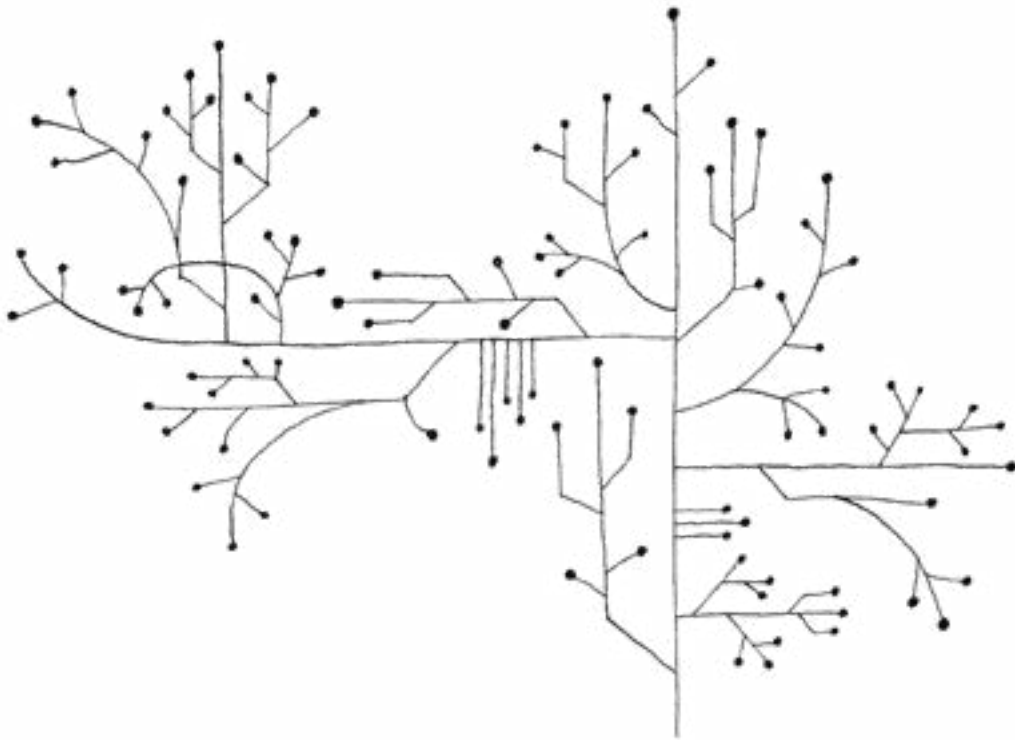


JYU DISSERTATIONS 526

Ville Vakkuri

Implementing AI Ethics in Software Development



UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION
TECHNOLOGY

JYU DISSERTATIONS 526

Ville Vakkuri

Implementing AI Ethics in Software Development

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston vanhassa juhlasalissa S212
toukokuun 27. päivänä 2022 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Seminarium, Old Festival Hall S212, on May 27, 2022, at 12 o'clock.



JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2022

Editors

Marja-Leena Rantalainen

Faculty of Information Technology, University of Jyväskylä

Päivi Vuorio

Open Science Centre, University of Jyväskylä

Cover picture by Hinako Sano.

Copyright © 2022, by University of Jyväskylä

ISBN 978-951-39-9170-8 (PDF)

URN:ISBN:978-951-39-9170-8

ISSN 2489-9003

Permanent link to this publication: <http://urn.fi/URN:ISBN:978-951-39-9170-8>

ABSTRACT

Vakkuri, Ville

Implementing AI Ethics in Software Development

Jyväskylä: University of Jyväskylä, 2022, 72 p.

(JYU Dissertations

ISSN 2489-9003; 526)

ISBN 978-951-39-9170-8 (PDF)

Technical advances in Artificial Intelligence (AI) have made AI-powered systems a part of our everyday life. This is seen across applications from targeted advertising to autonomous vehicles. AI systems have grown from the confines of laboratories and are now applied in diverse societal contexts. Alongside the success stories of AI applications, false promises of the technology and growing numbers of incidents related to AI systems have highlighted the need to address ethical considerations. With current technology, the previously hypothetical threats and ethical issues related to development, application and use of AI have now become a reality. It is foreseeable that more issues will arise as the level of maturity increases.

AI systems are still far from perfect. This realization has led to a call for action towards the advancement of AI ethics in the field, resulting in a high demand for principle-based approaches to AI ethics. Various stakeholders have addressed AI ethics via guidelines, laws, and regulations. AI principles in the form of guidelines have lacked actionability and developers have struggled with implementing abstract ethical guidelines into concrete actions. Transferring principles into practice is a major challenge for AI ethics. The question remains about how to influence developers to identify and consider ethical issues in stages of development. Although methods in the area exist, empirically tested AI ethics methods are still in need to bridge the gap of research and practice.

The lack of actionability in proposed high-level solutions has prompted this research. This dissertation offers a means of addressing this research problem via an empirically grounded understanding on how to implement ethics in software development processes. The focus is on operational tools used in software development that transform philosophical thinking tools and principles to development practices. The research includes five qualitative research articles: two conference papers, one magazine article and two journal articles. The results of the dissertation further our understanding of how to implement ethics in software development. The method ECCOLA (Article V) helps to raise awareness of ethical issues and offers a process to implement ethical considerations into software development.

Keywords: Artificial Intelligence, Ethics, AI ethics, Software Development, Methods, Implementing

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Vakkuri, Ville

Tekoälyn etiikan soveltaminen osaksi ohjelmistokehitystä

Jyväskylä: Jyväskylän yliopisto, 2022, 72 s.

(JYU Dissertations

ISSN 2489-9003; 526)

ISBN 978-951-39-9170-8 (PDF)

Teknologinen kehitys on tuonut tekoälyä hyödyntävät ohjelmistoratkaisut osaksi arkipäivää ja niitä löytyy niin kohdennetuista mainoksista kuin autoista. Enää tekoälyä ei hyödynnetä vain laboratorioden rajatuissa konteksteissa, vaan sen käyttö on levinnyt kaikille elämän osa-alueille. Uusien mahdollisuuksien rinnalla ei ole voitu välttää tekoälyyn liittyviltä ongelmilta. Uutisoinnit erilaisista tekoälyjärjestelmien epäonnistumisista ja niistä seuranneista onnettomuuksista ovat tuoneet teknologian käyttöön liittyvät eettiset haasteet yleiseen tietoisuuteen.

Havahtuminen siihen, että kyvykkäät tekoälyjärjestelmät tuovat väistämättä mukanaan uusia haasteita, on kiihdyttänyt tekoälyn etiikan tutkimusta. Viimeisen vuosikymmenen aikana tekoälyn eettisiä ulottuvuuksia koskeva tieteellinen sekä yleinen keskustelu on lisääntynyt ja osallistujia on ollut niin tietentekijöistä, yrityksistä, valtiollisista toimijoista kuin kansainvälisistä järjestöistä. Tämän tuloksena on syntynyt useita keskenään kilpailevia tekoälyn eettisiä periaatteita, joita on julkaistu tekoälyn käyttöä koskevissa ohjeistuksissa. Periaatteiden saavuttamasta suosiosta huolimatta niiden on todettu olevan haasteellisia tai jopa käyttökelvottomia käytännön tasolla. Periaatteiden soveltamisesta käytäntöön on tullut yksi tekoälyn etiikan keskeisimmistä haasteista. Kysymys siitä, kuinka soveltaa tekoälyn etiikkaa käytäntöön siten, että ohjelmistokehittäjät sekä tunnistaisivat että kykenisivät ratkaisemaan työhönsä liittyviä eettisiä haasteita, on edelleen avoin.

Tämä väitöskirja pyrkii vastaamaan tekoälyn etiikan käytäntöön soveltamisen haasteisiin hyödyntämällä ohjelmistokehityksen ja kehittäjien näkökulmaa. Väitöskirja koostuu viidestä laadullisen tutkimuksen artikkelista: kahdesta konferenssijulkaisusta, yhdestä lehtiartikkelista sekä kahdesta tieteellisestä aikakausjulkaisusta. Väitöskirjassa esitetty tutkimus auttaa ymmärtämään, miten tekoälyn etiikkaa voi soveltaa osaksi ohjelmistokehitystä. Artikkelissa V esiteltävän ECCOLA-menetelmän avulla voidaan lisätä tietoisuutta tekoälyn etiikan kysymyksistä sekä jalkauttaa eettisten haasteiden käsittely osaksi ohjelmistokehitystä.

Avainsanat: Tekoäly, Etiikka, Tekoälyn etiikka, Ohjelmistokehityskäytänteet, Ohjelmistonkehitysmenetelmä, Implementointi

Author

Ville Vakkuri
Faculty of Information Technology
University of Jyväskylä
Finland
ORCID: 0000-0002-1550-1110

Supervisors

Professor Pekka Abrahamsson
Faculty of Information Technology
University of Jyväskylä
Finland

Professor Mikko Siponen
Faculty of Information Technology
University of Jyväskylä
Finland

Reviewers

Docent Patrik Floréen
Department of Computer Science,
University of Helsinki
Finland

Professor Daniela Soares Cruzes
Department of Computer Science
Norwegian University of Science and Technology
Norway

Opponents

Professor Thomas Olsson
Faculty of Information Technology and
Communication Sciences
Tampere University
Finland

Professor Kai Petersen
School of Business
Flensburg University of Applied Sciences
Germany

ACKNOWLEDGEMENTS

Now, in May of 2022 I have been working with AI ethics almost for six years. It is hard to believe how fast these years have gone and how many amazing people, events, and opportunities I have had and met during these years. There are plenty to be grateful of and many to thank for.

Even with difficulties, studying and learning have always been a passion for me. It was natural for me that after graduating from university of Helsinki I would apply for doctoral studies. I simply crave for more. Already, while I was finishing my social ethics master thesis, I was inspired by the ethical challenges related to use of technology. Especially in 2016 I was captivated by the rapidly progressing application of artificial intelligence and the new context that it provided to applied ethics.

Despite my extensive background work on AI ethics and enthusiasm towards the matter, it was hard for me to find a professor who would be willing to supervise dissertation focusing on AI ethics. On many occasions, when I was meeting with different professors, I faced answers that, the topic is interesting and relevant, but not under my current research interest or expertise. Some even thought at the time, that AI ethics was not a topic worth pursuing for. Therefore, it is clear that I first need to express my gratitude to my first supervisor, Professor Pekka Abrahamsson who was at first in place willing to take me under his supervision and saw the potential in AI ethics. Pekka has done a great job in reminding me to consider, what ethics means in practice and for practitioners. Thank you Pekka for believing in me and AI ethics!

I would also like to thank my second supervisor, Professor Mikko Siponen. With your philosophically oriented mindset and critical thinking you have reminded me to ask myself what is in the core ethics and what is just software engineering in my work. Having the possibility to teach with you was great way to finalize the doctoral studies. The teaching experience gave me possibility to share my experiences to others!

I also want to thank those people who gave from their time to make this dissertation to what it is now. I would like to thank both of my preliminary examiners, Professor Daniela Soares Cruzes, and Docent Patrik Floréen, for taking their time to review this dissertation and to provide suggestions on how to improve it. For the cover image and chapter illustration I want to thank Hinako Sano, who have seen this process as single articles transfer to dissertation. どうもありがとう! Thank you, Rebekah Rousi and Ida Vakkuri, you both provided helpful comments for the language. I am sure that your feedback has made this dissertation more accessible both in English and Finnish!

Working with Professor Pekka Abrahamsson have always meant to me that I am not just working by myself but also as part of a community, sharing and learning from others. This attitude has brought many colleagues and co-writers to my path along the way. Among my colleagues, I first and foremost need to thank Kai-Kristian Kemell for all the extended co-writing sessions and award-winning papers. You were the colleague to talk to when I needed to make sense

out of the doctoral studies. I am grateful of how effortless and smooth the working, writing, and sharing ideas have been and still is with you! I hope other researchers would find colleague like you during their careers. When talking about the role of community, I cannot over emphasize the role of the AI ethics team that I have had chance to coordinate since its birth. The team with its weekly meetings and research projects have been fruitful meeting places for researchers young and more seasoned with versatile backgrounds to pass ideas and further AI ethics research. Thank you, past and current AI ethics team members, of sharing the enthusiasm towards ethically minded, human centric and socially aware software engineering.

I would also like to thank everyone at the JYU Startup Lab. Especially I would like to highlight Joni Kultanen. Alongside few others Joni was crazy enough to apply for the AI Ethics team as we were starting the lab. Afterwards Joni have shifted more towards general duties at JYU Startup Lab and have grown to be the go-to guy for the lab practicalities and savior of many projects. In this vein, it is natural to thank the multiple founding institutions and programs making this research possible. This work has been conducted on the founding provided by multiple research programs, expect for the first six month start grant from JYU IT faculty. This founding sets an important example that AI ethics is valued both by the academics and the industry.

As exciting research with AI ethics would have not been enough, I have entered priesthood during my studies, and it has been also a big part of my life. Here I want to extend my gratitude to Karjasillan seurakunta – Karjasilta parish – for being very supportive for my studies and research.

Finally, I want to express my gratitude to my parents Tia and Matti Vakkuri and family. Things to be grateful of overruns given page limit and my capacity to find words. May these small words of thank you express the wave of gratitude. Tia and Matti you have made this process possible in unbelievably many ways. Your support has been something that I have learned to rely on. Similar attitude towards work and life have made it easy to share my work with you. My wife, Miia-Liisa Vakkuri, liked or not, you have experienced the back seat view to my research and studies since much of the work have been done remotely wherever our family has been. You have stretched to the need for our family when my research has taken me away from home. You have walked with me the whole process, seen the crisis and had patience to understand what my research is about. Miia, thank you for always asking me to reach the potential that I have. And yes, we can name our summer cottage's outhouse ECCOLA.

I want to end by thanking Luukas Vakkuri for constantly reminding me by asking "Isn't it ready yet!" to finish the dissertation process. This process is now over but only for the dissertation as my work on AI ethics continues!

27.5.2022

Ville Vakkuri

LIST OF INCLUDED ARTICLES

- I Vakkuri, V., & Abrahamsson, P. (2018). The key concepts of ethics of artificial intelligence. In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)* (pp. 1–6). IEEE.
- II Vakkuri, V., Kemell, K. K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2022). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. arXiv: 1906.07946. To be published in *Journal of Business Ethics and Organization Studies EJB*.
- III Vakkuri, V., Kemell, K. K., Kultanen, J., & Abrahamsson, P. (2020). The current state of industrial practice in artificial intelligence ethics. *IEEE Software*, 37(4), 50–57.
- IV Vakkuri, V., & Kemell, K. K. (2019). Implementing AI ethics in practice: An empirical evaluation of the RESOLVEDD strategy. In *International Conference on Software Business* (pp. 260–275). Springer.
- V Vakkuri, V., Kemell, K. K., Jantunen, M., Halme, E., & Abrahamsson, P. (2021). ECCOLA – A method for implementing ethically aligned AI systems. *Journal of Systems and Software*, 182, 111067.

The author's contribution to the articles is as follows. The author of this dissertation served as the main author in all included articles. For Article I he was responsible for the research idea, literature review, data collection, analysis, and the writing process. The co-author reviewed and improved the article, provided constructive comments, and helped with its structure.

For Article II, the author of this dissertation was responsible for the research idea, design, literature review and data analysis. The data collection was performed by the third author. The second author assisted with validation of the data analysis results. The article was written as a joint effort by the first and second author. The fourth and fifth authors reviewed the article, provided constructive comments, and helped with the presentation of the results.

For Article III, the authors worked jointly on the design of the research framework, the survey questions, and writing the article. The author of this dissertation was responsible for organizing the survey and conducting data analysis. The third author was responsible of the pre-study and data collection.

For Article IV, the author of this dissertation was responsible for the overall research idea and design, organizing data collection, and formulating the data analysis framework. Alongside the first author, the co-author of the article contributed to the literature review and data analysis to validate the results. The final writing of the article was done in collaboration with all the authors.

Article V is the result of four years of close collaboration between the author of this dissertation and the second author of the article. They worked together to build the method and testing, and further developed the method through various research cycles. The writing of the article was a joint effort among researchers participating in the last steps of the model development led by the author of this dissertation.

FIGURES

FIGURE 1	Operational tools used in software development	15
FIGURE 2	Research questions and their relationship to the included articles.....	16
FIGURE 3	Three categories of AI ethics.....	17
FIGURE 4	Two components of ethical approach	27
FIGURE 5	Pragmatic cycle of empirical research.....	29
FIGURE 6	Survey questions	39
FIGURE 7	Example card from the ECCOLA method.....	43
FIGURE 8	ECCOLA cards 0 and 1.....	62
FIGURE 9	ECCOLA cards 2 and 3.....	63
FIGURE 10	ECCOLA cards 4 and 5.....	64
FIGURE 11	ECCOLA cards 6 and 7.....	65
FIGURE 12	ECCOLA cards 8 and 9.....	66
FIGURE 13	ECCOLA cards 10 and 11.....	67
FIGURE 14	ECCOLA cards 12 and 13.....	68
FIGURE 15	ECCOLA cards 14 and 15.....	69
FIGURE 16	ECCOLA cards 16 and 17.....	70
FIGURE 17	ECCOLA cards 18 and 19.....	71
FIGURE 18	ECCOLA card 20	72

TABLES

TABLE 1	Set of artificial intelligence technologies.....	22
TABLE 2	Typology of AI Ethics Methods	25
TABLE 3	Methodological aspects of articles.....	30
TABLE 4	Collected data	31
TABLE 5	Analysis, methods, and processes	33
TABLE 6	Primary empirical conclusions of the study.....	37

CONTENTS

ABSTRACT

TIIVISTELMÄ (ABSTRACT IN FINNISH)

ACKNOWLEDGEMENTS

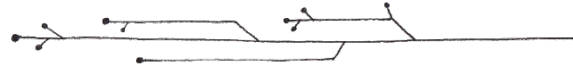
LIST OF INCLUDED ARTICLES

FIGURES AND TABLES

CONTENTS

1	INTRODUCTION	13
1.1	Demand for Ethical AI	13
1.2	Research Objectives and Scope	15
1.3	Structure of this Work.....	18
1.4	Other Scientific Contributions by the Author	18
2	THEORETICAL BACKGROUND	20
2.1	Artificial Intelligence as a Set of Technologies	20
2.2	AI Ethics and Methods Used in AI Ethics.....	23
2.3	Applied Ethics Approach to AI Ethics	26
3	RESEARCH METHODOLOGY	28
3.1	Research Approach and Design	28
3.2	Data Collection - Empirical Evidence	30
3.3	Data Analysis	32
4	OVERVIEW OF THE ARTICLES.....	34
4.1	Article I: The Key Concepts of Ethics of Artificial Intelligence.....	34
4.2	Article II: Ethically Aligned Design of Autonomous Systems: Industry Viewpoint and an Empirical Study	35
4.3	Article III: The Current State of Industrial Practice in Artificial Intelligence Ethics	38
4.4	Article IV: Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy	40
4.5	Article V: ECCOLA – A Method for Implementing Ethically Aligned AI Systems	41
5	RESULTS AND CONTRIBUTIONS	44
5.1	Results	44
5.2	Validity Threats.....	45
5.2.1	Reliability	46
5.2.2	Construct Validity	47
5.2.3	Internal Threats to Validity.....	48
5.2.4	External Threats to Validity.....	49
5.3	Contributions.....	49

5.3.1	Theoretical contributions	50
5.3.2	Practical Contributions.....	51
5.3.3	Limitations	52
5.4	Further Studies.....	53
YHTEENVETO (SUMMARY IN FINNISH)		54
REFERENCES.....		56
APPENDIX: ECCOLA CARDS.....		61
ORIGINAL PAPERS		



1 INTRODUCTION

This chapter describes the research area of the dissertation and introduces the scope of research, including research questions and outlines the dissertation structure.

1.1 Demand for Ethical AI

With Artificial Intelligence (AI) technology development progressing rapidly and bringing prominent breakthroughs, AI-powered systems have become increasingly prevalent in our lives, while at the same time having a profound and widespread impact on society. The deployment of AI systems from the private to public domains has led to us witnessing a number of AI system failures. At the same time, the ethical challenges of AI systems are becoming increasingly more evident. Individual but regrettable failures of AI systems have been brought to the public's attention. These are issues that academics have emphasized as potential ethical AI technology concerns. For example, these systems are reported to have led to unintended, but harmful, consequences such as intrusions of privacy, discrimination, and opaque decision-making (Zhang et al., 2021). Among many others, one prominent topic of concern has been the use of facial recognition technology, which has raised alarm among the general public,¹ as well as among policymakers.²

Incidents involving AI systems and their use have resulted in collective learning experiences, providing an understanding that the systems that have been developed are still far from problem-free. Especially when considering deployment in the versatile context of human lives and society. Ethical issues seem to persist, and more issues arise as the level of sophistication of AI

¹ Resisting the rise of facial recognition <https://www.nature.com/articles/d41586-020-03188-2>

² Regulating facial recognition in the EU: [https://www.europarl.europa.eu/thinktank/en/document/EPRS_IDA\(2021\)698021](https://www.europarl.europa.eu/thinktank/en/document/EPRS_IDA(2021)698021)

technologies increases. Aside from the obvious issue of the capacity to cause physical harm that embodied AI such as robots and autonomous vehicles have, there are many other application areas of AI systems that are full of indirect ethical issues. These indirect ethical issues range from well-known topics such as data handling to the complex societal impact from AI systems and the still unforeseen future systems. These issues call for action towards advancing the work of AI ethics as a field of study. This is coupled by the necessity to reinforce software engineering skills in order to resolve potential ethical problems while projects are still in early stages of development.

