the qualitative codes and their distribution among the scenarios can be found in an external repository, to which a link is provided at the end of this research paper.

The themes in AIE research inspired the three themes, more concretely the ECCOLA method for implementing ethically aligned design in AI systems [28]. The method was also utilized in the analysis of concerns by Rousi et al. [23] as a framework in their data analysis. The ECCOLA method cards have eight themes: analysis, transparency, safety and security, fairness, data, agency and oversight, well-being, and accountability. The Social & Ethical category was inspired by the ECCOLA themes of fairness, accountability, and well-being. ECCOLA's safety and security inspired the Safety

Table 5: Results of the content analysis. Codes labeled as 'effects' are in italics.

Qualitative code	Occurrences
physical harm	47
training data limitation	41
human rights	39
model inaccuracy	39
bias & unfairness	38
privacy violation	34
psychological harm	30
lack of model adaptability	29
unwanted societal effects	28
unfavorable transference of human qualities	26
negative business effects	23
social harm	21
transparency	20
model incompetence	19
legal issues	17
loss of human agency or self-determination	14
faulty ethical priorities	14
unclear accountability	9
system vulnerability	9
financial harm	7
accessibility	6
missed potential	6
technical or physical problem	4

& Security category and agency and oversight. The category Technical & Design consists of the ECCOLA theme of Data and mixed themes that fall out of the method's immediate grasp. The names of the ECCOLA themes and the qualitative codes sorted under the categories used in this study may differ due to different interpretations and descriptions of the words.

The descriptions of the categories are as follows:

- **Social & Ethical**: Issues related to social and societal well-being and alignment with ethical or moral codes, issues that are intertwined with human behavior or have primarily a social or societal significance.
- **Technical & Design**: Issues that have a technical origin and context that have to do with the design or programming of the AI system or issues that are related to the functionality of its physical parts.
- **Safety & Security**: Issues related to physical, psychological or information-related safety and security of people and other entities.

The most prevalent category was the Social & Ethical category, with 241 codes; the second most prevalent Technical & Design, with 152 codes; and the third Safety & Security, with 127 codes. All in all, the largest part of the concerns that emerged in the study appear to be related to social and ethical themes.

Table 6: Results of the final categorizing analysis

Code name	Total occurrences	Category
Human rights; Agency; Social harm Bias & unfairness; Unwanted societal effects; Negative business effects; Transparency; Legal issues; Accountability; Accessibility; Unfavorable transfer of human qualities	241	Social & Ethical impacts
Training data limitation; Inaccuracy; Adaptability; Incompetence; Priorities; Missed potential; Physical problem	152	Technical & Design issues
Physical harm; Privacy; Psychological harm; Exploitability; Financial harm;	127	Safety & Security

5 DISCUSSION

The interpretation of these results from an AIE perspective is positive; the findings suggest that AI researchers are well attuned to AI systems' humane, social, and societal effects. The participants mostly had a background in computer science, which traditionally has no explicit connection to humanistic sciences. In this context, the prevalence of expertise in AIE topics tells a promising tale: AI researchers from various fields without a particular emphasis on AIE in their work (as indicated by Table 3) are thinking about ethics.

The overlap of concerns and established topics in AIE is clear. The codes categorized under the Social & Ethical category have significant overlap with common themes in AIE research; human rights, avoidance of bias, accountability, and transparency are among the most prevalent themes in AIE guidelines [15]. Many of these themes and the codes related to negative effects on humans and the legality of AI systems are also found in the IEEE Ethically Aligned Design (EAD) document [1], which thoroughly discusses priorities for ethical and trustworthy autonomous and intelligent systems. For example, accountability is among the general principles [1].

The prevalence of the Social & Ethical category is not the sole indicator of the familiarity of the participants to AIE: the themes in the other categories (Technical & Design, Safety & Security) relate to these themes as well. For example, the code exploitability is related to the general principles of Awareness of Misuse and Competence in the EAD [1], physical and psychological harm attack human well-being [1], and training data limitation has an ethical connotation. This concern aligns with both technological and social issues. Data quality relates heavily to ethical questions in the context of AI, particularly on the deliberate or indeliberate introduction of built-in bias to an AI system through its training data, as several types of bias

can easily find their way into the system through misrepresentative data [22].

Compared to the results of Rousi et al. [23], who also studied researchers' ethical concerns in an adjacent topic despite the scale and topic differences in the two studies, there are some similarities. Rousi et al. also found themes of accountability, privacy, security concerns, and data-related issues. The two studies, therefore, validate each other in this regard.