Over the past decade as AI-related technology progresses, discussions in the field of AI ethics have soared, resulting in high demand for principle-based approaches to AI ethics. Discussions have led to a state in which many of the relevant AI ethics issues have been addressed with widely acknowledged key principles, although the debate on what these principles encompass is still open (Morley et al., 2021). The principles cover a wide range of subjects, such as the goal of building responsible AI (Dignum, 2017) and demands for AI systems to be explainable (Rudin, 2019). Additionally, they cover the alignment with human rights and well-being (IEEE Global Initiative, 2019). Partly due to the ambiguity of these principles, transferring these principles into practice has proved a major challenge for principle-based approaches to AI ethics. That is, the question remains regarding how to influence developers in order for them to identify and think through ethical issues during the development of these systems (Canca, 2021).

Recently, the preferred means to address AI ethics has been through guidelines, laws, and regulations. For example, the EU has launch proposal for shared regulation for use and application of AI (EU AI Act³). Guidelines have been published by various stakeholders such as companies (e.g., Google,⁴ IBM,⁵ Sony⁶), governments (e.g., EU [HLEG, 2019], Canada⁷), and standardization organizations (e.g., IEEE [IEEE Global Initiative, 2019], ISO⁸). Despite their popularity, AI principles presented in the form of guidelines have been generally lacking in actionability (Canca, 2021). Developers are reported to have struggled while implementing abstract ethical guidelines in practice (McNamara et al., 2018). On the other hand, even if ethical guidelines were in place, there are no procedures to guarantee that they would influence the actual decision-making of developers (Hagendorff, 2020).

³ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>

⁴ AI at Google: our principles <https://www.blog.google/technology/ai/ai-principles/>

⁵ IBM's multidisciplinary, multidimensional approach to trustworthy AI: <https://www.ibm.com/artificial-intelligence/ethics>

⁶ AI Engagement within Sony Group: https://www.sony.com/en/SonyInfo/csr_report/humanrights/AI_Engagement_within_Sony_Group.pdf

⁷ Responsible use of artificial intelligence (AI): <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html>

⁸ ISO/IEC JTC 1/SC 42 technical committee: <https://www.iso.org/committee/6794475.html>

Methods and practices in the area are: (a) highly technical and focus on, for example, specific machine learning issues (Morley et al., 2021); (b) developed and administered externally to the software engineering (SE) domain (e.g., VSD [Friedman et al., 2008]); or (c) developed outside the academic community without empirical testing (e.g., AI Ethics Cards at 33A⁹). While highly technical methods are useful in their specific contexts, they broadly offer only limited help for companies in the design and development process. On the other hand, commercial methods may be appealing for design and development, but they often lack rigor and evidence to support their claims. Therefore, other approaches such as empirically tested development methods for ethical AI are still required to bridge the gap between research and practice in the area.

1.2 Research Objectives and Scope

Considering the demand for including ethical considerations into AI systems and the lack of actionability of proposed high-level solutions among the guidelines, laws, and regulations, this dissertation aims to answer the problem from the viewpoint of software development. More precisely, the focus is on operational tools used in software development that transform from philosophical thinking tools and principles to development practices (FIGURE 1). AI system development, much like software development in general, is an activity where humans developing it play a significant role in deciding how the system behaves.

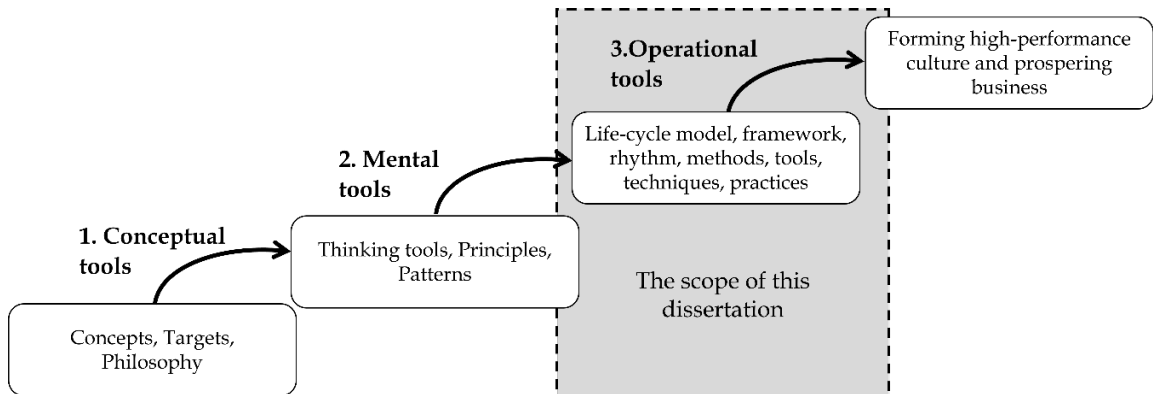


FIGURE 1 Operational tools used in software development (Ebert et al., 2012)

⁹ <https://www.33a.ai/ethics>

Therefore, to make AI guidelines and principles actionable, the main research question of this dissertation is:

- How to implement AI ethics in software development?

The supporting sub-questions are presented as follows:

1. How can the field of AI ethics be made sense of and organized?
2. How is the software industry implementing ethics today?
3. How can industry be supported in addressing its AI ethics concerns?

To address the research questions, this dissertation presents five completed research articles as a path for developing methods to actionalize implement AI ethics in software development. FIGURE 2 describes how the research questions and articles connect with relevant theoretical background components.

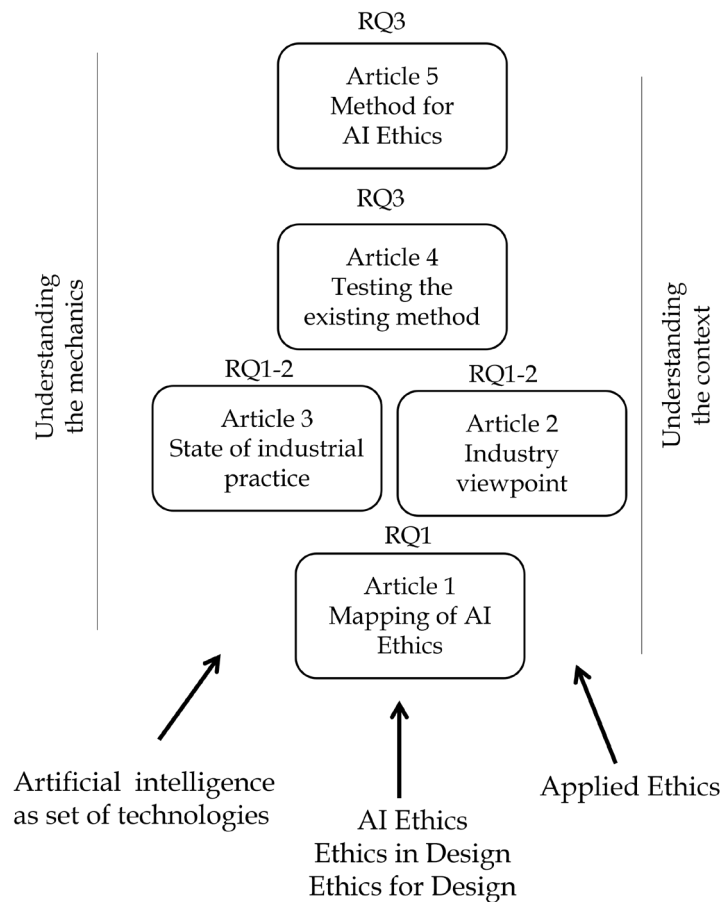


FIGURE 2 Research questions and their relationship to the included articles

The scope of this dissertation is defined by three theoretical background components: AI as a set of technologies, AI ethics, and the practical application of ethical considerations (applied ethics). In this work, AI is seen as a set of technologies, such as machine learning, natural language processing, and computer vision, that provide increasing capabilities for software systems and

hence set the context of this study. On the ethics side, this work does not debate moral theories in light of normative ethics nor does it discuss metaethical deliberation of the nature of morals. Rather, the work relates to the field of applied ethics, which addresses how to apply ethical theory in a particular situation. Following the tradition of computer ethics (Van den Hoven, 2008), which studies the moral questions that are associated with the development, application, and use of computers and the process of computer science research, this work focuses on questions concerning the development, application, and use of AI systems. The third theoretical component, AI ethics, is a vast field of study that includes research ranging from philosophical debate to specific technical improvements of algorithmic equality. To categorize the field of AI ethics, Dignum (2018) presented three categories: (a) ethics by design (integrating ethics into system behavior), (b) ethics in design (software development methods etc. to support the implementation of ethics), and (c) ethics for design (standards etc. that ensure the integrity of developers and users) (FIGURE 3). This dissertation relates to the levels of ethics in design and ethics for design. The main scope of this study is an examination of the development of software systems utilizing AI solutions and the developers of these systems. All of the aforementioned theoretical background components are discussed in detail in Chapter 2.

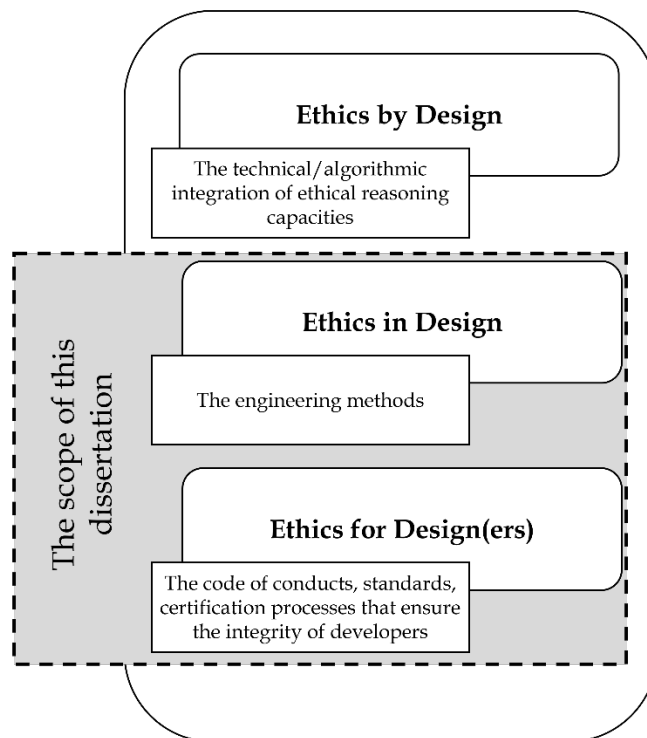


FIGURE 3 Three categories of AI ethics (Dignum, 2018)

1.3 Structure of this Work

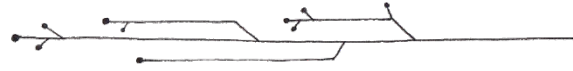
The dissertation is structured in following way. The next chapter, Chapter 2 reviews related literature and summarizes the key concepts of the dissertation. Chapter 3 describes the research process through the research design, research approach, and selected methodologies. Chapter 4 presents the results of each article. Concluding chapter, Chapter 5 discusses the results in light of their theoretical and practical contributions, as well as the limitations of this study and future research topics. Additionally, summary in Finnish, the original articles and Appendix including ECCOLA cards can be found at the end.

1.4 Other Scientific Contributions by the Author

This section lists other scientific contributions by the author related to this dissertation. These articles contribute to the field of AI ethics:

- Halme, E., Vakkuri, V., Kultanen, J., Jantunen, M., Kemell, K. K., Rousi, R., & Abrahamsson, P. (2021). How to write ethical user stories? Impacts of the ECCOLA method. In *International Conference on Agile Software Development* (pp. 36–52). Springer.
- Vakkuri, V., Jantunen, M., Halme, E., Kemell, K. K., Nguyen-Duc, A., Mikkonen, T., & Abrahamsson, P. (2021). Time for AI (ethics) maturity model is now. *SafeAI@AAAI 2021*.
- Agbese, M., Alanen, H. K., Antikainen, J., Halme, E., Isomäki, H., Jantunen, M., ... & Vakkuri, V. (2021). Governance of ethical and trustworthy AI systems: Research gaps in the ECCOLA method. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 224–229). IEEE.
- Antikainen, J., Agbese, M., Alanen, H. K., Halme, E., Isomäki, H., Jantunen, M., ... & Vakkuri, V. (2021). A deployment model to extend ethically aligned AI implementation method ECCOLA. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 230–235). IEEE.
- Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2020). ECCOLA-a method for implementing ethically aligned AI systems. In *46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 195–204). IEEE.
- Vakkuri, V., Kemell, K. K., Jantunen, M., & Abrahamsson, P. (2020) “This is just a prototype”: How ethics are ignored in software startup-like environments. In *International Conference on Agile Software Development* (pp. 195–210). Springer, Cham.
- Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019). AI ethics in industry: A research framework. In *Proceedings of the Third Seminar on Technology Ethics. (Tethics)* RWTH Aachen University. CEUR Workshop Proceedings, 2505
- Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019). Ethically aligned design: An empirical evaluation of the RESOLVEDD strategy in software and

- systems development context. In *45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 46–50). IEEE.
- Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019). Implementing ethics in AI: Initial results of an industrial multiple case study. In *International Conference on Product-Focused Software Process Improvement* (pp. 331–338). Springer.
 - Koivisto, R., Leikas, J., Auvinen, H., Vakkuri, V., Saariluoma, P., Hakkarainen, J., & Koulu, R. (2019). *Tekoäly viranomaistoiminnassa-eettiset kysymykset ja yhteiskunnallinen hyväksyttävöyys* [Artificial intelligence in authority use - ethical and societal acceptance issues]. Publications of the Government's analysis, assessment and research activities 14/2019. Prime Minister's Office.
 - Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2021). Technical briefing: Hands-on session on the development of trustworthy AI software. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)* (pp. 332–333). IEEE.
 - Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2020). Tutorial on developer-focused method for managing AI ethics in agile software development. *XP2020, 21st International Conference on Agile Software Development 2020*.
 - Vakkuri, V., & Kemell, K. K. (2019). Implementing artificial intelligence ethics: A tutorial. In *International Conference on Software Business* (pp. 439–442). Springer.



2 THEORETICAL BACKGROUND

The theoretical background for this dissertation is based on three key elements: (a) AI as set of technologies, (b) AI ethics, and (c) applied ethics. This chapter presents these three building blocks and the relevant literature that form the theoretical foundation of the research. The first section outlines the definition used for AI, the second focuses to the main theoretical context of this research, AI ethics, and the third section presents the ethical approach used, namely applied ethics.

2.1 Artificial Intelligence as a Set of Technologies

The question of what AI is can be overwhelming due to its vast number of definitions. It may not even be possible to find a precise definition that would suit all the parties and forms implicated under the banner of AI (Bringsjord & Govindarajulu, 2020). Hence when dealing with AI, researchers often need to approach the question by defining what AI is within the scope of their own research. At least three intertwining categories of definitions can be found: AI as a goal, as a field of study, and as a set of technologies. When AI is perceived as a goal, it is seen as the aim of various efforts that strive to create artificial (non-biological) computational abilities to achieve goals in the world (McCarthy, 2007). Relatively closely connected to this is the definition of AI as a field of study pursuant towards AI. Russell and Norvig (2020, pp. 1-2) described AI as a field that encompasses a huge variety of subfields, ranging from the general to the specific, that aspire for AI that is either: (a) thinking humanly, (b) thinking rationally, (c) acting humanly, or (d) acting rationally. Here, “humanly” relates to human performance, and “rationally” to right measure against an ideal performance measure (Russell & Norvig, 2020). To show the challenge of identifying a consensus definition for AI as a field, Bringsjord and Govindarajulu (2020) noted that researchers working in one of the categories of Russell and Norvig’s taxonomy can see their work providing a central component or

capability for another category. The third definition of AI, as a set of technologies, is in contrast to the previous definition's approach to AI. This is not in terms of the components that constitute the aim of AI, but in terms of what AI is when applied to software systems. This definition can be found from industry parties applying AI, such as those of IBM,¹⁰ Amazon,¹¹ and Google.¹²

From the three described definitions, the third, that AI is set of technologies, represents the understanding of AI that is applied in the context of this dissertation. Although this definition has its critics. Yet, there is common ground to be found under the umbrella of the definition. For example, as shown on (TABLE 1) the current calls for papers to top AI conferences in addition to one of the most commonly used AI textbooks, demonstrate this persistent use of AI as an umbrella term for various sets of technologies.

Additionally, it should be noted that in the context of this dissertation AI technologies are seen as software. The adaption of AI technologies as part of software systems has clouded established practices in software engineering, as if AI would be understood to be something different than software in its core. For example, the lack of emphasis placed on transparency in the Blackbox AI solution (e.g., Rudin, 2019) has become the accepted way of working, discarding the decades-long tradition that understanding the inner workings of a system is considered the key to any software engineering endeavor. Ultimately, software systems that utilize AI technology are still software and are affected by largely the same requirements as any other software system (Mikkonen et al., 2021; Sculley et al., 2015).

¹⁰ <https://aws.amazon.com/machine-learning/what-is-ai/>

¹¹ <https://www.ibm.com/design/ai/basics/ai/>

¹² <https://ai.google/about/>

TABLE 1 Set of artificial intelligence technologies

IJCAI-19 ¹³	AAAI21 ¹⁴	Russell and Norvig, 2020
-	Computer vision	Computer Vision
Knowledge representation	Knowledge representation	Knowledge, reasoning, and planning (logical agents, first-order logic, inference in first-order logic, knowledge representation, automated planning)
Machine learning	Machine learning (deep learning, statistical learning, etc.)	Machine learning (learning from examples, learning probabilistic models, deep learning, reinforcement learning)
Multiagent systems	Multiagent systems	Multiagent decision-making
Natural language processing	Natural language processing	Natural language processing (deep learning for natural language processing)
Perception	Perception	Communicating, perceiving, and acting (natural language processing, deep learning for natural language processing, computer vision, robotics)
Planning	Planning	Automated planning
Reasoning	Reasoning	Uncertain knowledge and reasoning (quantifying uncertainty, probabilistic reasoning, probabilistic reasoning over tie, probabilistic programming, making simple decisions, making complex decisions)
Robotics	Robotics	Robotics
Search	Search	Problem-solving (solving problems by searching, search in complex environments, adversarial search and games, constraint satisfaction problems)
Constraint satisfaction	Data mining	-
-	Human-in-the-loop AI	-

¹³ <https://www.ijcai19.org/call-for-papers.html>

¹⁴ <https://aaai.org/Conferences/AAAI-22/aaai22call/>

2.2 AI Ethics and Methods Used in AI Ethics

Scholars have been interested in AI ethics since the birth of computer ethics. In the past, the debate has been limited to a small number of scholars who have focused on hypothetical future scenarios that would result from technological progress. Over the last decade, these hypothetical future scenarios have started to become reality, and the topic of AI ethics has grown increasingly active and appealing for researchers (Borenstein et al., 2021).

Much of the AI ethics research has focused on theory, and specifically on solving ethical issues through guiding principles. Amid the abundance of proposed principles, some of these have become largely agreed-upon, as Jobin et al. (2019) discovered in their profound mapping of principles and guidelines on ethical AI. They reported that although various principles had been interpreted differently across a number of documents, the five most prominent of them are transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al., 2019). Similarly, Morley et al. (2021) revealed in their analysis of existing AI ethics guidelines the key principles that may be considered to be central, based on their occurrence in the guidelines. These are “transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity.” (Morley et al., 2021)

The principle of transparency provides a good example of the work that has been conducted in relation to these principles. Transparency can be considered one of the central AI ethical principles (Dignum, 2017; Jobin et al., 2019; Morley et al. 2021). For example, it is included as a key principle in the high-profile guidelines of the EU (HLEG, 2019) and the IEEE (IEEE Global Initiative, 2019). Transparency has been approached at least from the viewpoints of understanding how AI systems work and how they have been developed (Ananny & Crawford, 2018; Dignum, 2017). Transparency has even been argued to be the very foundation for ethical consideration in AI ethics. If an understanding of how a system works is not given, the system cannot be made ethical (Turilli & Floridi, 2009). There has also been discussion on how to achieve transparency in AI systems and what transparency means for such systems. For example, Ananny and Crawford (2018) discussed the limitations of transparency in terms of system complexity brought on by machine learning. Nevertheless, although there has been emphasis on the principle of transparency, there also exists criticism for naming it as a key principle. Core principles encapsulate intrinsic values, in contrast to instrumental principles (including transparency), which gain their value from instrumental tools for promoting intrinsic values (Canca, 2021).

The discussion of principles is ultimately just an avenue for bringing attention to potential ethical issues proposed by AI-powered systems. Privacy issues, for example, have been among the most noticeable topics of discussion both in academia and the media following various examples of ethical failures

and failures of AI systems. Privacy issues have been discussed in relation to data handling, and challenges faced in the application of technologies such as facial recognition. However, the privacy issues are hardly a matter of discussion that is unique to the field of AI ethics, as the past studies for example in intrusiveness of IT systems (Introna 1997) and information privacy (Floridi 2006) have highlighted outside the context of AI ethics discussion.

Guidelines have gained popularity for bridging the gap between research and practice and as a tool for distilling discussion in academia and industry. However, past research has shown that guidelines only have limited or potentially no effect as a tool for influencing software engineering. McNamara et al. (2018) studied the impact that the ACM Code of Ethics¹⁵ has had on practice in the area, finding that it was little to none. In her promotion of codes of ethics as a solution for AI ethics issues, Boddington (2017) noted that such codes are not a straightforward solution, as they can also make the situation worse in various ways, such as by creating moral passivism or leading ethical consideration to be outsourced to somebody else. Canca (2021) highlighted that principles themselves are of little use to developers without their prior knowledge of philosophical considerations. This gap between principles and practice and the issues with guidelines was also acknowledged by Johnson and Smith (2021) in their gap analysis.

Morley et al. (2021) highlighted in their systematic review of the field that there are already various methods and tools for implementing AI ethics. The 104 reported tools in their study are largely methods for the technical side of AI systems development, such as machine learning toolsets. The study pointed out two main challenges regarding AI ethics tools: (a) the researched tools were immature, as they lacked usability and required adjustment in practice to suit the needs of developers; and (b) existing AI ethics tools and methods are not evenly distributed across the applied AI ethics typology (TABLE 2). Methods encompassing the whole development process do not exist. For instance, many of the tools are intended for application during a specific phase of development (Morley et al., 2021).

¹⁵ <https://www.acm.org/code-of-ethics>

TABLE 2 Typology of AI Ethics Methods (Morley et al., 2021)

	Business and use-case development	Design phase	Training and testing data procurement	Building	Testing	Deployment	Monitoring
Beneficence, Non-maleficence, Autonomy, Explicability, Justice	Problems or improvements are defined and use of AI is proposed.	The business case is turned into design requirements for engineers.	Initial data sets are obtained to train and test the model.	AI application is built.	The system is tested.	The AI system goes live.	Performance of the system is assessed.
<p>104 tools/methods were assigned to correspond to the intersection of ethical requirements and ML algorithm development to recognize coverage of existing tools/methods and guide developers in finding tools to aid in ethical considerations.</p>							

The review of AI ethics tools and methods by Morley et al. (2021) was conducted in 2019, after which time work on AI ethics methods has continued. For example, the usability of existing guidelines has been evolved (ALTAI, 2020). Yet, many of the issues remain (Ayling, 2021). New frameworks have been developed outside the software engineering domain, such as design-focused approaches (Peters et al. 2020) or developed outside the academic community without empirical validation (e.g., IDEO's AI Ethics Cards¹⁶). From the software engineering domain new preliminary study of method (RE4AI Ethical Guide) addressing AI ethics through ethical requirements has been published expressing the need for software development team and project level tools (Siqueira de Cerqueira et al. 2022)

2.3 Applied Ethics Approach to AI Ethics

In this dissertation, ethics are seen through the lens of applied ethics, which is to be distinguished from other fields of ethics such as metaethics (nature of moral truth, language, judgments) and normative ethics (schools of thought/ethical theories to identify moral truths) (Archard & Lippert-Rasmussen, 2013). Applied ethics, also sometimes referred to as practical ethics, seeks answers to practical questions regarding what to do in a given context at individual, professional, and societal levels (Tännsjö, 2011). Subfields of applied ethics include medical ethics, bioethics, business ethics, environmental ethics, and professional ethics (Archard & Lippert-Rasmussen, 2013). For example, one of the most prominent subfields of applied ethics, bioethics, focuses on the ethical implications and applications of issues related to healthcare and life sciences, from policy development to clinical ethics (Flynn, 2021). This dissertation continues the applied ethics tradition of computer ethics (Bynum, 2006; Moor, 1985). AI-powered software development is approached in the same manner as that of computer ethics, whereby it studies moral questions that are associated with the development, application, and use of computers and computer science (Van den Hoven, 2008). Also in this dissertation, the relevant themes of professional ethics are addressed from the developer's point of view. Professional ethics is required when members of a profession possess certain capacities that others lack. In the case of AI, technical knowledge gives AI system developers power and authority over others through technical means (Boddington 2017, pp. 39, 59).

In this dissertation, applied ethics are presented on two levels, since the main research question (*How to implement AI ethics in software development?*) has two main components, namely "How to implement" and "AI ethics in software development." It can be said that asking *How?* is connected to the mechanical level and the concept of *AI ethics software development* connects to the contextual level (FIGURE 4). The term "mechanical level" refers to the level of ethical analysis in applied ethics, such as how to perform an ethical analysis (Allhoff,

¹⁶ <https://www.ideo.com/post/ai-ethics-collaborative-activities-for-designers>

2011), how to turn principles to practices (Canca, 2021; Morley et al., 2021), and how to justify actions taken in a given context (Pfeiffer & Forsberg, 1993). The term “contextual level” refers to the specific context from which ethical issues are raised – in the case of this study, it refers to the development of AI systems. More detail on the context can be found in Section 2.2.

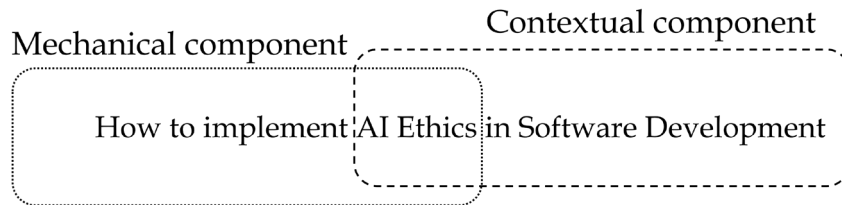
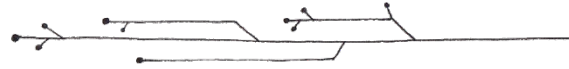


FIGURE 4 Two components of ethical approach

This dissertation seeks to understand how ethical issues that arise from a given context (development of AI systems) can be addressed by means of applied ethics, such as by providing tools for developers to: (a) formulate reasoned and robust conclusions with regard to whether a certain practice is morally right or wrong; and, (b) describe the relevant conflicting moral considerations. The historical approach to solving issues in applied ethics has been to select theories from normative ethics and then apply those theories to specific cases, to thus enable the application of ethics (Van den Hoven, 2008). When operating in a real-life context, the questions of ethical challenges and dilemmas are more diverse than isolated philosophical thought experiments. They are therefore made to be relevant aspects of applied ethics through recognition of the problem, asking how the problem came into being, and considering if it could have been solved by means of design (Van den Hoven, 2008).



3 RESEARCH METHODOLOGY

This chapter presents the methodological aspects of the dissertation. The section begins with an introduction to and justification of the selected research approaches and their theoretical foundations. This is continued by a description of the collection of empirical evidence and its analysis.

3.1 Research Approach and Design

The research approach follows the tradition of an empirical software engineering established by Kitchenham et al. (2004), Rombach (2013), and Basili (2013), where a evidence-driven approach includes scientific use of quantitative and qualitative data to understand and improve software engineering. Empirical software engineering addresses questions such as how empirical observations mature into empirical laws and theories. Moreover, it calls for software engineering results and practices to be critically evaluated in light of the scientific evidence (Rombach, 2013). Likewise, Basili (2013) emphasized that no technique should be published without it first being tested, and the trial application of new ideas should not solely involve having data to demonstrate their application. Rather, it should also open up avenues for identifying boundaries and limits, in addition to pointing out ideas for improvement.

This dissertation seeks to understand how to implement AI ethics in software development and instill them within the practices of developers by adopting the theoretical approach of an empirical software engineering. In this approach, knowledge creation is understood to be a pragmatic cycle: the iterative process of generating empirically grounded and tested theories from a repetition of induction, abduction, and deduction until a useful theoretical maturity has been reached (Fernández & Passoth, 2019). The pragmatic cycle (FIGURE 5) is based on the two underlying assumptions of empirical research: (a) There is no one universal way of scientific practice, but rather multiple ways of undertaking research; and (b) No single empirically inquired point of view will ever provide

us with an entire picture when interpreting relevant phenomena (Fernández & Passoth, 2019). Therefore, a cycle of approaches and combination of techniques is needed. In the context of this dissertation, there are three approaches included to complement an empirically grounded understanding of the implementation of AI ethics in software development. These approaches are: (a) creating theoretical knowledge, (b) applying theoretical knowledge, and (c) testing theoretical knowledge. TABLE 5 describes how the five different articles in this study are related to a range of theoretical approaches. Theoretical knowledge was created by multiple means, via reviewing and synthesizing theoretical knowledge (Article I) and through induction from empirical evidence (Articles II, III, and IV). Theoretical knowledge was applied and tested when existing ethical tools were implemented (Articles IV and V) and when existing frameworks of AI ethics were reflected with empirical evidence from industry.

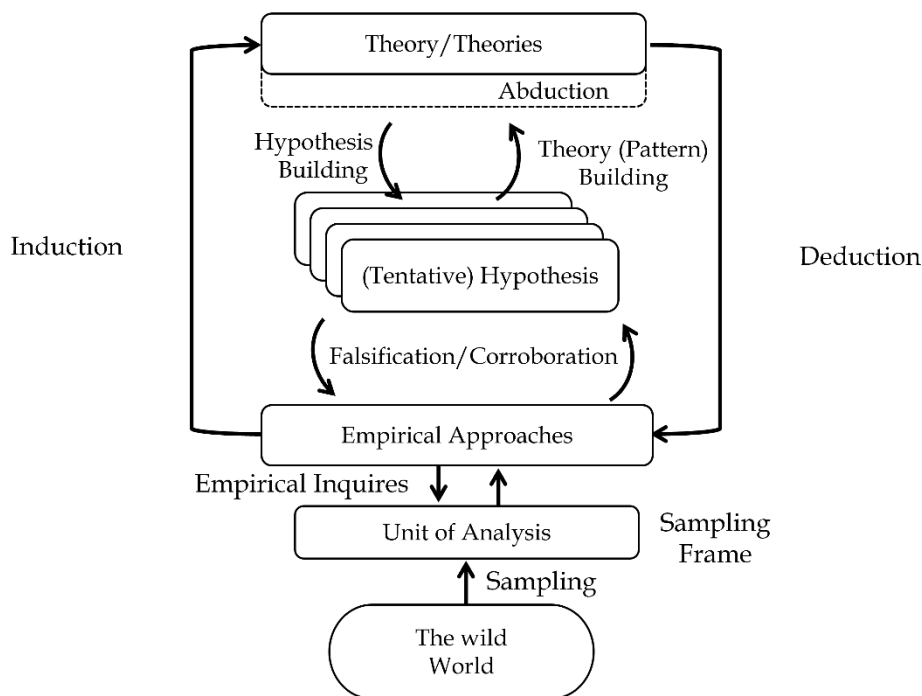


FIGURE 5 Pragmatic cycle of empirical research (Fernández & Passoth, 2019)

TABLE 3 Methodological aspects of articles

<i>Article</i>	<i>Approach</i>	<i>Method</i>
1	Creating theoretical knowledge	Literature review: Systematic mapping study
2	Creating theoretical knowledge	Qualitative research: Case study
3	Testing and creating theoretical knowledge	Mixed methods: Survey
4	Applying and testing theoretical	Qualitative research: Case study
5	Applying, testing, and creating theoretical knowledge	Qualitative research: Cyclical action research

In addition to a systematic mapping study (Article I) that has served to provide structure and insight regarding the research area, this dissertation has applied the methodological approaches of case study, survey, and action research to form an iterative process for generating empirically grounded and tested knowledge. The application of ethical considerations in systems development is an ongoing contemporary phenomenon, whereby often the actions taken in the real-life context give rise to the ethical issues. In this light, the case studies (Articles II and IV) provided an apt methodology for their capacity to address contemporary phenomena outside the isolation of a laboratory environment (Runeson & Höst, 2009). To complement the deeper understanding of the case studies, a survey (Article III) provided a wider perspective for the phenomenon. As a method, a survey utilizes a large population to develop theoretical and analytical models to explain behavior, attitudes, or values (Joye et al., 2016). Though surveys are often seen as tools to quantify given phenomena, excluding qualitative aspects (see e.g., Kitchenham & Pfleeger, 2008), this dissertation applied a mixed-method approach to surveys by including open-ended questions to address the contemporary nature and immature state of AI ethics. Since the objective of this research was to present empirically tested theoretical knowledge in the form of a solution, the method of cyclical action research (Article V) was used to combine and further develop the knowledge provided by other methods used in this dissertation. Susman and Evered (1978) described action research as being a well-suited process for using different data collection methods across different contexts. The application of action research involves solving organizational problems (out-of-lab problems) through intervention while at the same time contributing to knowledge (Davison et al., 2004). Hence, action research was not just a method for applying, testing, and creating theoretical knowledge for the context of this study (industry), but also a means to produce relevant results.

3.2 Data Collection - Empirical Evidence

The collection of empirical evidence was done using a range of data collection tools over several phases between 2017 and 2021. Each article used different data

collection instruments with a specific focus on the field of AI ethics. The data collection methods, collection times, data types, and quantities are presented in TABLE 4.

TABLE 4 Collected data

Article	Collection of data	Data collection time	Unit of analysis (N)
1	Keyword search in selected scientific databases	2017	Retrieved papers: 1062, included papers: 83
2	Semi-structured interview	2018	6 respondents from 5 software companies
3	Survey instrument including demographic, Likert scale, and open-ended questions.	2018-2019	249 respondents from 211 software companies,
4	Semi-structured interview	2018	5 groups of 4-5 students
5	Semi-structured interviews, researcher's and user's notes, work-products (of ECCOLA users), unstructured participant interviews, workshop recordings, unstructured software developer interviews, project documentation	2018-2021	6 action cycles

Article I followed the systematic mapping study (SMS) method (Kitchenham & Charters, 2007; Petersen et al., 2015) for formulating search strings to collect and classify research papers from academic databases. SMS was chosen as the research method due to its capability to deal with the wide and loosely defined areas of study that comprise AI ethics. SMS aims to produce an overview of the field and reveals which topics have been covered to a certain extent.

Data collection for Articles II and IV was based on qualitative semi-structured interviews (Yin, 2015). Two sets of interviews were conducted. To understand the awareness of AI ethics issues and existing practices in AI ethics in an industrial context, the first set of interviews included respondents from companies that utilized AI solutions (Article II). For comprehension of the feasibility of existing ethics methods in an AI ethics context, the second set of interviews included student project members of a university course where students were developing a prototype for a futuristic innovation (Article IV). The flexible format of semi-structured interviews afforded the possibility to gain information from outside the search framework while still providing a clear structure for the interviews. This exploratory approach was seen to be highly important, as the research topics were novel with only a limited amount of existing literature available. Also, using multiple cases made it possible to have multiple data sources with rich in-depth analysis within each case and cross-referencing across the cases to validate the observations (Yin, 2015).

The survey data collection for Article III was conducted in multiple ways. It was undertaken either through face-to-face structured interviews (interviewer asking survey questions and writing down the answers) or via an online survey to enable an understanding of the current state of industrial practice in AI ethics.

This was further by probing how AI ethics guidelines have been adopted by the software industry for developing AI solutions. As the aim was to gain a broad industry viewpoint, the survey was seen to be the most effective and feasible data collection tool. It was also convenient for the respondents, as it required less of their time than in-depth interviews. The survey instrument was built upon known principles of AI ethics, namely transparency, accountability, responsibility, and predictability. It also included three types of questions: demographic questions about the organization, Likert scale questions, and open-ended questions. In the Likert scale questions, the participants were told to evaluate the importance of AI ethics principles and were asked practical questions, such as whether they had faced ethical issues with their software. The data were collected from different sized organizations from 20 countries. Most of these companies were either US or Finland based.

For Article V, data collection was conducted over a period of four years using multiple data collection tools, which were semi-structured interviews, researcher and user notes, work products (for ECCOLA users), unstructured participant interviews, workshops recordings, unstructured software developer interviews, and project documentation. As the method used in Article V, cyclical action research (Davison, 2004; Susman & Evered 1978), is responsive and reflective, the data collection toolset also needed to react to changes in cycles. Therefore, not only the best suited actions but also the best suited data collection tools were selected for each cycle in Article V. In the cyclical action research process, the role of data collection was to provide material for the evaluation and reflection of the quality of the performed intervention to improve the method in development. User data for the action research were collected in an augmenting manner, starting from student testing and progressing through academic workshops to make the method mature before industry testing.

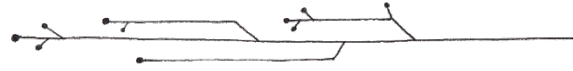
3.3 Data Analysis

This dissertation has utilized four analysis methods: content analysis (Article I), grounded theory (Articles II and V), thematical analysis (Articles III and V), and descriptive analysis (Article III). Descriptions of the various article analysis processes can be found in TABLE 5. The content analysis method allows the sorting, quantifying, and comparison of versatile datapoints into recurring themes or categories. It is therefore suitable for the mapping process described in the method guides of Kitchenham & Charters (2007) and Petersen et al. (2015). Grounded theory was used based on its capacity to analyze and organize interview data as well as pictures and diagrams (see the “All is data” approach of Holton & Walsh 2017) for new insights and theory-building with minimal advanced knowledge of the phenomenon (Stol et al., 2016). Thematic analysis, as reported by Cruzes & Dybå (2011), is one of the most commonly used analytical methods in software engineering research. It was used in this dissertation in cases where the limited interpretative power of thematic analysis could be supported

by the existing theoretical framework. The survey's descriptive analysis (Fisher & Marshall, 2009) was used to provide numerical and graphical presentations and summarize the information before comparing it with the open-ended questions.

TABLE 5 Analysis, methods, and processes

Article and method	Aim of analysis	Analysis process
1: SMS	Identifying key AI ethics concepts	Extracting, reading, comparing, quantification, and categorizing keywords for recurring themes to draw conclusions
2: Case studies	Recognizing AI ethics topic and existing ethical practices	Stage 1: Interview data transcription, reading, coding (open and selective), comparing codes, and categorizing codes; finding relationships among categories and theory-building Stage 2: Cross-referencing across cases to validate the observations
3: Survey	Describing survey sample	Collecting and describing survey data in graphical form; comparison of open-ended questions
4: Case studies	Recognizing well-performing and underperforming features	Stage 1: Interview data transcription, reading, coding, comparing codes and categorizing codes; Stage 2: Cross-referencing across cases to validate the observations
5: Action research to develop method	Providing insights for the evaluation and reflection of quality of the performed intervention	Iterative process of diagnosis, action planning, intervention, evaluation, reflection; e.g., grounded theory was used to analyze each cycle's data sources for insights



4 OVERVIEW OF THE ARTICLES

4.1 Article I: The Key Concepts of Ethics of Artificial Intelligence

Vakkuri, V., & Abrahamsson, P. (2018). The key concepts of ethics of artificial intelligence. In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)* (pp. 1-6). IEEE.

Research Objectives

As described in the background chapter, the fields of AI and AI ethics research are by nature versatile and multidisciplinary, including a wide range of different approaches and viewpoints to the topic of AI ethics. To understand this field, common vocabulary is needed. Without knowledge of what AI ethics is, the goal of implementing it is a hard or even pointless task. In this article, conceptualization of the main concepts was seen as a solution for understanding and progressing the field of AI ethics. The paper claimed that shared concepts paved the way for the practical implementation of the ethics of AI.

Before taking the first steps towards conceptualization, the main concepts used in AI ethics need to be identified. Therefore, the goal of this paper was to identify and categorize keywords used in academic papers in AI ethics discourse. The focus was on the author's selected keywords as the best means to describe important concepts from author's own viewpoint. A keyword-based systematic mapping study of the keywords used in AI and ethics was conducted to help identify, define, and compare the main concepts used in the current AI ethics discourse. The study was conducted with the following focuses:

1. Recognizing keywords used in the field
2. Extracting potential keywords for future research
3. Comparing keywords to proposed concepts in academic literature.

Findings

The keyword-based mapping study identified 37 reoccurring keywords in 83 academic papers focusing on AI and ethics. Based on keyword listing, three main patterns were discovered: (a) a lack of different branches of AI in keywords; (b) the minor role of technology-based keywords; and, (c) a great variance in the formulation of keywords, even though they could be classified under known topics. A lack of different branches of AI, such as machine learning, natural language processing, and pattern recognition, in the listed keywords seemed to imply that if not nonexistent, the relevant discussion of ethical aspects was done under separate AI branches rather than the broader topic of AI. This would also be a fitting explanation, as the subfields of AI are known to be rather independent. The minor role of the technology-based keywords in the list could be seen as a sign that the discussion on AI ethics was held on a more abstract level. Furthermore, this was seemingly not among researchers from a technological or technical perspective, at least during the time window of the mapping (2012–2018). The variance in the formulation of keywords touching upon the same topic showed that the use of AI ethics-related keywords was not settled, and therefore, the concepts and shared vocabulary is still in formation. Overall, the study revealed that defining the field of AI ethics is still a challenging task and that the immaturity of AI ethics discussions (at the time of the study) could pose major challenges for implementing ethics, as the discussion of what should be implemented is still open.

Connection to the Objectives of the Dissertation

The mapping study of AI ethics keywords contributed to answering Research Question 1 by providing a review of how AI ethics is seen as a field of research. The results of this study provided background knowledge and a starting point for Articles II-V.