While the results show clear concern for ethical aspects of AI, some dangers (such as direct physical harm) are more prevalent in the minds of researchers than other issues. These initially surprising results highlight a need for sensitisation towards different types of risk and harm which may not be as immediately obvious as poor medical advice or accidents in which autonomous vehicles were involved. Due to the variety of applications for AI, with many of them not relating to health-critical areas of human life, it is important to highlight less obvious ways AI could harm its users and society at a large. This is of particular importance for industrial use cases. With the rise of LLMs, generative AI and other AI-based services as well as the dominance of private companies developing foundation models, many developers of applications are disconnected from traditional, academic research and the ethical regulations it is subjected to. As such, this need for increased sensitisation for types of social and mental harm is not solely with researchers, but also of increasing importance during the education of future software and service developers to take ethical considerations into account where legal guidelines have not yet caught up with the state of the art and the possibilities offered to software development and AI-driven services.

All scenarios present cases for which private business either already offer AI-driven services or are currently developing applications that allow for commercial use. With the ubiquity of AI and often slow legislative and academic processes, companies hold a high degree of responsibility which is not yet subject to the same degree of regulation and consumer protection as more traditional fields of business.

While the results indicate that researchers are aware of the ethical impacts of AI systems, research also shows that urging the use of ethical frameworks, such as the ACM code of ethics, has varying effects, ranging from none [17] to unexpected positive effects [14]. In order to create solutions, one must first understand the challenge. Is the disconnect between research into ethics and the application of guidelines caused by a lack of knowledge among developers, hindered by profit-driven superiors or are other factors at fault?

Garrett et al. [11] analyzed AI ethics in university classes and suggest that integrating ethics into the technical practice of AI building could make ethical issues more tangible, and ethics related topics in AI education are already increasing [25]. It has been suggested that people in STEM may find ethics to be a problem for somebody else to solve [5]. However, based on the results of this study, it seems that at least in research, the capabilities for ethical consideration are strong. Now, it's time to find ways to empower AI systems development stakeholders to apply their ethical thinking skills. In the light of these findings, the imbalance of ethical requirements and their consideration in SE practices might see an improvement with increased inclusion of researchers in the industry.

Another research opportunity to dig deeper into the topic could be to conduct a similar style study with a controlled sample of participants from different fields in a way that would enable a meaningful comparison between, e.g., researcher experiences in different fields or the amount of experience in research. The scope of this study was purposefully restricted to enable a more detailed qualitative analysis, but a study with a larger sample with shorter response units would enable a larger scale quantitative study on the topic. A study with access to comparisons between fields could find out if there are any significant differences in the way developers or researchers in different fields approach the topic. A study with scenarios that are planned to systematically include specific themes could enable the analysis of a more controlled, specific range of ethical considerations.

5.1 Validity Threats

When it comes to the validity of the study, the qualitative nature of the analysis and individual biases of the researchers add subjectivity to the results of the study. These threats to validity were mitigated with a partial validating analysis by a researcher not included in the research process. The comparison between the results can be found at the external repository. Due to the method of inviting participants to fill the survey, it was not possible to assess the response rate or confirm the identities of all of the participants. The study operated with the trust that the participants answered the survey with honesty.

While the research method of a text-based survey enabled a larger sample size, the limitations posed by text-based anonymous responses apply. Some in-depth methods, such as interviews, could be used in the future to complement the results of this study. In retrospect, some of the questions in the demographic section of the survey could have been phrased with more precision to extract more defined responses. For example, we now know how much time the participants assess to having used to consider societal impacts of AI in their work, but we do not know where the motivation to do that came from - was it internal or external, *why* did they, or did not, consider these questions.

Lastly, the scenarios chosen for the survey cover only a portion of the whole range of themes discussed in AI ethics, which can lead to uneven distribution of themes discovered by the participants. However, this would be difficult to mitigate without creating a long and time-consuming survey that may have led to significantly fewer participants willing to fill it.

5.2 Conclusion

The study investigated what concerns AI researchers have on the impacts of AI systems, approaching the topic from the perspective of AIE. The data were analyzed qualitatively applying the content analysis method [3], resulting in single-concept qualitative codes and broader categories of those codes to assess the overall types of concerns the participants had. The categorized results show that the most commonly considered themes the participants wrote down were related to the social and ethical impacts of AI systems. The data analysis also suggests that the participants thought of both the negative impacts of AI systems and the issues that cause them.

The results indicate that even though most participating AI researchers dedicated little or no time to considering the impacts of AI in their work, they are overall very conscious of ethical issues and the risks of the negative social impact caused by AI. More specifically, AI researchers are concerned about several negative impacts of AI systems related to themes such as causing physical, psychological, financial, or social harm, negative effects on human autonomy and agency, bias and unfairness, privacy violations, unwanted societal effects, negative business effects, and legal problems. The most common concerns identified in the study are regarding social and ethical problems related to AI systems. The AI researchers' concerns in the study have a solid connection to AIE. The participants appear to be familiar with and concerned about topics established in AIE.