4.2 Article II: Ethically Aligned Design of Autonomous Systems: Industry Viewpoint and an Empirical Study

Vakkuri, V., Kemell, K. K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2022). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. arXiv: 1906.07946. To be published in *Journal of Business Ethics and Organization Studies EJB*.

Research objectives

Both the academic and public discussion of AI ethics have accelerated, but the state of practice in the area remains unclear. To better understand the gap between theoretical contributions and practices in AI ethics, a multiple case study of five case companies was conducted. Specifically, the focus was on the current industry mindset in relation to AI ethics from the point of view of some of the most common AI ethics principles discussed in AI ethics guidelines. The goal of

this study was to understand which practices, tools, and methods, if any, industry professionals utilized to implement ethics in AI design and development.

Findings

Exploring the industry mindset in relation to AI ethics revealed multiple findings that can be utilized when aiming to implement AI ethics in practice. This study provided insights from practice in three ways by; (a) providing new knowledge for a novel research area, (b) empirically validating the existing literature, and (c) showing evidence that contradicted existing literature. These findings are reported in nine primary empirical conclusions (PEC) in TABLE 6. To summarize these findings, the gap in academic discussion and industrial practice is prominent, as none of the case companies utilized any guidelines, tools, or methodologies to implement AI ethics. Developers considered ethics to be important in principle but considered it impractical and distant from the issues they faced in their work. Despite there being various existing AI ethics guidelines, none of these were used by the industry experts, who rather considered them to be unactionable. In contrast to the popular trend of compiling different AI ethics guidelines, such guidelines may not be the way to proceed if the aim is to aid AI systems developers. To implement AI ethics guidelines in practice, the main goal should first be to make them actionable.

Connection to the Objectives of the Dissertation

By providing an industry viewpoint with practical insight on AI ethics, this study contributed to Research Question 2. This study also served as a pre-study for the survey in Article III, and the results of this study were used during the diagnostic stage of the first action cycles for Article V.

TABLE 6 Primary empirical conclusions of the study

PEC	Theoretical component	Description	Contribution
1	Conceptual	Ethics is considered important in principle, but as a construct it is considered detached from the current issues of the field by developers.	Empirically validates existing literature
2	Conceptual	Regulations force developers to take into account ethical issues while also raising their awareness of them.	Empirically validates existing literature
3	Transparency	Developers have a perception that end-users are not tech-savvy enough to gain anything out of technical system details.	Contradicts existing literature
4	Transparency	Documentation and audits are established software engineering project practices that form the basis for producing transparency in AI/AS projects.	Empirically validates existing literature
5	Transparency	Machine learning is considered to inevitably result in some degree of unpredictability. Developers need to explicitly acknowledge and accept heightened odds of unpredictability.	Empirically validates existing literature
6	Responsibility, accountability	Developers consider the harm potential of a system primarily in terms of physical harm or harm towards humans.	New knowledge
7	Responsibility, accountability	Physical harm potential motivates personal drivers for responsibility.	Empirically validates existing literature
8	Responsibility, accountability	Main responsibility is outsourced to the user, regardless of the degree of responsibility exhibited by the developer.	New knowledge
9	Responsibility, accountability	Developers typically approach responsibility pragmatically from a financial, customer relations, or legislative point of view, rather than an ethical one.	New knowledge

4.3 Article III: The Current State of Industrial Practice in Artificial Intelligence Ethics

Vakkuri, V., Kemell, K. K., Kultanen, J., & Abrahamsson, P. (2020). The current state of industrial practice in artificial intelligence ethics. *IEEE Software*, 37(4), 50–57.

Research Objectives

This study sought to understand whether the public and academic AI ethics discussion has had an impact on the AI industry on a wider scale, and if AI ethics guidelines have been adopted by industry. During the time of the survey, no other surveys utilizing data from company respondents on the current state of practice in AI ethics existed. The existing surveys relied on public opinion or on document data such as guidelines or project documentation. This survey had two high-level goals: (a) to help understand the state of the industry in terms of AI ethics; and (b) to provide data for benchmarking where different organizations stand in relation to AI ethics. Additionally, among the objectives of the survey was a comparison of how well-versed AI companies were in AI ethics compared to other software companies.

Findings

The survey data collected from over 200 companies revealed that AI ethics implementation was still in its infancy. The comparison of AI companies and other software companies resulted in largely similar responses. Based on geographical data, there were no notable differences in the survey. Overall, the responses indicated a mixed level of maturity in terms of implementing AI ethics. Responses to some of the questions directly indicated immaturity in relation to AI ethics. For example, over one-third of respondents skipped or answered “I don’t know” to the liability question, which implied that it was an unfamiliar or overlooked theme. For the question on misuse, most organizations (51%) felt their system could not be misused. Nevertheless, there were still questions that indicated some level of maturity. For example, on predictability, the companies indicated more concern towards AI ethics-related issues.

Looking at the questions related to the guideline topics, it seems that the various AI ethics guidelines have not had a notable impact on practice. For example, many respondents either said that they did not have a fallback plan in place for unexpected system behavior or that they did not know whether they had one, although many guidelines call for it. The same was seen in relation to transparency, which was understood in terms of data and algorithms but not in terms of transparency in system development. The respondents also outsourced responsibility for systems use to the users, although on the other hand they felt responsible for any harm caused by their software. Meeting mandatory regulatory standards was seen as sufficient in terms of responsibility. This would imply that although standards raise the quality of a system, they limit the space for ethical considerations by setting a level of satisfaction.

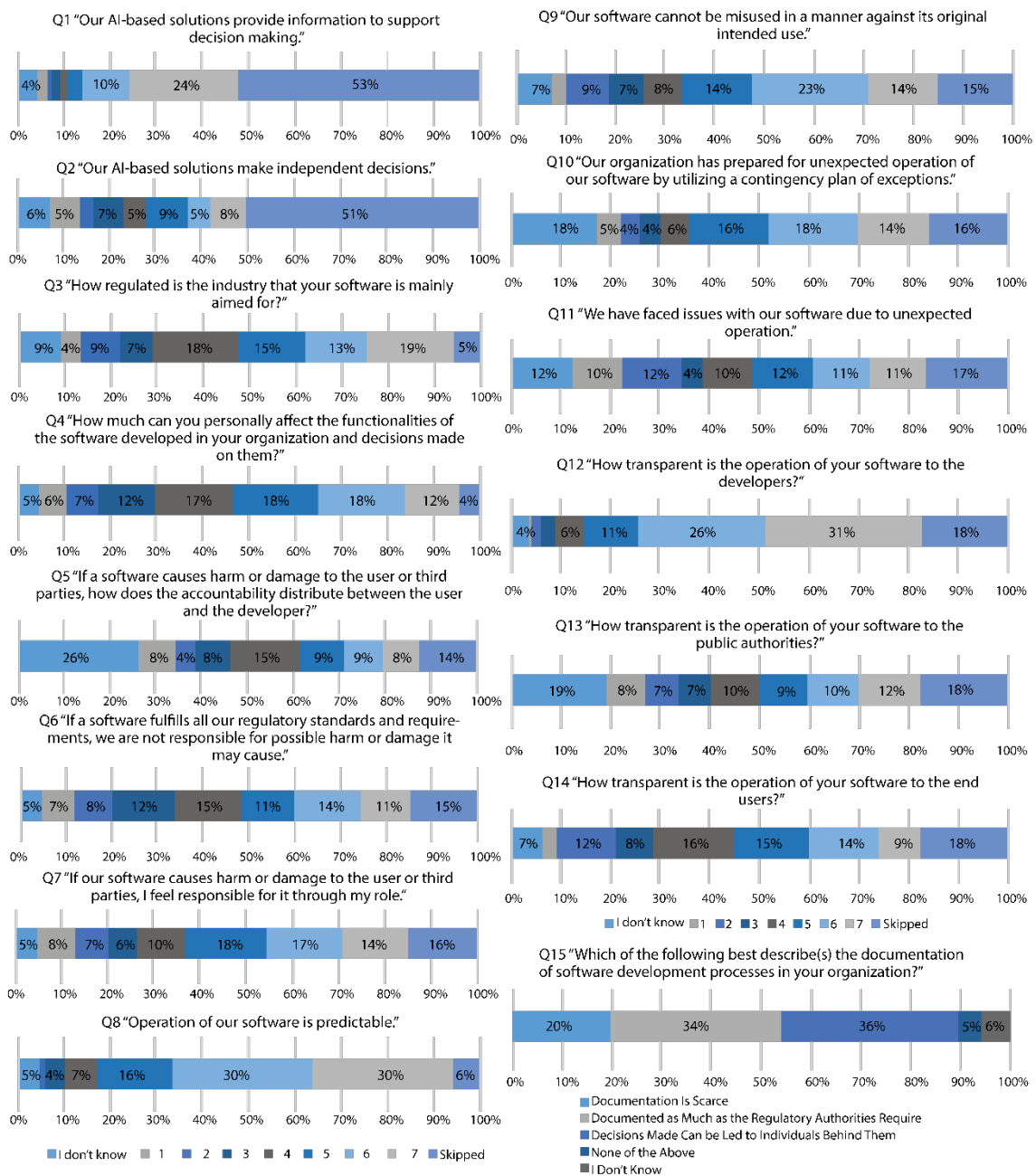


FIGURE 6 Survey questions

Connection to the Objectives of the Dissertation

The article contributed to Research Question 1 by depicting how the industry sees AI ethics and to Research Question 2 by showing the lack of existing AI ethics practices in the industry. As with Article II, this article supported the diagnostic stage of first action cycles of Article V.

4.4 Article IV: Implementing AI Ethics in Practice: An Empirical Evaluation of the RESOLVEDD Strategy

Vakkuri, V., & Kemell, K. K. (2019). Implementing AI ethics in practice: An empirical evaluation of the RESOLVEDD strategy. In *International Conference on Software Business* (pp. 260–275). Springer.

Research Objectives

Despite the calls for ethical considerations in the design and development of AI-based systems, little knowledge exists on how to provide useful and tangible tools that can help software developers and designers implement ethical considerations into practice. To address this issue, this study presented multiple case studies that assessed the use of the RESOLVEDD strategy, a tool for business ethics, in the context of AI systems design and prototyping. The goal of this case study was: (a) to understand how the introduction of an ethical tool would affect developers' ethical consideration in the design process; and (b) to empirically evaluate how the RESOLVEDD strategy, as an existing ethical tool, would serve in the given AI context. This study used three AI ethics constructs, accountability, responsibility, and transparency, as its basis. The study was undertaken to identify the possible relations between the constructs, as well as relationships to other constructs that may be involved in the process. This included ethical considerations for AI system design and prototyping.

Findings

Having introduced an existing ethical tool, RESOLVEDD-strategy to AI system design and prototyping, and by requiring its use as a system requirement, five findings were made:

1. While normative pressure for the use of ethically aligned design brings immediate results, it will cease to exist when external pressure is removed
2. RESOLVEDD increased transparency in the design process
3. RESOLVEDD did not deliver accountability
4. Requiring developers to use ethical tools increased their sense of responsibility
5. The mere presence of an ethical tool had an effect on the ethical consideration exerted by developers, creating more responsibility even when the use of the method is not voluntary.

In summary, the findings indicated that simply the presence of an ethical tool had an effect on ethical consideration, even in instances where the tool's use was not intrinsically motivated. Based on these findings and the lessons learned in the practice of using the method, the RESOLVEDD strategy formed a useful starting point. However, it needed to be adjusted to suit the context of software design and development. For example, one distinct and practical adjustment done by teams using the method was the introduction of group discussions as a primary means to include ethical consideration in their way of working.

Connection to the Objectives of the Dissertation

This study, which empirically evaluated an existing ethical tool, was used to collect the information needed for the foundation of answering Research Question 3. The results of this study served as a crucial baseline for the research behind Article V.

4.5 Article V: ECCOLA – A Method for Implementing Ethically Aligned AI Systems

Vakkuri, V., Kemell, K. K., Jantunen, M., Halme, E., & Abrahamsson, P. (2021). ECCOLA – A method for implementing ethically aligned AI systems. *Journal of Systems and Software*, 182, 111067

Research Objectives

The boom of ethical principles and guidelines, and the public demand for these, has led to the advent of questions such as how the given principles and values should be converted into requirements in practice for AI systems? And, what should professionals and organizations developing these systems do? This paper addressed the shortage of practical AI ethics tools for industry and the lack of actionability of AI ethics principles and guidelines. This was achieved by proposing a method designed to facilitate ethical thinking in AI and autonomous systems development. The objective was to develop a method capable of: (a) creating awareness of AI ethics and its importance; (b) devising a modular method suitable for a wide variety of software engineering contexts; and (c) making the method suitable for agile development.

Findings

This paper presented a method for implementing AI ethics, ECCOLA, which is based on a sprint-by-sprint process and designed to be used together with existing software engineering methods. Following the software engineering custom in which physical tools are commonly used to deploy methods in practice, the method took shape in the form of a deck of 21 cards (example card in FIGURE 7; a full card deck can be found in the Appendix) that were subsequently split into eight AI ethics themes. Following a cyclical action research approach and drawing inspiration from the EU and IEEE's EAD guidelines, ECCOLA was iteratively developed over the course of multiple years in collaboration with both researchers and practitioners.

The iterative developing and testing of the ECCOLA method led to a state where it provides a starting point for implementing ethics in AI. To address practicality, ECCOLA was designed to be incorporated into any existing method. Based on data from testing the method, its users preferred taking the approach of applying it together with existing methods. It seemed to work in agile development, as the companies utilizing it were all agile and had no issue incorporating the method into their way of working. This idea of conjoining

methods was designed to lower the barrier to adoption for an ethical tool. While addressing awareness of AI ethics and providing tools for users to tackle ethical issues, ECCOLA does not provide any direct answers to ethical problems. Rather, it asks questions that would make the organization consider various ethical issues related to AI systems. In this way, ECCOLA empowers its users in terms of enabling and inspiring ethical thinking.

Based on the user testing, it can be argued that ECCOLA facilitates the implementation of AI ethics in two confirmable ways. First, it raises awareness of AI ethics by making its users aware of different ethical issues and facilitating ethical discussion among the teams using it. This could be seen from the notes of ECCOLA users during the different stages of its development, as well as from the discussions and interviews had with its users. Second, ECCOLA produces transparency in systems development. When utilizing the method, the users were asked to produce documentation of their ethical decision-making by means of making notes, providing documentation to the existing project documentation platform, or by forming non-functional requirements for the project's needs. Therefore, compared to a baseline where no ethical methods are used, it can be argued that ECCOLA increased ethical consideration during development.

Connection to the Objectives of the Dissertation

This paper fulfilled the research objective described in Research Question 3. In this study, a comprehensive answer to how can industry be supported in addressing its AI ethics concerns was given in the form of the ECCOLA method.

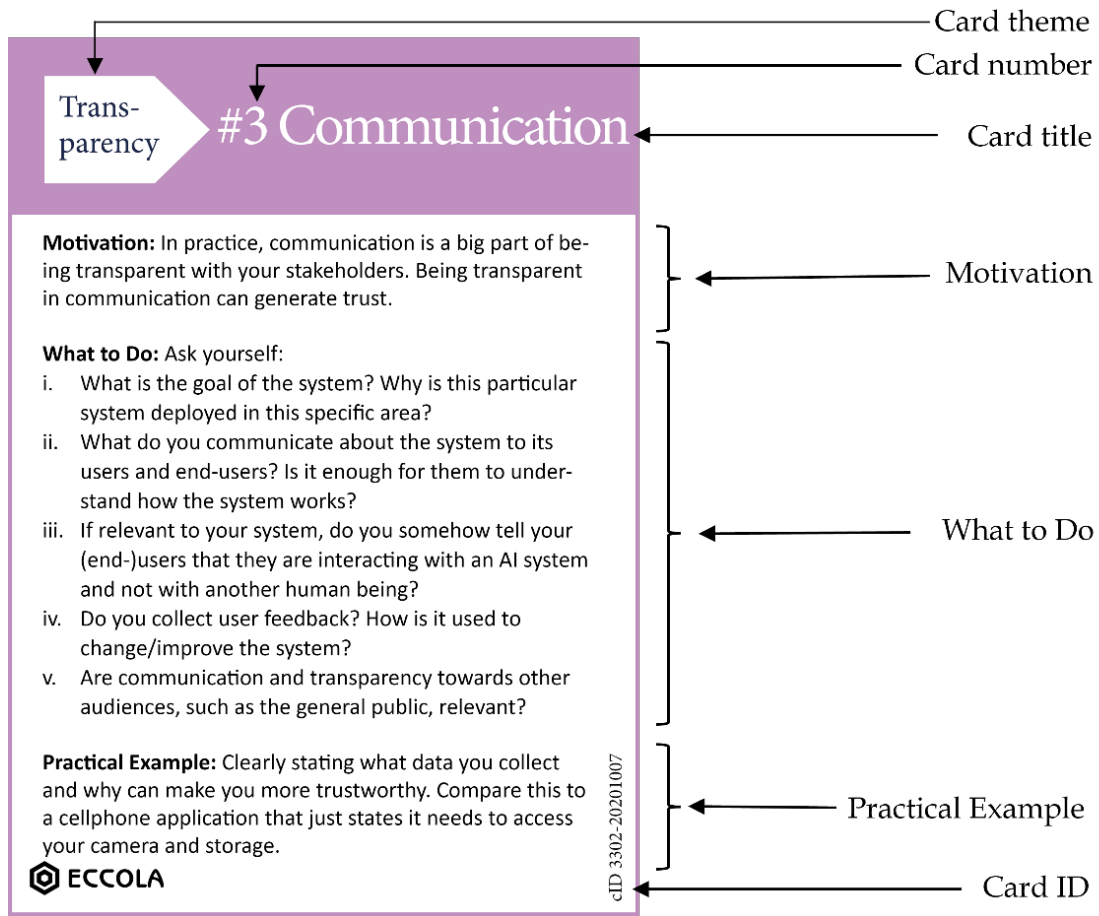
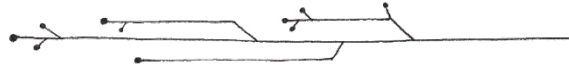


FIGURE 7 Example card from the ECCOLA method (Card #3 Communication)



5 RESULTS AND CONTRIBUTIONS

In this section, the results of the research are summarized and the contributions to the literature and to practice are outlined. Finally, the limitations and further research ideas are presented.

5.1 Results

The key results of this dissertation aim to fulfill the research objective of providing answers to the main research question (how to implement AI ethics in software development?) in the form of an empirically grounded method. In the following sections, the results are summarized under the relevant research questions.

1. How can the field of AI ethics be made sense and organized?

The mapping of AI ethics-related keywords revealed that although the AI ethics discussion is active, it still remains at an immature level. Related keywords have not been established and shared concepts and vocabulary are lacking (Article I). Notable too is the lack of technically oriented keywords in relation to the concept of AI ethics. This implies that the discussion of AI ethics is held on a more abstract level or without researchers from a technological or technical perspective (Article I).

2. How is the software industry implementing ethics today?

The industry cases revealed that AI ethics is seen as important in principle, but this does not carry over into practice. Indeed, ethics is considered to be impractical and distant from the issues that companies and professionals face in their work (Article II). AI ethics guidelines provide some understanding of AI ethics to the industry, but the guidelines themselves are not used by industry experts, as they are seen to be unactionable (Articles II and III). The current state of the industry in terms of AI ethics implementation is still in its infancy. Although, companies with a progressive attitude toward the topic do exist. This

implies some mixed maturity in the implementation of AI ethics (Article III). The gap between academic discussion and industrial practice is prominent. None of the case companies utilized any guidelines, tools, or methodologies to implement AI ethics (Articles II and III). The main principles from AI ethics guidelines have not had a notable impact on practice (Article III). Moreover, industry values guidelines that are made actionable for developers (Article II).

3. How can the industry be supported in addressing its AI ethics concerns?

Empowering developers and development teams to address ethical considerations in their work is a valid way of implementing ethics in software development (Articles IV and V). Providing means to transform abstract ethical considerations into tangible concepts and actions is an effective tool, as simply the presence of an ethical tool influences ethical consideration in a positive manner (Article IV). To further enhance the acceptance of ethically-driven methods, the tool should be tailored to suit software design and development - the given context of this study (Articles IV and V).

The following characteristics for a method to aid the implementation of AI ethics in software development were recognized. (a) Practicality and simplicity are highly valued. The method should work together with existing software engineering methods, which lowers the barrier to adoption for ethical methods. (b) The method should be modular/agile. Software development processes and ways of working vary, and a method that cannot adapt is seen to be irrelevant from the viewpoint of the primary task, which is software development. (c) Capability to empower users. By facilitating users, such as by providing the necessary vocabulary to address ethical considerations, they recognize their own role in the implementation. (d) Implementation of ethics should be measurable or auditable. The method should provide transparency to the process, such as in the form of documentation (Article V).

To address these requirements for an implementation method, and as a result of the iterative process of diagnosis, planning, intervention in an empirical setting, and evaluation, a method for implementing AI ethics in software development, ECCOLA, has been developed and tested (Article V).

5.2 Validity Threats

This section addresses the limitations of the dissertation through validity threats. Runeson and Höst (2009) defined validity as, "trustworthiness of the results, to what extent the results are true and not biased by the researchers' subjective point of view." Several categorizations for validity threats and their evaluation exist in the field of software engineering, based on different worldviews and methods used (Petersen & Gencel, 2013). In this dissertation, the validity categorization of Runeson and Höst (2009) is followed due to its applicability for empirical software engineering research. In this categorization, validity consists of reliability, construct validity, internal validity, and external validity.

5.2.1 Reliability

Reliability describes the extent to which the research itself is independent from the researchers conducting it (Runeson & Höst, 2009). In qualitative research, reliability consists of two layers, reliability of the results and reliability of the research process and is seen as parallel to the replicability of the research (Coghlan & Brydon-Miller, 2014). To provide reliability, one solution is to provide transparency in the form of documentation and reporting of the research process so that the data collection and interpretation can be evaluated and, if needed, repeated. In this way, another researcher could hypothetically conduct the same study and the result would be similar (Runeson & Höst, 2009).

The multiple methods used in this dissertation (see TABLE 5) presents various threats to reliability. The systematic mapping procedure followed in Article I is by nature easily repeatable. The steps taken and search strings used were well documented, and the data-gathering focused on databases rather than human subjects. Nevertheless, some variation as to how the search algorithms acted in the academic database were observed.

For the chosen research approach for Articles II-V, whereby the research subjects were learning and adapting humans, different challenges for reliability were raised. For example, in the case of the interviews, learning and behavioral changes were noted among those who had been interviewed after they had been exposed to set of questions on the novel subject of AI ethics. In the case of action research for Article V, a particular challenge of replicability was presented, as the aim of the study was to influence and change the research target (organizations developing software). Therefore, it is not possible to carry out subsequent studies using that study in the same context.

In the case of articles relying on data collection from interviews and questionnaires (Articles II-V), uncertainties were raised from the interview and questionnaire responses. The answers may have varied depending on internal reasons, such as willingness, readiness, understanding, or mood of the participants, and external, contextual reasons such as the presentation of the interview or questionnaire, the timing, or the language and concepts used.

To mitigate such reliability threats, the following measures were enacted. (a) An open science policy was used as widely as possible, such as making used questionnaires and some parts of the data sets openly available. (b) Separate plans for data collection and analysis were made for each article (see data collection types, TABLE 4). (c) Emphasis was placed on the researchers' distinct role and distance from the participants, so that while collecting the data, the researchers focused on maintaining their role and avoiding advising the participants or leading them in any direction. (d) Analysis results were cross-validated between two or three researchers to limit researcher error and bias. Additionally, the participant-related uncertainties may have been reduced through larger sample sizes and a comparison between respondents and cross-referencing between cases, as was done in the multiple case study of Article III. For the multi-stage process of action research, an audit trail was used, as recommended by Coghlan & Brydon-Miller (2014). In the case of Article V, an

audit trail of past publications (Vakkuri et al., 2020; Vakkuri et al., 2021; Vakkuri & Kemell, 2019) was used.

5.2.2 Construct Validity

The extent to which the selected concepts and studied measures represented what the researchers were aiming for and had invested in through the research questions is described by construct validity (Runeson & Höst, 2009). This study had three primary threats for construct validity: (a) the research strategy, (b) the construct of context (AI and ethics), and (c) the construct of method. The research approaches that were used, namely, SMS, survey, case study, and cyclical action research, are typical research approaches in the software engineering research context. The design behind the research strategies was based on a recognized approach, the industry-as-a-lab approach by Potts (1993). A later example of this approach for method development can be found, for example, in Fagerholm et al. (2017).

As highlighted in the background sections, the concept of AI and ethics can be understood differently in academic discussion and within the industry. To tackle this proposed potential threat to the validity of understanding the key concepts, two separate strategies were followed. Firstly, when addressing high-level concepts such as AI through interviews and survey, the respondents were always given room to elaborate on how they saw such concepts. Secondly, sub-concepts were used to narrow the scope and make the phenomenon more approachable. In the case of ethics, rather than directly asking about ethics itself, the questions addressed concepts grounded in existing research, such as responsibility or transparency, or even practices such as documentation. In particular, direct use of the word “ethics” or “ethical” was avoided after it was noticed that the respondents often had strong and differing preconceptions of such wording.

The concept of method was used in at least two ways in this study: method in ethical consideration and method in software development. Methods in software engineering describe ways of working and how work should be carried out. They consist of practices in the context of software engineering (Jacobson et al., 2012) and techniques in context of information systems (Tolvanen, 1998). Method in context of applied ethics is seen as a decision-making tool or support for its users to take ethical issues into account and aid in mitigating them (Pfeiffer & Forsberg, 1993). In the context of this dissertation, the possible conceptual misunderstanding was addressed by synthesis. In the research, the aim was the focus on both aspects of the method as a concept in order to provide a method fitting for the software engineering definition, but still carrying ethical consideration for software development. To guide the work, past studies addressing methods in software engineering were followed (Abrahamsson & Iivari, 2002 and Jacobson et al., 2012).

5.2.3 Internal Threats to Validity

Internal validity threats consist of threats to the examined causal relations where investigated relations of examined factors are affected by additional factors (Runeson & Höst, 2009). In other words, if additional factors are not accounted for, or their affects are not limited, there is a possibility that they could obscure the examined phenomena. The main threat to internal validity is the certainty of the relation of the investigated ethical methods, RESOLVEDD (Article IV) and ECCOLA (Article V), to the development of ethical AI systems. Therefore, there is no explicit claim that these methods produce ethical AI systems, but there is evidence that they produced ethical behavior (e.g., consideration of responsibility questions) when people were exposed to the methods. This is not only a challenge in the cases presented in this dissertation (Articles IV and V), but also on a more general level - there are no widely accepted benchmarks or measures for ethical AI systems. It is also noteworthy to mention how these methods worked. They were not standardized processes with checklists that forced developers to include precise ethical system features. Rather, the methods were mediators for the developers, and hence the actions that were taken as a result of their ethical considerations were ultimately up to the developers and organizations. In the case of both methods, it can be argued from the evidence presented in this dissertation (Articles IV and V) that the methods in question help implement AI ethics and produce more ethical consideration during development, compared to a situation where no ethical method is used.

To investigate the methods, data were collected from different contexts and using different collection methods (TABLE 4). All the data on RESEOLVEDD and most of the data on ECCOLA were collected in a real-life setting (classroom, industry), rather than from controlled experiments, and after influencing the subjects in some way. This limited the evaluation of internal validity. To exclude the additional influencing factors and raise the level of internal validity, there could have been some test-setting to include data both before and after influence by the ethical tool. Moreover, there could have been a better-defined test setting. Based on the selected research strategy, preliminary inquiries on the topic were avoided and settings as close as possible to a normal software development setting were preferred. Limiting preliminary questions beforehand was avoided, as not to direct the subjects into any line of thinking in relation to AI ethics. Having research subjects work as normally as possible while utilizing the additional ethical method was preferred in order to identify how they used the method in a real-life use-case scenario and with real stakes at play. To mitigate the internal validity threats of the selected research strategy in multiple cases, different research contexts and data collection methods from various sources were used (see Runeson & Höst, 2009). For example, Article V included more than 100 test users in six action cycles.

5.2.4 External Threats to Validity

In qualitative research, external validity concerns the extent to which the findings are generalizable and relevant outside the conducted research when there is no statistically representative sample (Runeson & Höst, 2009). The role of generalizability has even been questioned, such as in case studies under the notion that the aim is to understand a single case well in the given context (Mills et al., 2010). It has been claimed that the generalizability of a result to other cases (software companies in the case of this dissertation) can be considered only on a theoretical level (Yin, 2003). Given the qualitative nature of this dissertation, the generalizability should be looked at from the viewpoint highlighted by case study research, where only analytical generalization is possible and the results are extendable to cases that have common characteristics, and hence for which the findings are relevant (Runeson & Höst, 2009). To further the external validity and generalizability, it can be said that this dissertation provides a necessary exploratory stage for the research of novel phenomena upon which more generalizable quantitative research could be based (see Mills et al., 2010).

When assessing the external validity of empirical studies in AI ethics, the scarcity of current studies and existing gap in the field should not be overlooked. For instance, empirical studies on AI ethics methods in the area hardly exist (Section 2.2). Eisenhardt (1989) argued that for novel research areas, a low number of cases can be an acceptable number. While this notion was made for case studies in particular, the issue of generalizability is also still present in other qualitative research approaches. For external validity, it is not just the number of cases or sample size that is relevant, but also the quality. Article IV and the first cycles of action research for Article V are based on student data, which could pose a threat to validity. Using student data from the classroom provided an avenue to: (a) test a novel solution even before it is adopted in the industry; (b) test and refine an ethical method at a general level, such as the visual presentation and whether the suggested processes of the method made sense to software developers; and (c) minimize risks to companies. For example, in a student project, the shortcomings of an immature method would not result in monetary loss for the relevant parties. In empirical software engineering, the claim that using students would resolve automatically to low external validity has been challenged, and the use of students is seen as a valid simplification of reality to advance software engineering theories (Falessi et al., 2018).

5.3 Contributions

This research contributes to the formulation of the field of AI ethics and represents a paradigm change from principles-based AI ethics. This paradigm change applies to the implementation of AI ethics by focusing on developers and organizations' viewpoint towards AI ethics. This empirically based research provides theoretical contributions to the field of AI ethics by highlighting

challenges in this emerging field, and by theoretically formulating what is needed for ethical principles to transform into actionable methods. Additionally, this research provides a practical contribution to AI ethics implementation in the form of a method, ECCOLA, supported within the current state of practice in the industry.

5.3.1 Theoretical contributions

The theoretical foundation of this study is based on two separate fields of research, AI ethics and applied ethics, as described in Sections 2.2 and 2.3. These form the contextual and mechanical levels of the research, and each has its own contribution. The field of AI ethics has been gaining increasing popularity among scholars from various backgrounds (Borenstein et al., 2021). Despite this, or because of it, structured discussion and shared concepts are needed. Sub-research question 1 provided tools for this in the form of a systematic mapping study, revealing that the field is still at an immature state (Article I). On a more detailed level, the findings of Article I provide classifications to aid the formulation of other literature reviews and mapping studies (e.g., Harris & Anthis, 2021; Xie et al., 2021).

Regarding sub-research questions 2 and 3, most of the work in the field of AI ethics has been theory-oriented and without empirical validation. This is a shortcoming that this dissertation seeks to rectify. An investigation into the software industry's understanding of AI ethics issues revealed that high-level discussion of AI ethics in its current form does not transfer into practice (Articles II and III). Comparing AI ethics discussion with empirical industrial data has made it possible to empirically validate the existing literature, find areas where practice contradicts existing understanding in the literature, and provide new knowledge to the discussion. For example, in terms of contradictory practices, AI systems developers did not find it meaningful to provide explanations of their systems' inner workings to end-users (Article II), although the guidelines emphasize the importance of transparency and explainability (HLEG, 2019; IEEE Global Initiative, 2019; Jobin et al., 2019).

Hallamaa & Kalliokoski (2022) describe in their analysis of AI ethics from the viewpoint of applied ethics that AI ethics is seen as ideal of applying moral principles to practical problems. Additionally, they note that AI ethics as applied ethics should be enriched with methodological and practice-oriented approaches (Hallamaa & Kalliokoski, 2022). The change from simply providing ethical principles for AI to demanding proven methods for practice has been present for some years (Mittelstadt, 2019). Despite the discussion and recognized issues, change has been slow and proven practice-oriented approaches are scarce (Morley et al., 2021). The industry studies (Articles II and III) contribute to the recognized issues of guidelines and principles by providing empirical evidence to support the claims. For example, the case studies for Article II were noted for their empirical industry insights in the systematic review of the field by Morley et al. (2021). The method presented in this work, ECCOLA (Article V), and in the founding work (Article IV) contributes to this change by: (a) theorizing the

requirements of what is needed for a method to be implemented and used; and, (b) providing empirical evidence to support the theory. Furthermore, the method ECCOLA has inspired other researchers develop and validate practice-oriented methods for implementing AI ethics to software development (Siqueira de Cerqueira et al. 2022).

Contribution to applied ethics comes from the focus on ethical considerations for furthering methods under sub-research question 3. The empirical evidence supports Pfeiffer and Forsberg's (1993) claim that empowering ethical method users (in the case of this research, developers, and development teams) to address ethical considerations in their work is a valid way to implement ethics in practice. Barry and Ohland (2009), in their literature review on the best methods for teaching applied ethics to engineering, health, business, and law professionals, noted case-based and context-relevant examples as the preferred method. This dissertation provides empirical validation for this claim through the findings of Articles IV and V. To further enhance the acceptance of an ethical method, it should be tailored to suit the given context (Article IV and V).

5.3.2 Practical Contributions

The research approach of this dissertation, empirical research in an industrial context, has provided a platform not just for developing the theorizing of the topic but also for testing the relevance of the research in industry and gaining practitioner feedback during the process. The strongest example of this is the action research process used in Article V. The findings contribute to practice through the potential implications for a versatile group of stakeholders, from practitioners who develop AI systems to policymakers who draft regulations for these systems. The views on how the software industry currently implements ethics, provided by Articles II and III, can aid companies to benchmark their own progress on the topic and recognize working practices. For instance, by designing AI guidelines and principles, the current state of the industry can help to recognize challenges in the formulation and presentation of these alongside the effectiveness of such tools.

The main practical contribution of this work is the ECCOLA method, which is seen as a solution for supporting the industry in addressing its AI ethics concerns (Article V). During the development process, four relevant stakeholder groups were recognized. The first group is developers and designers of AI and software systems. For them, the ethical method was able to work together with existing software engineering methods, lowering the barrier to its adoption as part of the existing way of working. The second group is product owners to whom the vocabulary and presentation of the method would assist in communicating the ethical aspects of their software systems to customers and their development team. For this group, the tool can also be seen as a means to include ethical requirements in the development backlog. The third group is companies buying AI systems. Sharing the vocabulary of ethical issues and key aspects of AI ethics principles could aid customers who were not technically

oriented to demand consideration of ethical issues from the companies developing their systems, so that the ethical requirements would be included in the bid requests. The fourth group is consultants. Despite the method not aiming to encompass all aspects of AI ethics discussion, it provides an evaluation framework for understanding how a system can meet the expectations of the most common AI ethics principles. Overall, the challenge for the industry is a lack of awareness and tools for addressing ethical issues (Articles II and III). To this end, it can be claimed that the most important practical contribution of ECCOLA is its capacity to raise awareness of ethical issues (Article V).

5.3.3 Limitations

There are several limitations regarding the findings and contribution of this dissertation. The first is the selected research framework and understanding of AI ethics. Section 2.2 describes the abundance of different guidelines and sets of principles for building ethical AI systems. The AI ethics principles used in this dissertation are primarily defined by the IEEE's Ethically Aligned Design guidelines (IEEE Global Initiative, 2019) and the EU's Trustworthy AI guidelines (HLEG, 2019). Much of the existing research on various principles have been distilled into these guidelines, and they are widely known. For example, the ECCOLA method (Article V) does not cover all the principles discussed in the field, nor does it aim to. The method was designed to be a responsive and adaptive tool that can be modified to fit use-case needs and spark further ones while including new cards based on new principles.

Second, the articles are based on qualitative empirical research with limited research samples and versatile real-life settings. This limits the empirical evidence of this dissertation. The majority of the companies involved in this research were Finnish or Finnish branches of international firms. In contrast, the survey (Article III) included companies from more than 20 countries and the action research (Article V) included cycles with feedback from international conferences.

The third limitation concerns the data-gathering tools. Given the novelty of the topic and lack of empirical research, there was only a limited opportunity to adapt the wording for the questionnaires and interview instruments from prior studies. These instruments evolved with the research. Likewise, the medium of the data-gathering tool itself can be seen as a limiting factor. For example, in the survey (Article III), online questionnaires were used in the research process, which made it difficult to ensure the quality and applicability of the responses, compared to visiting the case companies.

Other limiting factors relating to the quality of this dissertation and actions taken to mitigate these factors are discussed in the articles and in Section 5.2. Despite these limiting factors, this dissertation provides answers to its research questions and complements the field of study with empirically validated theoretical knowledge and industry-relevant practical insights.

5.4 Further Studies

The dissertation and its findings open up new avenues for future research in the field of AI ethics and the application of ethical considerations as part of software engineering. The findings of this dissertation present a means to address recognized issues in the practical implementation of AI ethics, such as actionable tools and methods (Canca, 2021; Morley et al., 2021). The next step on the path to the successful implementation of ethics would be to gain an understanding of the cost of ethical consideration in AI systems development. Cost can be described as a comprehensive concept including straightforward cost from resource allocation to the cost of the psychological threshold that developers need to overcome to include ethical considerations into the development process.

The main focus of this dissertation is on the developer level and that of the development team. Further studies should expand this view to the whole organization to understand how developing organizations can increase their maturity by dealing with ethical aspects of software and AI systems. One possible approach could be a maturity model for AI ethics. Also, existing processes from the product management could be used to extend the ethical consideration further in the organization. Preliminary examples of this can be seen in Siqueira de Cerqueira's et al. (2022) work on ethical requirements elicitation and Halme's et al. (2021) work on ethical user stories.

The method presented in this dissertation used the approach of presenting ethical methods in the format of physical cards. This was borrowed from the tradition of software engineering practices where the use of physical tools is encouraged in general, as seen in the adoption of methods such as Planning Poker in Agile. From an applied ethics point of view, it would be useful to understand and empirically validate if the use of physical cards is only an efficient approach in the software engineering practices context, with its tradition of card-based methods, or if the format would be viable in other professional contexts.

YHTEENVETO (SUMMARY IN FINNISH)

Miten huomioida tekoälyn etiikka osana ohjelmistokehitystä?

Tekoälyratkaisujen määrä ja niiden hyödyntäminen osana ohjelmistoratkaisuja ovat kasvaneet kiihtyvään tahtiin viime vuosina. Tekoälyn hyödyntäminen on levinnyt elämän eri osa-alueille akateemisista ja teollisista testeistä arkipäivän sovelluksiin. Samalla näiden järjestelmien vaikutus on kasvanut niin yksittäisten ihmisten elämässä kuin koko yhteiskunnassa. Tekoälyn laajan hyödyntämisen rinnalla myös useat tekoälyjärjestelmien epäonnistumiset ja niistä johtuneet onnettomuudet ovat tulleet ihmisten tietoisuuteen lukuisten uutisotsikoiden kautta. Osaltaan nämä tapaukset ovat akateemisen tiedeyhteisön näkökulmien rinnalla korostaneet niitä eettisiä ongelmia, joita tekoälyn hyödyntämiseen liittyy. Esimerkiksi erilaiset suosittelu- sekä kuvantunnistusjärjestelmissä piilevät vääristymät (bias) ovat herättäneet huolta. Samalla onnettomuudet, jotka liittyvät tekoälyjärjestelmien käyttöön, ovat toimineet muistutuksena siitä, että nämä järjestelmät ovat vielä kaukana ideaaleista. Näiden järjestelmien kehittyessä yhä kompleksemmiksi eettisten haasteiden voidaan vain olettaa kasvavan. Eettisten haasteiden huomiointia voidaan pitää jo kiinteänä osana tekoälyjärjestelmien hyödyntämistä, jonka tulisi näkyä myös osana näiden järjestelmien kehitystä sekä käyttöönottoa.

Vastauksena tekoälyn asettamiin haasteisiin niin akateeminen kuin yritysten tekemä työ tekoälyn etiikan (AI ethics) parissa on johtanut erinäisiin kokoelmiin tekoälyn etiikan keskeisiä periaatteita. Näihin periaatteisiin kuuluu itseisarvoina pidettäviä asioita, kuten ihmisten hyvinvointi; välinearvoja, kuten vaatimus järjestelmien selitettävyyteen sekä näitä arvoja heijastelevia käytänteitä. Ongelma erilaisten tekoälyn periaatteiden kanssa on ollut niiden muuntaminen käytännön toimintaohjeiksi tai praktiikoiksi ja siten todellinen vaikuttaminen ohjelmistokehitykseen. Periaatteet ovat pitkälti näkyneet erilaisten ohjeiden, lakien sekä säännösten muodossa. Erityisesti erilaisten tekoälyn etiikan ohjeistusten (guidelines) laatimisesta on muodostunut tunnettu menettelytapa ja niitä ovat laatineet niin yritykset, valtiot kuin kansainväliset järjestötkin, kuten IEEE. Suosiosta huolimatta erilaisista periaatteista koottujen ohjeistusten on todettu olevan käytännön ohjelmistokehityksen kannalta haastavia tai jopa käyttökelvottomia. Vaikka yksittäisiä spesifejä työkaluja yksittäisten eettisten kysymysten avuksi on kehitetty, ei niistä juuri ole apua järjestelmien suunnittelu- ja kehitystyössä kokonaisuudesta käsin katsottuna. Tämän väitöskirjan tavoitteena on vastata tekoälyn hyödyntämisen eettisten periaatteiden esille nostamien teoreettisten kysymysten ja ohjelmistokehityksen käytännön yhteensovittamiseen liittyviin haasteisiin. Päättökysymyksenä on: Miten huomioida tekoälyn etiikka osana ohjelmistokehitystä? Väitöskirjassa esitellään katsaus tekoälyn eettisten kysymysten huomioimiseen teollisuudessa sekä väitöskirjaprosessin aikana kehitetty monien testivaiheiden läpi käynyt menetelmä, ECCOLA, tekoälyn etiikan haasteiden huomioimiseen ohjelmistokehityksessä. Tutkimuksessa on keskitytty erityisesti ohjelmistokehityksen käytännön näkökulmaan antamalla ääni myös

vähemmän esillä olleille sovellusten kehittäjille ja heidän haasteilleen. Tutkimusaineiston kokoamisessa keskeisessä roolissa ovat olleet ohjelmistokehittäjien ja ohjelmistokehitystiimien näkökulmat. Tutkimusaineisto koostuu laadullisesta ja empiirisestä aineistosta, joka on raportoitu viidessä tieteellisessä artikkelissa.