In this regard, the results of this survey confirm previous findings of smaller-scaled studies and cement the ubiquity of ethical concerns relating to AI beyond specialized researchers. Therefore, these findings further highlight the importance of ethical considerations during the academic and commercial development of AI applications as an additional dimension that developers and systems architects need to take into consideration.

EXTERNAL RESOURCE

More materials used in this study can be accessed at an external repository, found at: https://zenodo.org/record/8247163

REFERENCES

[1] 2019. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

[2] Blair Attard-Frost, Andrés De los Ríos, and Deneille R Walters. 2022. The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics* (2022), 1–18.

[3] Mariette Bengtsson. 2016. How to plan and perform a qualitative study using content analysis. *NursingPlus open* 2 (2016), 8–14.

[4] Clive Roland Boddy. 2016. Sample size for qualitative research. *Qualitative Market Research: An International Journal* 19, 4 (2016), 426–432.

[5] Jason Borenstein and Ayanna Howard. 2021. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics* 1, 1 (2021), 61–65.

[6] Nick Bostrom and Eliezer Yudkowsky. 2018. The ethics of artificial intelligence. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 57–69.

[7] Virginia Dignum. 2018. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20, 1 (2018), 1–3. https://doi.org/10.1007/s10676-018-9450-z"

[8] Ethics Guidelines for Trustworthy AI 2019. Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[9] Amitai Etzioni and Oren Etzioni. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21 (2017), 403–418.

[10] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (2019), 261–262.

[11] Natalie Garrett, Nathan Beard, and Casey Fiesler. 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*. 272–278.

[12] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research* 62 (2018), 729–754.

[13] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* (2020), 1–22.

[14] Erika Halme, Ville Vakkuri, Joni Kultanen, Marianna Jantunen, Kai-Kristian Kemell, Rebekah Rousi, and Pekka Abrahamsson. 2021. How to write ethical user stories? impacts of the ECCOLA method. In *Agile Processes in Software Engineering and Extreme Programming: 22nd International Conference on Agile Software Development, XP 2021, Virtual Event, June 14–18, 2021, Proceedings*. Springer International Publishing Cham, 36–52.

[15] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[16] LD Jones, D Golan, SA Hanna, and M Ramachandran. 2018. Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? *Bone & joint research* 7, 3 (2018), 223–225.

[17] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. 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*. 729–733.

[18] Forrest E Morgan, Benjamin Boudreaux, Andrew J Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. 2020. *Military applications of artificial intelligence: ethical concerns in an uncertain world.* Technical Report. RAND PROJECT AIR FORCE SANTA MONICA CA SANTA MONICA United States.

[19] Jessica Morley, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. 2021. Ethics as a service: a pragmatic operationalisation of AI ethics. *Minds and Machines* 31, 2 (2021), 239–256.

[20] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics* 26, 4 (2020), 2141–2168.

[21] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2021. Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY* (2021), 1–13.

[22] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[23] Rebekah Rousi, Ville Vakkuri, Paulius Daubaris, Simo Linkola, Hooman Samani, Niko Mäkitalo, Erika Halme, Mamia Agbese, Rahul Mohanani, Tommi Mikkonen, et al. 2022. Beyond 100 Ethical Concerns in the Development of Robot-to-Robot Cooperation. In *2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*. IEEE, 420–426.

[24] Stuart Russell, Sabine Hauert, Russ Altman, and Manuela Veloso. 2015. Ethics of artificial intelligence. *Nature* 521, 7553 (2015), 415–416.

[25] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)* 19, 4 (2019), 1–26.

[26] V. Vakkuri, K. Kemell, J. Kultanen, and P. Abrahamsson. 2020. The Current State of Industrial Practice in Artificial Intelligence Ethics. *IEEE Software* 37, 4 (2020), 50–57.

[27] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson. 2020. "This is just a prototype": How ethics are ignored in software startup-like environments. In *Agile Processes in Software Engineering and Extreme Programming: 21st International Conference on Agile Software Development, XP 2020, Copenhagen, Denmark, June 8–12, 2020, Proceedings 21*. Springer International Publishing, 195–210.

[28] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. 2021. ECCOLA—A method for implementing ethically aligned AI systems. *Journal of Systems and Software* 182 (2021), 111067.

[29] Pei Wang. 2019. On defining artificial intelligence. *Journal of Artificial General Intelligence* 10, 2 (2019), 1–37.

[30] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 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*. 195–200.

[31] Baobao Zhang, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C Horowitz, and Allan Dafoe. 2021. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research* 71 (2021), 591–666.