Tutkimus on ottanut vahvasti vaikutteita empiirisestä ohjelmistokehitystutkimuksesta (empirical software engineering), jonka tavoitteena ei ole pelkästään havaita käytännön kautta ohjelmistokehityksen ilmiöitä, vaan myös empirian keinoin todentaa kirjallisuudessa esitettyjä hypoteeseja sekä etsiä niihin vastauksia empiirisesti testaamalla. Yksittäisistä tutkimusmenetelmistä keskeisin on ollut toimintatutkimus (cyclical action research), jonka tutkimussyklit kokoavat sisälleen tässä väitöskirjassa esiteltyjen tutkimusartikkelien havaintoja sekä tuloksia. Kootun laadullisen tutkimusaineiston analyysissä on hyödynnetty pääosin sisällönanalyysin menetelmiä, kuten ankkuroitua teoriaa (Grounded Theory). Tässä väitöskirjassa esiteltyjen empiiristen aineistojen ja aiemman tutkimuksen perustella voidaan todeta, että käsitys tekoälyn etiikasta – siitä, mitä se on ja miten sitä tulisi lähestyä - on elänyt viime vuosina niin teollisuudessa kuin akateemisessa keskustelussa. Tietoisuus teollisuudessa eettisten kysymysten huomioimisen merkittäväydestä on kasvanut, mutta menetelmiä näiden kysymysten huomiointiin ei juuri ole käytössä. Käytännössä asian huomioiminen on vielä vähäistä, vaikka kiinnostusta aiheeseen on. Väitöskirjan tutkimustulokset osoittavat, että tekoälyn eettisten kysymysten onnistunut huomioiminen edellyttää sisällöltään alaspesifejä menetelmiä tukemaan kehittäjien tietoisuutta sekä valmiuksia eettisten kysymysten käsittelyyn. Toiminnaltaan tällaisten menetelmien tulee olla kevyitä, ketteriä ja luontevasti ohjelmistokehityksen osaksi soveltuvia, jotka mukautuvat erilaisten kehitysmenetelmien sekä tiimien tarpeisiin. Nämä näkökulmat on pyritty yhdistämään tässä väitöskirjassa esitellyssä ECCOLA-menetelmässä. Tutkimuksen perusteella suositellaan tekoälyn eettisiä haasteita ratkaistaessa huomioimaan olemassa olevat ohjelmistokehityksen menetelmät sekä prosessit. Tällä tavalla voidaan edistää periaatteiden toteutumista käytännön tasolla. Etiikka ei voi tulla huomioiduksi tyhjiössä erillään muusta kehitystyöstä, vaan sen tulee nivoutua osaksi sitä.

REFERENCES

- Abrahamsson, P., & Iivari, N. (2002). Commitment in software process improvement-in search of the process. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 3239–3248). IEEE.
- Allhoff, F. (2011). What are applied ethics? *Science and Engineering Ethics*, 17(1), 1–19.
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989.
- Archard, D., & Lippert-Rasmussen, K. (2013). Applied ethics. *The international Encyclopedia of Ethics*, 10, 9781444367072.
- Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*, 1–25.
- ALTAI, (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Europa. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- Basili, V. R. (2013). A personal perspective on the evolution of empirical software engineering. In *Perspectives on the Future of Software Engineering* (pp. 255–273). Springer, Berlin, Heidelberg.
- Barry, B. E., & Ohland, M. W. (2009). Applied ethics in the engineering, health, business, and law professions: A comparison. *Journal of Engineering Education*, 98(4), 377–388.
- Boddington, P. (2017). Codes of professional ethics. In *Towards a code of ethics for artificial intelligence* (pp. 39–57). Springer.
- Borenstein, J., Grodzinsky, F. S., Howard, A., Miller, K. W., & Wolf, M. J. (2021). AI ethics: A long history and a recent burst of attention. *Computer*, 54(1), 96–102.
- Bringsjord, S. & Govindarajulu, N. S. (2020). Artificial intelligence. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* (Summer 2020 ed.) <https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence>
- Bynum, T. W. (2006). Flourishing ethics. *Ethics and Information Technology*, 8(4), 157–173.
- Canca, C. (2020). Operationalizing AI ethics principles. *Communications of the ACM*, 63(12), 18–21.
- Coghlan, D., & Brydon-Miller, M. (Eds.) (2014). *The SAGE encyclopedia of action research*. (Vols. 1–2). SAGE Publications.
- Cruzes, D. S., & Dybå, T. (2011). Recommended steps for thematic synthesis in software engineering. In *2011 International Symposium on Empirical Software Engineering and Measurement* (pp. 275–284). IEEE.
- Davison, R., Martinsons, M. G., & Kock, N. (2004). Principles of canonical action research. *Information Systems Journal*, 14(1), 65–86.

- Dignum, V. (2017). Responsible artificial intelligence: designing AI for human values.
- Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, 20(1), 1–3.
- Ebert, C., Abrahamsson, P., & Oza, N. (2012). Lean software development. *IEEE Software*, 29(5), 22–25.
- Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, 14(4), 532–550.
- Fagerholm, F., Guinea, A. S., Mäenpää, H., & Münch, J. (2017). The RIGHT model for continuous experimentation. *Journal of Systems and Software*, 123, 292–305.
- Falessi, D., Juristo, N., Wohlin, C., Turhan, B., Münch, J., Jedlitschka, A., & Oivo, M. (2018). Empirical software engineering experts on the use of students and professionals in experiments. *Empirical Software Engineering*, 23(1), 452–489.
- Fernández, D. M., & Passoth, J. H. (2019). Empirical software engineering: From discipline to interdiscipline. *Journal of Systems and Software*, 148, 170–179.
- Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. *Australian Critical Care*, 22(2), 93–97.
- Floridi, L. (2006). Four challenges for a theory of informational privacy. *Ethics and Information technology*, 8(3), 109–119.
- Flynn, J. (2021). Theory and bioethics. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*.
<https://plato.stanford.edu/archives/spr2021/entries/theory-bioethics>
- Friedman, B., Kahn, P. H., Borning, A. (2008). Value sensitive design and information systems. In *The Handbook of Information and Computer Ethics* (pp. 69–101).
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120.
- Halme, E., Vakkuri, V., Kultanen, J., Jantunen, M., Kemell, K. K., Rousi, R., & Abrahamsson, P. (2021). How to write Ethical user stories? Impacts of the ECCOLA Method. In *International Conference on Agile Software Development* (pp. 36–52). Springer, Cham.
- Hallamaa, J., & Kalliokoski, T. (2022). AI Ethics as Applied Ethics. *Frontiers in Computer Science*, 4, 12.
- Harris, J., & Anthis, J. R. (2021). The moral consideration of artificial entities: A literature review. *Science and Engineering Ethics*, 27(4), 1–95.
- HLEG, High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. *Europa*. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- Holton, J. & Walsh, I. (2017). 5 finding your data. In *Classic grounded theory: Applications with qualitative and quantitative data* (pp. 57–75)
- IEEE Global Initiative. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (1st ed.)

<https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>

- Introna, L. D. (1997). Privacy and the computer: why we need privacy in the information society. *Metaphilosophy*, 28(3), 259-275.
- Jacobson, I., Ng, P. W., McMahon, P. E., Spence, I., & Lidman, S. (2012). The essence of software engineering: The SEMAT kernel. *Communications of the ACM*, 55(12), 42-49.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399.
- Johnson, B., & Smith, J. (2021). Towards ethical data-driven software: Filling the gaps in ethics research & practice. In *2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics)* (pp. 18-25). IEEE.
- Joye, D., Wolf, C., Smith, T. W., & Fu, Y. C. (2016). Survey methodology: Challenges and principles. *The SAGE handbook of survey methodology* (pp. 3-15).
- Kitchenham B. and Charters S. (2007). Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University Joint Report.
- Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004). Evidence-based software engineering. In *Proceedings. 26th International Conference on Software Engineering* (pp. 273-281). IEEE.
- Kitchenham, B. A., & Pfleeger, S. L. (2008). Personal opinion surveys. In *Guide to advanced empirical software engineering* (pp. 63-92). Springer.
- McCarthy, J. (2007). What is artificial intelligence?
<http://jmc.stanford.edu/articles/whatisai/whatisai.pdf>
- McNamara, A., Smith, J., & 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* (pp. 729-733).
- Mikkonen, T., Nurminen, J. K., Raatikainen, M., Fronza, I., Mäkitalo, N., & Männistö, T. (2021). Is machine learning software just software: A maintainability view. In *International Conference on Software Quality* (pp. 94-105). Springer.
- Mills, A. J., Durepos, G., & Wiebe, E. (Eds.) (2010). *Encyclopedia of case study research*. SAGE Publications.
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501-507.
- Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, 16(4), 266-275.
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, governance, and policies in artificial intelligence* (pp. 153-183). Springer.

- Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI—Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1), 34–47.
- Petersen, K., & Gencel, C. (2013). Worldviews, research methods, and their relationship to validity in empirical software engineering research. In *2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement* (pp. 81–89). IEEE.
- Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64, 1–18.
- Pfeiffer, R. S., & Forsberg, R. P. (1993). *Ethics on the job: Cases and strategies*. Cengage Learning.
- Potts, C. (1993). Software-engineering research revisited. *IEEE Software*, 10(5), 19–28.
- Rombach, D. (2013). Empirical software engineering models: Can they become the equivalent of physical laws in traditional engineering? In *Perspectives on the Future of Software Engineering* (pp. 1–12). Springer.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2), 131–164.
- Russell, S., & Norvig, P. (2020). *Artificial intelligence: a modern approach*. Pearson
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28.
- Seaman, C. B. (1999). Qualitative methods in empirical studies of software engineering. *IEEE Transactions on Software Engineering*, 25(4), 557–572.
- Siqueira de Cerqueira, J. A., de Azevedo, A.P., Tives, H.A., Canedo E.D. (2022) Guide for Artificial Intelligence Ethical Requirements Elicitation-RE4AI Ethical Guide. In *Proceedings of the 55th Annual Hawaii International Conference on System Sciences*.
- Stol, K. J., Ralph, P., & Fitzgerald, B. (2016). Grounded theory in software engineering research: A critical review and guidelines. In *Proceedings of the 38th International Conference on Software Engineering* (pp. 120–131).
- Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... & Teller, A. (2016). *Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence*. Stanford University
- Susman, G. I., & Evered, R. D. (1978). An assessment of the scientific merits of action research. *Administrative Science Quarterly*, 582–603.
- Tännsjö, T. (2011). Applied ethics. A defence. *Ethical Theory and Moral Practice*, 14(4), 397–406. <https://doi.org/10.1007/s10677-011-9293-8>

- Tolvanen, J. P. (1998). Incremental method engineering with modeling tools: Theoretical principles and empirical evidence. Ph.D Thesis, University of Jyvaskyla.
- Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, 11(2), 105–112.
- Vakkuri, V., & Kemell, K. K. (2019). Implementing artificial intelligence ethics: A tutorial. In *International Conference on Software Business* (pp. 439–442). Springer.
- Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2020). ECCOLA-a method for implementing ethically aligned AI systems. In *46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 195–204). IEEE.
- Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2021). Technical briefing: Hands-on session on the development of trustworthy AI software. In *IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)* (pp. 332–333). IEEE.
- Van den Hoven, J. (2008). Moral methodology and information technology. *The Handbook of Information and Computer Ethics*, 49.
- Xie, Y., Cruz, L., Heck, P., & Rellermeier, J. S. (2021). Systematic mapping study on the machine learning lifecycle. In *IEEE/ACM 1st Workshop on AI Engineering–Software Engineering for AI (WAIN)* (pp. 70–73). IEEE.
- Yin, R. K. (2003). *Case study research: Design and methods* (Vol. 5). SAGE.
- Yin, R. K. (2015). *Qualitative research from start to finish*. Guilford publications.
- Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., ... & Perrault, R. (2021). The AI index 2021 annual report. arXiv. <https://doi.org/10.48550/arXiv.2103.06312>

APPENDIX: ECCOLA CARDS

This appendix showcases the card presentation of ECCOLA method presented in Article V. ECCOLA is intended to provide developers an actionable tool for implementing AI ethics. ECCOLA method is meant to be evolving tool that is iteratively updated based on empirical research and the latest version of the can be accessed from external repository:

<https://doi.org/10.6084/m9.figshare.12136308>

Analyze

#0 Stakeholder Analysis


Motivation: In order to understand the big picture, it is important to first understand who the system can affect and how. Try to also think past the obvious, direct stakeholders such as your end-users.

What to Do: Identify stakeholders.

- Who does the system affect, and how? Stakeholders are not simply users, developers and customers.
- How are the various stakeholders linked together?
- Can these different stakeholders influence the development of the system? How?
- Remember that a user is often an organization and the end-user is an individual. Similarly, AI systems can treat people as objects for data collection.

Practical Example: Autonomous cars don't just affect their passengers. Anyone nearby is affected; some even change the way they drive. If at one point half of the traffic consists of self-driving cars, what are the societal impacts of such systems? E.g., regulations arising from such systems also affect everyone.

cID 2102-20201007



Transparency

#1 Types of Transparency

Motivation: When considering transparency, it is important to understand who you are being transparent towards, and what you are being transparent about.

What to Do: Consider the following...

- Are you trying to understand something? (Internal transparency)
- Are you trying to explain something? (External transparency)
- Are you trying to understand or explain how the system works? (Transparency of algorithms and data)
- Are you trying to understand or explain why the system was made to be the way it is now? (Transparency of system development)
- External stakeholders to consider, among others: (end-)users, safety certification agencies, accident investigators, lawyers or expert witnesses, and society at large for disruptive technologies

cID 3102-20201007




FIGURE 8 ECCOLA cards 0 and 1

Trans-
parency


#2 Explainability

Motivation: If we cannot understand the reasons behind the actions of the AI, it is difficult to trust it.

What to Do: Ask yourself:

- Is explainability a goal for your system? How do you plan to ensure it?
- How well can each decision of the system be understood? By both developers and (end-)users.
- Did you try to use the simplest and most interpretable model possible for the context?
- Did you make trade-offs between explainability and accuracy? What kind of? Why?
- How familiar are you with your training or testing data? Can you change it when needed?
- If you utilize third party components in the system, how well do you understand them?

Practical Example: When interacting with a robot, users could ideally ask the robot “why did you do that?” and receive an understandable response. This would make it much easier for them to trust a system.

 ECCOLA

cID 3202-20201007

Trans-
parency


#3 Communication

Motivation: In practice, communication is a big part of being transparent with your stakeholders. Being transparent in communication can generate trust.

What to Do: Ask yourself:

- What is the goal of the system? Why is this particular system deployed in this specific area?
- What do you communicate about the system to its users and end-users? Is it enough for them to understand how the system works?
- If relevant to your system, do you somehow tell your (end-)users that they are interacting with an AI system and not with another human being?
- Do you collect user feedback? How is it used to change/improve the system?
- Are communication and transparency towards other audiences, such as the general public, relevant?

Practical Example: Clearly stating what data you collect and why can make you more trustworthy. Compare this to a cellphone application that just states it needs to access your camera and storage.

 ECCOLA

cID 3302-20201007

FIGURE 9 ECCOLA cards 2 and 3

Trans-
parency


#4 Documenting Trade-offs

Motivation: One important part of transparent system development is the documentation of trade-offs. Whenever you make a decision, you choose one option over other alternatives. However, documenting *why* and *what* the alternatives were is important.

What to Do: Ask yourself:

- i. Are relevant interests and values implicated by the system and potential trade-offs between them identified and documented?
- ii. Who decides on such trade-offs (e.g. between two competing solutions) and how? Did you ensure that the trade-off decision and the reasons behind it were documented?

Practical Example: E.g., choosing machine learning algorithm is often a trade-off between accuracy and explainability. Documenting trade-offs can improve your customer relationship, allowing you to better explain why certain decisions were made over others. Moreover, it can reduce the responsibility placed on the individual developer(s) from an ethical point of view.

 ECCOLA

cID 3402-20201007

Trans-
parency


#5 Traceability

Motivation: Traceability supports explainability. It helps us understand why the AI acts the way it does.

What to Do: Document. Different types of documentation (code, project etc.) are typically key in producing transparency.

- i. How have you documented the development of the system, both in terms of code and decision-making? How was the model built or the AI trained?
- ii. How have you documented the testing and validation process? In terms of data and scenarios used etc.
- iii. How do you document the actions of the system? What about different actions in mostly similar scenarios (e.g. if the user was different but the situation otherwise the same)?

Practical Example: When the system starts making mistakes, by aiming for traceability, it will be easier to find out the cause. Consequently, it will also be faster and possibly easier to start fixing the underlying issue from an ethical point of view.

 ECCOLA

cID 3502-20201007

FIGURE 10 ECCOLA cards 4 and 5

Trans-
parency


#6 System Reliability

Motivation: Transparency makes ethical development possible in the first place. To make it ethical, we must understand how the system works and why it makes certain decisions.

What to Do: Ask yourself:

- How do you test if the system fulfills its goals?
- Have you tested the system comprehensively, including unlikely scenarios? Have the tests been documented?
- When the system fails in a certain scenario, will you be able to tell why? Can you replicate the failure?
- How do you assure the (end-)user of the system's reliability?

Practical Example: An autonomous coffee machine successfully brews coffee 8 times out of 10. While this is a decent success rate, we are left wondering what happened the 2 times it failed to do so, and why. Errors are inevitable, but we must understand the causes behind them and be able to replicate them to fix them.

 ECCOLA

cID 3601-20200415

Data


#7 Privacy and Data

Motivation: Privacy is a rising trend in the wake of various recent data misuse reveals. People are now increasingly conscious about handing out personal data. Similarly, regulations such as the General Data Protection Regulation (GDPR) now affect data handling.

What to Do: Ask yourself:

- What data are used by the system?
- Does the system use or collect personal data? Why? How is the personal data used?
- Do you clearly inform your (end-)users about any personal data collection? E.g., ask for consent, provide an opportunity to revoke it etc.
- Have you taken measures to enhance (end-user) privacy, such as encryption or anonymization?
- Who makes the decisions regarding data use and collection? Do you have organizational policies for it?

Practical Example: Rather than collecting and selling data, appealing to privacy can also be profitable. Regulations are making it increasingly difficult to collect lots of personal data for profit. Privacy can be an alternate selling point in today's climate.

 ECCOLA

cID 4102-20201007

FIGURE 11 ECCOLA cards 6 and 7


Data **#8 Data Quality**

Motivation: As AI are trained using data, the data used directly affects how the system operates. The nature, the quality, and integrity of the data used have to align with the goals of the system.

What to Do: Ask yourself:

- i. What are good or poor-quality data in the context of your system?
- ii. How do you evaluate the quality and integrity of your own data? Are there alternative ways?
- iii. If you utilize data from external sources, how do you control their quality?
- iv. Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?
- v. How can you tell if your data sets have been compromised? E.g., data pollution.
- vi. Who handles the data collection, storage, and use?

Practical Example: In 2017, Amazon scrapped its recruitment AI because of data. They used past recruitment data to teach the AI. As they had mostly hired men, the AI began to consider women undesirable based on the data.

 ECCOLA

cID 4202-20201007


Data **#9 Access to Data**

Motivation: Aside from carefully planning what data you collect and how, it is also important to plan how it can or will be used and by whom.

What to Do: Ask yourself:

- i. Who can access the users' data, and under what circumstances?
- ii. How do you ensure that the people who access the data: 1) have a valid reason to do so; and 2) adhere to the regulations and policies related to the data?
- iii. Do you keep logs of who accesses the data and when? Do the logs also tell why?
- iv. Do you use existing data governance frameworks or protocols? Does your organization have its own?

Practical Example: Third parties you give access to the data can misuse it. A prominent example of this is the case of Cambridge Analytica and Facebook, in which data from Facebook was used questionably. However, such incidents can also paint your organization in a bad light even if you were not the ones misusing the data.

 ECCOLA

cID 4301-20200415

FIGURE 12 ECCOLA cards 8 and 9


Agency & Oversight #10 Human Agency

Motivation: People interacting with the system or using it should be able to understand it sufficiently. Users should be able to make informed decisions based on its suggestions, or to challenge its suggestions. AI systems should let humans make independent choices.

What to Do: Ask yourself:

- i. Does the system interact with decisions by human actors, i.e. end users (e.g. recommending users actions or decisions, or presenting options)?
- ii. Does the system communicate to its (end) users that a decision, content or outcome is the result of an algorithmic decision? Into how much detail does it go?
- iii. In the system's use context, what tasks are done by the system and what tasks are done by humans?
- iv. Have you taken measures to prevent overconfidence or overreliance on the system?

Practical Example: A medical system recommends diagnoses. How does the system communicate to doctors why it made a recommendation? How should the doctors know when to challenge the system? Does the system somehow change how patients and doctors interact?

 ECCOLA

cID 5101-20200415


Agency & Oversight #11 Human Oversight

Motivation: AI systems should support human decision-making. They should not undermine human autonomy by making decisions for us, meaning they should be subject to human oversight.

What to Do: Ask yourself:

- i. Who can control the system and how? In what situations?
- ii. What would be the appropriate level of human control for this particular system and its use cases?
- iii. Related to the Safety and Security cards: how do you detect and respond if something goes wrong? Does the system then stop entirely, partially, or would control be delegated to a human? Why?

Practical Example: Assuming control is especially related to cyber-physical systems such as drones or other vehicles. For purely digital systems, the focus should be on *supporting* human decision-making instead of directing it.

 ECCOLA

cID 5201-20200415

FIGURE 13 ECCOLA cards 10 and 11

Safety & Security


#12 System Security

Motivation: While cybersecurity is important in any system, AI systems present new challenges. Cyber-physical systems can even cause fatalities in the hands of malicious actors.

What to Do: Ask yourself:

- Did you assess potential forms of attacks to which the system could be vulnerable? Did you consider ones that are unique or more relevant to AI systems?
- Did you consider different types of vulnerabilities, such as data pollution and physical infrastructure?
- Have you verified how your system behaves in unexpected situations and environments?
- Does your organization have cybersecurity personnel? Are they involved in this system?

Practical Example: The autonomous nature of AI systems makes new vectors of attack possible. A white line drawn across a road can confuse a self-driving vehicle. The case of Microsoft's Tay Twitter bot, who began to exhibit extreme views after being bombarded with such, is one example of a new type of attack.

 ECCOLA

cID 6102-20201007

Safety & Security


#13 System Safety

Motivation: AI systems exert notable influence on the physical world whether they are cyber-physical or not. Various risks and their consequences should be considered, thinking ahead to the operational life of the system.

What to Do: Ask yourself:

- What kind of risks does the system involve? What kind of damage could it cause?
- How do you measure and assess risks and safety?
- What fallback plans does your system have? Have they been tested?
- In what conditions do the fallback plans trigger? Are they automatic or do they require human input?
- Is there a plan to mitigate or manage technological errors, accidents, or malicious misuse? What if the systems provides wrong results, becomes unavailable, or provides societally unacceptable results?
- What liability and consumer protection laws apply to

Practical Example: AI systems can aid automating various organizational tasks, making it possible to reduce personnel. However, if a customer organization becomes reliant on your AI system to handle a portion of its operations, what happens if that AI stops functioning for even a few days? What could you do to alleviate the impact?

 ECCOLA

cID 6201-20200415

FIGURE 14 ECCOLA cards 12 and 13


Fairness #14 Accessibility

Motivation: Technology can be discriminating in various ways. Given the enormous impact AI systems can have, ensuring equal access to their positive impacts is ethically important.

What to Do: Ask yourself:

- i. Does the system consider a wide range of individual preferences and abilities? If not, why?
- ii. Is the system usable by those with special needs or disabilities, those at risk of exclusion, or those using assistive technologies?
- iii. Were people representing various groups somehow involved in the development of the system?
- iv. How is the potential user audience taken into account?
- v. Is the team involved in building the system representative of your target user audience? Is it representative of the general population?
- vi. Did you assess whether there could be (groups of) people who might be disproportionately affected by the negative implications of the system?

Practical Example: AI tends to benefit those who are already technologically capable, resulting in increased inequality.

 ECCOLA

cID 7102-20201007


Fairness #15 Stakeholder Participation

Motivation: As AI systems have notable impacts, their stakeholders are also numerous. Though the system affects these various holders in various ways, they are often not involved in the development. Yet, e.g. when using a decision-making system, its users have to trust the system while also being critical of it.

What to Do: Check your stakeholder analysis (card #0):

- i. Which stakeholders are stakeholders in system development?
- ii. How are the different stakeholders of the system involved in the development of the system? If they aren't, why?
- iii. How do you inform your external and internal stakeholders of the system's development?

Practical Example: Often the people an AI system is used on are individuals who are simply objects for the system. For example, a medical system is developed for hospitals, used by doctors, but ultimately used on patients. Why not talk to the patients too?

 ECCOLA

cID 7202-20201007

FIGURE 15 ECCOLA cards 14 and 15


Wellbeing **#16 Environmental Impact**

Motivation: Past the general wellbeing implications, ecological consciousness is a current trend. Being ecological can be a selling point for your organization.

What to Do: Ask yourself:

- i. Did you assess the environmental impact of the system's development, deployment, and use? E.g., the type of energy used by the data centers.
- ii. Did you consider the environmental impact when selecting specific technical solutions?
- iii. Did you ensure measures to reduce the environmental impact of your system's life cycle?

Practical Example: If you are hosting on a third party cloud, try to ascertain the sustainability of the service provider's services. If you are using hardware, are you processing the data in each physical device of your own or are you processing it in the cloud?

 ECCOLA

cID 8101-20200415


Wellbeing **#17 Societal Effects**

Motivation: The impacts of a system go beyond its user-base. A system may affect negatively even those who do not use it nor wish to use it.

What to Do: Ask yourself:

- i. Did you assess the broader societal impact of the AI system's use beyond the individual (end-)users? Consider stakeholders who might be indirectly affected by the system.
- ii. How will the systems affect society when in use?
- iii. What kind of systemic effects could the system have?

Practical Example: Surveillance technology utilizing facial recognition AI has long-reaching impacts. People may wish to avoid areas that utilize such surveillance, negatively affecting businesses in said area. People may become stressed at the mere thought of such surveillance. Some may even emigrate as a result.

 ECCOLA

cID 8202-20201007

FIGURE 16 ECCOLA cards 16 and 17

Account-ability


#18 Auditability

Motivation: Regulations affecting AI and data may necessitate audits of systems in the future. Similarly, if the system causes damage, an audit might be requested. It is good to have mechanisms in place beforehand.

What to Do: Ask yourself:

- Is the system auditable?
- Can an audit be conducted independently?
- Is the system available for inspection?
- What mechanisms facilitate the system's auditability?
How is traceability and logging of the system's processes and outcomes ensured?

Practical Example: In heavily regulated fields such as medicine, audits are typically required before a system can be utilized in the first place.

 ECCOLA

cID 9101-20200415

Account-ability


#19 Ability to Redress

Motivation: Making sure people know they can be compensated in some way in the event something goes wrong with the system is important in generating trust. Such scenarios should be planned in advance to what extent possible.

What to Do: Ask yourself:

- What is your (developer organization) responsibility if the system causes damage or otherwise has a negative impact?
- In the event of negative impact, can the ones affected seek redress?
- How do you inform users and other third parties about opportunities for redress?

Practical Example: AI systems can inconvenience users in unforeseen, unpredictable ways. Depending on the situation, the company may or may not be legally responsible for the inconvenience. Nonetheless, by offering a digital platform for seeking redress, your company can seem more trustworthy while also offering additional value to your users.

 ECCOLA

cID 9201-20200415

FIGURE 17 ECCOLA cards 18 and 19

Account-
ability

#20 Minimizing Negative Impacts

Motivation: Minimizing negative impacts of the system is financially important for any developer organization. Incidents are often costly.

What to Do:

- i. First, consider...
 - a. Is your stakeholder analysis up-to-date (Card #0)
 - b. Have you discussed risks? (Card #13)
 - c. Have you discussed auditability?
 - d. Have you discussed redress issues?
- ii. Are the people involved with the development of the system also involved with it during its operational life? If not, they may not feel as accountable.
- iii. Are you aware of laws related to the system?
- iv. Can users of the system somehow report vulnerabilities, risks, and other issues in the system?
- v. With whom have you discussed accountability and other ethical issues related to the system, including grey areas?



cID 9302-20201007

FIGURE 18 ECCOLA card 20



ORIGINAL PAPERS

I

THE KEY CONCEPTS OF ETHICS OF ARTIFICIAL INTELLIGENCE

Ville Vakkuri & Pekka Abrahamsson 2018

IEEE International Conference on Engineering, Technology and
Innovation (ICE/ITMC)

DOI 10.1109/ICE.2018.8436265

Reproduced with kind permission by IEEE.

The Key Concepts of Ethics of Artificial Intelligence

A Keyword based Systematic Mapping Study

Ville Vakkuri

[<https://orcid.org/0000-0002-1550-1110>]

Faculty of Information Technology

University of Jyväskylä

Jyväskylä, Finland

ville.vakkuri@jyu.fi

Pekka Abrahamsson

[<https://orcid.org/0000-0002-4360-2226>]

Faculty of Information Technology

University of Jyväskylä

Jyväskylä, Finland

pekka.abrahamsson@jyu.fi

Keywords — Artificial Intelligence; Ethics; AI ethics; Systematic Mapping Study

Abstract — The growing influence and decision-making capacities of Autonomous systems and Artificial Intelligence in our lives force us to consider the values embedded in these systems. But how ethics should be implemented into these systems? In this study, the solution is seen on philosophical conceptualization as a framework to form practical implementation model for ethics of AI. To take the first steps on conceptualization main concepts used on the field needs to be identified. A keyword based Systematic Mapping Study (SMS) on the keywords used in AI and ethics was conducted to help in identifying, defying and comparing main concepts used in current AI ethics discourse. Out of 1062 papers retrieved SMS discovered 37 re-occurring keywords in 83 academic papers. We suggest that the focus on finding keywords is the first step in guiding and providing direction for future research in the AI ethics field.

1. INTRODUCTION

By reviewing the latest accomplishment and increasing implementation of Autonomous systems (AS) and Artificial Intelligence (AI) systems have become more influential in our lives. By growing influence ethical questions related to these systems have become more and more obvious and actual.

For example, looking biased algorithms in social media[1], decision making systems of autonomous cars[2], or even social effects of automatization in whole transportation ecosystems like autonomous maritime[3] it is clear that system development is not anymore only about technological or engineering question. AI and AS are already in the surrounding world among us and the need of implementing ethics and our values into these systems is urgent.

Concerning ethics as a part of system design has also gained attention from governmental and standardization level, such as Federal Ministry of Transport and Digital Infrastructure in Germany[4] and IEEE[5]. The academic discussion on the relation of AI and ethics has been ongoing for decades, but the development of systems and ethical research have only slightly crossed[6]. The ethical research has been mainly focused on the potential of AI on theoretical level [7]. So, the question still remains open on application level: How ethics should be implemented in practice into these systems?

There can be little ethical implementation without understanding the consequences of developers' own actions, open dialogue and ethical aspects considered in AI and

autonomous system development, because of the multidisciplinary nature of AI ethics development [8]. As a solution for understanding the field of ethics of AI, philosophical conceptualization should be used. This method allows to discuss and to form cross-disciplinary definitions for key concepts and also initiate productive dialog merging philosophical and technological views to produce a common framework for implementing ethics in AI.

The goal of this paper is to identify and categorize keywords used in academic papers in the current AI ethics discourse and by that take first steps to identify, define and compare main concepts and terms used in discourse. To find the relevant papers and keywords, a preliminary Systematic Mapping Study was conducted with the following focus:

- Recognize keywords used in the field
- Extract potential keywords for future research
- Compare keywords to proposed concepts in academic literature

The Systematic Mapping Study based on keywords reveals 37 re-occurring author keywords found in 83 academic papers that are found from an initial set of 1062 papers in the field of AI and ethics. Cause of the preliminary nature of this study as Systematic Mapping Study it does not provide full comprehensive picture of the primary studies in the area, but it provides an important standpoint and relevant tools for future research on AI and ethics. By understanding the used concepts, research can shift from discussing concepts to defining

This is the author's version of the work. The definite version was published in Vakkuri, V., Abrahamsson, P. 2018, June. The Key Concepts of Ethics of Artificial Intelligence. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC). IEEE. <http://dx.doi.org/10.1109/ICE.2018.8436265>

concepts and this way aids the need of practical implementation of ethics into AI systems.

This paper is organized as follows: Section 2 describes the background and related work; Section 3 describes the research methodology and conducted keyword based search; Section 4 findings; Section 5 concludes the paper by discussing the findings with general presentation of AI ethics and summarizes the answers for research questions to set guidelines for future work.

2. BACKGROUND

The ethical discussion of Artificial Intelligence has been present from the start for AI research, but instead of focusing on the real use cases, the focus has been mainly on the theoretical work discussing the possibilities and future impacts of AI. In recent years there has been a major change in discussion of AI related ethics when new level of capabilities of AI have become reality and more influential in our lives due to the recent breakthroughs in AI development. Availability of low cost computing power and innovation like Big Data technologies have made AI more useable in solving complicated problems. [9] One milestone of AI development can be seen in year 2012 when Google's large-scale deep learning experimentation on brain simulation using 16000 CPU cores and deep learning was conducted[10]. The experiment significantly improved the state of the art on a standard image classification test. This year also serves as the starting point for the current AI ethics discussion in the context of this study.

Even though the academic discussion on the relation of AI and ethics has been ongoing for decades, there is no commonly shared definition of what AI ethics is or even how it should be named. As the defining concept, Machine ethics has arisen out the discussion but it has also been criticized. There has been a heated discussion on how does the concept of machine ethics also cover and include new branches in AI related ethics. [7, 11, 12]

There is only a handful of books that have comprehensive presentations covering the ethical issues of AI, such as Towards a Code of Ethics for Artificial Intelligence that mainly focuses on professional ethics[9]; robot ethics 2.0 covers ethics related to embodied AI[13] and Machine Ethics prior to the current discussion[14]. For defining the field of AI related ethics so called "six hot topics" have proposed [15]. The problem with these categorically wide topics is that they are not necessarily comprehensive or clear enough and not in balance with the overall discussion. Importantly, they are also not necessarily scientifically founded. For example of the wide scope of AI ethics discourse, the first AAAI/ACM Conference on AI, Ethics, and Society held in 2018 had broad set of 12 different topics from technical to social sciences[16].

Besides defining relevant concepts for a crucial problem in practical implementation of AI ethics is the limited co-operation and communication between the developers of the AI systems and ethics researchers.[6, 11] To reach the practical implementation of AI ethics, a multidisciplinary research approach is needed where AI developers can also see the use of ethics and results of the philosophical research on a practical level.

3. RESEARCH METODOLOGY AND MAPPING

As a multidisciplinary research area AI ethics covers a wide range of topics and the discussion of definitions still endures. To gain a better understanding of the research area, a Systematic Mapping Study was chosen as a research method due to its capability to deal with wide and loosely defined areas of study. SMS aims at producing an overview of the field and reveals concretely which topics have been covered to a certain extent. The present study is a keyword based systematic mapping study. Two main guidelines for systematic mapping study were combined aiming at recognizing primary studies and the used keywords therein. We consider this study, however, to be the first step since the mapping process is not executed to its full length. We needed first to gain a better understanding of relevant keywords for the PICO (Population, Intervention, Comparison and Outcomes) process. [17, 18]

A. Definition of Research Questions

The main research question for the present study is: What are the main author keywords used in academic papers in the current AI ethics discussion. To answer this question, four sub-questions were formed:

- Q1 What are the author keywords used?
- Q2 Which of the keywords are re-occurring and in which pattern?
- Q3 How can the author keywords be classified?
- Q4 How do the used keywords reflect the proposed concepts in academic AI ethics literature?

The purpose of Q1 is to produce a preliminary picture of the keywords used in the identified papers and gathering information together. Q2 aims at recognizing the main keywords by means of a quantitative analysis of the variance and appearance in the identified papers whereas Q3 aims at providing qualitative classification of the used keywords. With Q4 the intention is to understand how keywords fit into proposed concepts, how comprehensive they are and what type of new concepts they can potentially offer.

B. Conducted Search

Keywords were identified by conducting keyword search in selected scientific databases. The search string was formed from the main research question by combining both key concepts artificial intelligence/AI + ethics. The suggested PICO process was not used to identify search string keywords because of the lack of shared concepts in AI ethics for the reasons argued earlier.

The selected scientific databases on which search was performed are shown in Table II, along with the number of publications retrieved from each database (in the 11th of March, 2018). The selection of databases were guided by the need to gain a wide coverage of the multidisciplinary nature of AI research and databases ability to handle advanced queries. The used set of keyword search strings were customized as shown in Table I to adapt to the syntax of the particular database. Web of Science and ProQuest databases do not have specified search term for author keyword, therefore keyword including the topic and subject fields were used in the search queries.

TABLE I. DATABASES AND RESEACH STRINGS

Database	Search String
IEEE Xplore (ieeexplore.ieee.org)	(("Author Keywords":ethics) AND "Author Keywords":artificial intelligence)
ACM Digital Library (dl.acm.org/advsearch.cfm)	keywords.author.keyword:(+"artificial intelligence" +ethics)
Scopus (www.scopus.com)	KEY ("artificial intelligence" OR ai AND ethics)
Web of Science (wokinfo.com)	TOPIC: ("artificial intelligence") AND TOPIC: (ethics)
ProQuest (www.proquest.com)	(SU.exact("ETHICS") AND SU.exact("ARTIFICIAL INTELLIGENCE"))

TABLE II. DATABASES AND RETRIEVED PAPERS

Database	Papers
IEEE Xplore (ieeexplore.ieee.org)	15
ACM Digital Library (dl.acm.org/advsearch.cfm)	27
Scopus (www.scopus.com)	320
Web of Science (wokinfo.com)	83
ProQuest (www.proquest.com)	617
Total retrieved	1062

C. Screening of Relevant Papers

Papers were included from search results by following criteria:

- Scholarly Journal articles
- Written in English language
- Part of current discussion, published 2012 or after
- Related to ethics and artificial intelligence or related technologies
- Full-text available for reviewing
- Author keywords available for extraction

Pre-exclusion of document type, source type and article language was done automatically in databases, see Table III. From databases five different results lists were exported and combined to reference management tool RefWorks resulting list of 588 papers. For duplicate exclusion each papers metadata and title were reviewed with aid of the reference management tool. In manual metadata analysis, papers published before 2012 were excluded. In addition, non-scholarly journal articles, for example popular articles, which were not detected in pre-exclusion phase, were excluded in the manual screening process. In in-depth review of the remaining papers, abstracts were analyzed to determinate whether the paper is related to ethics and artificial intelligence or related technologies. In the last iteration of exclusion, papers were excluded if full-text and author keywording were not available. Resulting 83 papers included. Screening process and steps can be seen in Table III and distribution by year in Fig. 1.

TABLE III. EXCLUDED PAPERS

Rationale	Amount
Pre exclusion in Database:	
Document type	-365
Source type	-96
Not in English language	-13
Manual exclusion	
Duplicate	-148
Published before 2012	-237
Academic settings or Document type	-76
Not in English language	-1
No Full-text available	-27
No author keywording available	-16
Total retrieved	1062
Total excluded	979
Total included	83

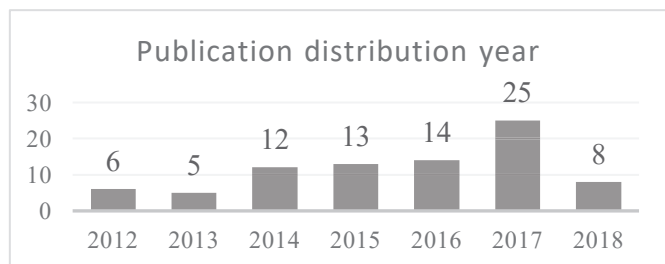


Fig. 1. Included publications distribution by year.

4. FINDINGS

For this study, the listing of the included keywords worked as a data extraction and no further keywording was conducted. To answer research questions Q1 and Q2, the keywords were listed and counted resulting in total of 324 different keywords in 83 papers. 37 of the 324 used keywords were re-occurring in two or more papers. Most frequently used keywords were Artificial intelligence/AI and ethics. This is a natural result due to the terms used in research strings, and therefore does not provide new information. These keywords were excluded from the listing. Re-occurring keywords and papers where these keywords were used can be seen on Table IV. The usage of keywords has considerable variance in incidence and spelling such as “Roboethics” and “Robot ethics” that may hinder search result. The variance in used keywords for one topic such as “Autonomous vehicle”, “Driverless cars”, “Self-driving cars” can be seen also as example of the immaturity of shared terms and undisclosed discussion what terms should be used in specific context.

TABLE IV. RE-OCCURRING KEYWORDS

Keyword	n	Found in
Machine ethics	16	[7, 12, 19-32]
Robotics	11	[20, 27, 33-41]
Robots	7	[37, 39, 42-46]
Autonomy	5	[23, 30, 44, 47, 48]
Responsibility	5	[22, 36, 49-51]
Roboethics	5	[22, 35, 41, 52, 53]
Robot ethics	4	[32, 54-56]
Artificial agents	4	[30, 50, 52, 57]
artificial general intelligence	3	[7, 24, 29]
Artificial moral agents	3	[32, 54, 58]
Automation	3	[34, 46, 59]
Consciousness	3	[31, 60, 61]
existential risk	3	[29, 62, 63]
free will	3	[52, 61, 64]
Moral agency	3	[53, 65, 66]
Moral patency	3	[53, 55, 65]
Self-driving cars	3	[58, 59, 67]
Value alignment	3	[19, 68, 69]
AI ethics	3	[19, 70, 71]
Anthropocentrism	2	[31, 36]
Artificial morality	2	[22, 58]
Autonomous agents	2	[22, 71]
Autonomous vehicle	2	[72, 73]
Driverless cars	2	[74, 75]
friendly AI	2	[29, 68]
Human rights	2	[20, 36]
human-robot interaction	2	[54, 56]
Information technology	2	[76, 77]
Machine Intelligence	2	[19, 76]
Moral status	2	[31, 70]
Personhood	2	[50, 78]
Regulation	2	[66, 79]
Rights	2	[49, 50]
Self	2	[47, 61]
Superintelligence	2	[24, 63]
Trust	2	[30, 57]
Virtue ethics	2	[46, 52]

The 37 re-occurring keywords were classified into 9 categories as shown in Table V. Classification of keywords was formed following four step process: 1) Linguistic similarity of keywords, for example similarity in spelling. 2) Ontological similarity of keyword as assumed reference for same concept. 3) Family resemblance of keywords. 4) Similarity in usage, from abstract to specific. After classification describing names were given to formed categories. [80]

The idea of classification was to outline re-occurring topics from the vast variance of keywords. This classification produced

a comparative set of more general topics relevant to AI ethics. By looking at the keywords listed it is surprising that different branches of AI such as Machine learning, Natural Language Processing or Pattern recognition were not found in the set keywords. This may imply that the relevant ethical discussion is done under separate AI branches and cannot be found through AI/artificial intelligence keyword.

Academic literature shows similarities in recurring terms when comparing keywords and the formed categories to proposed concepts and topics, but keyword listing is in some parts also partial. For example, technology based keywords and topics are underrepresented such as bias issues, fairness, transparency and controlling AI. Also socioethical topics like impact on society or workforce are lacking. [9, 13] Comparison of keyword classification reveals topics that are quite commonly shared in literary. Found keywords can be classified under the known topics even specified formulation of keywords in some parts varies considerably.

TABLE V. FORMED CATEGORIES

Category	Keywords
Conceptual	AI ethics, Machine ethics, Information technology Sports ethics, Virtue ethics, Friendly AI
Robotics	Robotics, Robots, Roboethics, Robot ethics, Automation
Generally Philosophical And Ethical	Autonomy, Autonomous agents, free will, Moral agency, Moral patiency, Moral status, Trust, Anthropocentrism Personhood, Self
AI specified Philosophical And Ethical	Artificial agents, Artificial moral agents Artificial morality
Law and Regulation	Regulation, Rights, Responsibility, Human rights
Autonomous vehicle	Autonomous vehicle, Driverless cars, Self-driving cars
AGI and AI risk	artificial general intelligence, superintelligence, existential risk
Human cognition	Intelligence, Consciousness, Machine Intelligence, human-robot interaction
Technology based	Value alignment

5. DISCUSSION & CONCLUSION

This study provided a set of AI ethics related keywords and listing of 37 re-occurring author keywords found in 83 academic papers. Re-occurring keywords were classified into 9 categories based on conceptual similarities of keywords to more general topics relevant to AI ethics. Keywords and formed categories were compared to concepts provided in academic literature to evaluate coverage of the systematic mapping study and listing. Three main differences were discovered: Lack of different branches of AI in keywords, technology based keywords have only minor role and there is a great variance in formulation of keywords even though keywords can be classified under the known topics. Recommendation for future research and systematic mapping studies: Different AI branches and different formulation for keywords extracted from known topics should be included in the keyword extraction process.

Keyword based systematic mapping study method used in this study has several weaknesses. Due to the focus on the keywords only, no primary studies of the field of AI ethics were recognized. The relevance of papers was evaluated in exclusion process and in the prevalence of keywords in the papers. Neither definitions of concepts that keywords represented were not analyzed. Despite the weaknesses, keyword based approach allowed to cover wide and loosely defined field of AI ethics to produce understanding of relevant keywords where no prior listing was available. This preliminary work also helps future systematic mapping studies by providing relevant keywords on AI ethics.

With wide variety of papers and keywords from different areas concerning AI ethics this study revealed that defining the field of AI ethics is still a challenging task. The comprehensive presentations have done a valuable work on setting definitions for expanding field of AI ethics. There is still a substantial amount of work to be done in the area. These presentations are not all inclusive and more comprehensive works are needed on the topic discussed on this paper. For example, by looking at the occurrence of different keywords, papers have different stress in different topics than comprehensive presentations have. Overall there is still research needed in the field of AI ethics on the concepts as such to see where AI ethics discourse is developing and how concepts can aid the need of practical implementation of ethics into AI systems.

This is the author's version of the work. The definite version was published in Vakkuri, V., Abrahamsson, P. 2018, June. The Key Concepts of Ethics of Artificial Intelligence. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC). IEEE. <http://dx.doi.org/10.1109/ICE.2018.8436265>

6. REFERENCES

- [1] W. Knight, "Biased Algorithms Are Everywhere, and No One Seems to Care," 2017, <https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/>, Retrieved April 21, 2018.
- [2] J.D. Greene, "Our driverless dilemma," *Science*, vol. 352, pp. 1514-1515, 2016.
- [3] Anonymous, "About Autonomous Shipping," <https://www.oneseaeosystem.net/about/about-autonomous-shipping/>, Retrieved April 21, 2018.
- [4] BMVI, "Ethics Commission's complete report on automated and connected driving," 2017, <https://www.bmvi.de/goto?id=354980>, Retrieved April 21, 2018.
- [5] IEEE Global Initiative, "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html, Retrieved April 21, 2018.
- [6] C. Allen, W. Wallach and I. Smit, "Why Machine Ethics?" *IEEE Intelligent Systems*, vol. 21, pp. 12-17, 2006.
- [7] M. Brundage, "Limitations and risks of machine ethics," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, pp. 355-372, 2014.
- [8] C. Mayer, "Developing Autonomous Systems in an Ethical Manner," *Issues for Defence Policymakers*, vol. 65, 2015.
- [9] P. Boddington, "Towards a Code of Ethics for Artificial Intelligence," Cham: Springer, 2017.
- [10] J. Dean, "Using large-scale brain simulations for machine learning and A.I." vol. 2018, June 26, 2012.
- [11] V. Charisi, L. Dennis, M.F.R. Lieck, A. Matthias, M.S.J. Sombetzki, A.F. Winfield and R. Yampolskiy, "Towards moral autonomous systems," *arXiv Preprint arXiv:1703.04741*, 2017.
- [12] R. Yampolskiy, "Safety Engineering for Artificial General Intelligence," *Topoi*, vol. 32, pp. 217-226, 2013.
- [13] P. Lin, "Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence," Oxford: Oxford University Press, 2017.
- [14] M. Anderson, S. L. Anderson, "Machine ethics," New York: Cambridge University Press, 2011.
- [15] D. Zeng, "AI Ethics: Science Fiction Meets Technological Reality," *Intelligent Systems, IEEE*, vol. 30, pp. 2-5, 2015.
- [16] Conference on Artificial Intelligence, Ethics and Society, "Call for papers (Main track)," <http://www.aies-conference.com/call-for-papers/>, Retrieved April 21, 2018.
- [17] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Keele University and Durham University Joint Report, 2007.
- [18] K. Petersen, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, vol. 64, pp. 1-18, 2015.
- [19] K. Bogosian, "Implementation of Moral Uncertainty in Intelligent Machines," *Minds and Machines*, vol. 27, pp. 591-608, 2017.
- [20] H. Ashrafian, "AlonAI: A Humanitarian Law of Artificial Intelligence and Robotics," *Sci.Eng.Ethics*, vol. 21, pp. 29-40, 2015.
- [21] J. Cervantes, "Autonomous Agents and Ethical Decision-Making," *Cognitive Computation*, vol. 8, pp. 278-296, 2016.
- [22] G. Dodig Crnkovic, "Robots: ethical by design," *Ethics and Information Technology*, vol. 14, pp. 61-71, 2012.
- [23] A. Etzioni, "The ethics of robotic caregivers," *Interaction Studies*, vol. 18, pp. 174-190, 2017.
- [24] B. Goertzel, "GOLEM: towards an AGI meta-architecture enabling both goal preservation and radical self-improvement," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, pp. 391-403, 2014.
- [25] M. Graves, "Shared Moral and Spiritual Development Among Human Persons and Artificially Intelligent Agents," *Theology and Science*, vol. 15, pp. 333-351, 2017.
- [26] D. Gunkel, "A Vindication of the Rights of Machines," *Philosophy & Technology*, vol. 27, pp. 113-132, 2014.
- [27] T. Hauer, "Society and the Second Age of Machines: Algorithms Versus Ethics," *Society*, pp. 1-7, 2018.
- [28] R.M. Omari and M. Mohammadian, "Rule based fuzzy cognitive maps and natural language processing in machine ethics," *Journal of Information, Communication and Ethics in Society*, vol. 14, pp. 231-253, 2016.
- [29] K. Sotala, "Responses to catastrophic agi risk: a survey," *Phys.Scripta*, vol. 90, pp. 018001, 2014.
- [30] H.T. Tavani, "Levels of Trust in the Context of Machine Ethics," *Philosophy & Technology* vol. 28.1, pp. 75-90, 2015.
- [31] S. Torrance, "Artificial agents and the expanding ethical circle," *AI & Society*, vol. 28, pp. 399-414, 2013.
- [32] A. van Wynsberghe and S. Robbins, "Critiquing the Reasons for Making Artificial Moral Agents," *Sci.Eng.Ethics*, 2018.
- [33] J. Bryson, "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems," *Computer*, vol. 50, pp. 116-119, 2017.
- [34] M. Coeckelbergh, "The tragedy of the master: automation, vulnerability, and distance," *Ethics and Information Technology*, vol. 17, pp. 219-229, 2015.
- [35] G. Ghilardi, "Post-human and scientific research: how engineering carried out the project," *Cuadernos De Bioetica : Revista Oficial De La Asociacion Espanola De Bioetica Y Etica Medica*, vol. 25, pp. 379, 2014.
- [36] L. Hin-Yan, "From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence," *Ethics and Information Technology*, pp. 1-13, 2017.
- [37] S. Russell, "Robotics: Ethics of artificial intelligence," *Nature*, vol. 521, pp. 415, 2015.
- [38] B. Schafer, "A fourth law of robotics? Copyright and the law and ethics of machine co-production," *Artificial Intelligence and Law*, vol. 23, pp. 217-240, 2015.
- [39] M. Szollosy, "Rapporteur's report," *Connect.Sci.*, vol. 29, pp. 254-263, 2017.
- [40] G. Tamburrini, "On the ethical framing of research programs in robotics," *AI & Society*, vol. 31, pp. 463-471, 2016.
- [41] A. Theodorou, "Designing and implementing transparency for real time inspection of autonomous robots," *Connect.Sci.*, vol. 29, pp. 230-241, 2017.
- [42] J.J. Bryson, "Of, for, and by the people: the legal lacuna of synthetic persons.(Special Issue: Machine Law)," *Artificial Intelligence and Law*, vol. 25, pp. 273, 2017.

This is the author's version of the work. The definite version was published in Vakkuri, V., Abrahamsson, P. 2018, June. The Key Concepts of Ethics of Artificial Intelligence. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC). IEEE. <http://dx.doi.org/10.1109/ICE.2018.8436265>

- [43] F. Javier Lopez Frias, "Will robots ever play sports?" *Sport, Ethics and Philosophy*, vol. 10, pp. 67-82, 2016.
- [44] D. Johnson, "Reframing AI Discourse," *Minds and Machines*, vol. 27, pp. 575-590, 2017.
- [45] P. Kopacek, "Roboethics," *IFAC Proceedings Volumes*, vol. 45, pp. 67-72, 2012.
- [46] S. Vallor, "Moral Deskillling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character," *Philosophy & Technology*, vol. 28.1, pp. 107-12, 2015.
- [47] M. Coeckelbergh, "Pervasion of what? Techno--human ecologies and their ubiquitous spirits.(Report)," *AI & Society*, vol. 28, pp. 55, 2013.
- [48] E. Colombetti, "Contemporary post-humanism: technological and human singularity," *Cuadernos De Bioetica : Revista Oficial De La Asociacion Espanola De Bioetica Y Etica Medica*, vol. 25, pp. 367, 2014.
- [49] H. Ashrafian, "Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights," *Sci.Eng.Ethics*, vol. 21, pp. 317-326, 2015.
- [50] M. Laukyte, "Artificial agents among us: Should we recognize them as agents proper?," *Ethics and Information Technology*, vol. 19, pp. 1-17, 2017.
- [51] H. Liu, "Irresponsibilities, inequalities and injustice for autonomous vehicles," *Ethics and Information Technology*, vol. 19, pp. 193-207, 2017.
- [52] V. Galanos, "Singularitarianism and schizophrenia," *AI & Society*, vol. 32, pp. 573-590, 2017.
- [53] D. Gunkel, "Introduction to the Special Issue on Machine Morality: The Machine as Moral Agent and Patient," *Philosophy & Technology*, vol. 27, pp. 5-8, 2014.
- [54] T. Arnold, "Against the moral Turing test: accountable design and the moral reasoning of autonomous systems," *Ethics and Information Technology*, vol. 18, pp. 103-115, 2016.
- [55] E. Neely, "Machines and the Moral Community," *Philosophy & Technology*, vol. 27, pp. 97-111, 2014.
- [56] K. Stowers, "Life or Death by Robot?" *Ergonomics in Design: The Quarterly of Human Factors Applications*, vol. 24, pp. 17-22, 2016.
- [57] F.S. Grodzinsky, "Developing Automated Deceptions and the Impact on Trust," *Philosophy & Technology* vol. 28.1, pp. 91-105, 2015.
- [58] C. Misselhorn, "Artificial Morality. Concepts, Issues and Challenges," *Society*, pp. 1-9, 2018.
- [59] J. Borenstein, "Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis," *Sci.Eng.Ethics*, pp. 1-16, 2017.
- [60] J.J. Bryson, "A ROLE FOR CONSCIOUSNESS IN ACTION SELECTION," *International Journal of Machine Consciousness*, vol. 4, pp. 471-482, 2012.
- [61] M.R. Waser, "SAFE/MORAL AUTOPOIESIS AND CONSCIOUSNESS," *International Journal of Machine Consciousness*, vol. 5, pp. 59-74, 2013.
- [62] K.S. Gill, "Data Driven Wave of Certainty- a question of ethical sustainability," *IFAC PapersOnLine*, vol. 49, pp. 117-122, 2016.
- [63] K. Sotola, "Superintelligence as a Cause or Cure for Risks of Astronomical Suffering," *Informatica*, vol. 41, pp. 389, 2017.
- [64] L. Frank, "Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable?" *Artificial Intelligence and Law*, vol. 25, pp. 305, 2017.
- [65] J. Bryson, "Patience is not a virtue: the design of intelligent systems and systems of ethics," *Ethics and Information Technology*, vol. 20, pp. 15-26, 2018.
- [66] D. Martin, "Who Should Decide How Machines Make Morally Laden Decisions?" *Sci.Eng.Ethics*, vol. 23, pp. 951-967, 2017.
- [67] A. Etzioni, "Incorporating Ethics into Artificial Intelligence," *The Journal of Ethics*, vol. 21, pp. 403-418, 2017.
- [68] G.P. Sarma, "Mammalian Value Systems," *Informatica*, vol. 41, pp. 441, 2017.
- [69] P. Vamplew, "Human-aligned artificial intelligence is a multiobjective problem," *Ethics and Information Technology*, vol. 20, pp. 27-40, 2018.
- [70] J. Basl, "Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines," *Philosophy & Technology*, vol. 27, pp. 79-96, 2014.
- [71] M. Wellman, "Ethical Issues for Autonomous Trading Agents," *Minds and Machines*, vol. 27, pp. 609-624, 2017.
- [72] G. Contissa, "The Ethical Knob: ethically-customisable automated vehicles and the law," *Artificial Intelligence and Law*, vol. 25, pp. 365-378, 2017.
- [73] K. Kinjo, "Optimal program for autonomous driving under Bentham- and Nash-type social welfare functions," *Procedia Computer Science*, vol. 112, pp. 61-70, 2017.
- [74] A. Etzioni, "AI assisted ethics," *Ethics and Information Technology*, vol. 18, pp. 149-156, 2016.
- [75] D. Purves, "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons," *Ethical Theory and Moral Practice*, vol. 18, pp. 851-872, 2015.
- [76] M. Loi, "Technological unemployment and human disenchantment," *Ethics and Information Technology*, vol. 17, pp. 201-210, 2015.
- [77] M. Rader, "The jobs of others: "speculative interdisciplinarity" as a pitfall for impact analysis," *Journal of Information, Communication and Ethics in Society*, vol. 10, pp. 4-18, 2012.
- [78] D.R. Lawrence, "Artificial Intelligence," *Quarterly of Healthcare Ethics* vol. 25.2, pp.250-261, 2016.
- [79] I. Rahwan, "Society-in-the-loop: programming the algorithmic social contract," *Ethics and Information Technology*, vol. 20, pp. 5-14, 2018.
- [80] A. Thomasson, "Categories," Mar 7, 2018, <https://plato.stanford.edu/archives/spr2018/entries/categories/>, Retrieved April 21, 2018.



II

ETHICALLY ALIGNED DESIGN OF AUTONOMOUS SYSTEMS: INDUSTRY VIEWPOINT AND AN EMPIRICAL STUDY

by

Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, Mikko Siponen & Pekka
Abrahamsson

To be published in electronic Journal of Business Ethics and Organization
Studies

DOI

Reproduced with kind permission by Business and Organization Ethics
Network (BON).

Ethically Aligned Design of Autonomous Systems: Industry Viewpoint and an Empirical Study

Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, Mikko Siponen, Pekka Abrahamsson

Abstract

Progress in the field of Artificial Intelligence (AI) has been accelerating rapidly in the past two decades. Various autonomous systems from purely digital ones to autonomous vehicles are being developed and deployed out on the field. As these systems exert a growing impact on society, ethics in relation to artificial intelligence and autonomous systems have recently seen growing attention among academia. However, the current literature on the topic has focused largely on theoretical contributions, and there is a gap between research and practice in the area. Though this gap has been acknowledged in existing studies, the exact issues resulting in this gap remain blurred. In order to better understand the gap in the area, we conduct a multiple case study of five case companies. Based on the data, we highlight a number of issues in the area in terms of implementing AI ethics in practice. We then propose ways to tackle this gap.

Keywords

Ethics, Artificial intelligence, Autonomous systems, Software development, Companies, Guidelines

Introduction

Artificial Intelligence (AI) systems and Autonomous Systems (AS) are becoming increasingly ubiquitous. Most inhabitants of the developed world interact with AI systems on a daily basis. The more sophisticated recommendation systems utilized by various B2C Software-as-a-Service media platforms such as YouTube utilize AI and Machine Learning (ML), and specifically Deep Learning (DL), to generate personalized recommendations for their users.

Autonomous Vehicles (AVs) operated by AI are slowly entering the public roads, AI-based surveillance systems armed with facial recognition capabilities are already being deployed, and various AI systems are being invested in and developed across fields such as medicine (Zhang, et al., 2022). In general, progress in AI has been far faster than anticipated by experts in the past.

One key difference between AI/AS and conventional software systems is that the idea of an active user is often blurred. One seldom uses AI systems as opposed to being an object to their data collection procedures or other actions. Whereas one can opt out of using conventional software systems, one often has little control over being targeted by AI systems. Moreover, some AI systems are Cyber-Physical Systems (CPS) that operate both in the digital and physical world. CPSs are various, ranging from security cameras to cargo ships, and exhibit various degrees of autonomy. CPSs such as AVs are now entering public spaces where they can interact with passers-by and cause physical damage rather than being confined to e.g., factories as factory robots (Charisi, et al., 2017).

Given their potentially enormous societal impact, AI systems should be designed while taking ethics into consideration (Bostrom & Yudkowsky, 2018; Bryson & Winfield, 2017; The IEEE Global Initiative, 2019). For example, when an AV gets into an accident, we should always be able to understand why. This is not always simple even with full access to the program code as ML systems can be highly complex even to their creators (Ananny & Crawford, 2018). Another factor that makes ethical consideration challenging at times is that the effects of the systems are not always direct (e.g., effects of individual AV on its surroundings vs. societal effects caused by 50% of the traffic being AVs).

Awareness of AI ethics issues has recently been growing in the wake of various practical incidents. For example, YLE, the Finnish national public broadcasting company, commissioned and deployed an AI-based moderation system to replace its human moderators for user comments. It was not until the system was deployed in practice and started making decisions that issues began to manifest to the point where the system was rather quickly decommissioned. This is but one of many incidents where an AI system is designed, developed, and deployed, only for it to prove unusable due to issues related to AI ethics. Similarly, users are becoming more aware of data privacy issues and are more conscious of what their data is being used for and whom it is being collected by.

As a result of the growing interest towards AI ethics related issues, a large number of guidelines have been devised to help organizations tackle AI ethics issues. These guidelines have been developed by companies, the academia, and governments (Jobin, et al., 2019). IEEE's Ethically Aligned Design (EAD) (The IEEE Global Initiative, 2019) is among these guidelines, and has been developed as a part of a particularly extensive initiative. As methods in the area remain highly technical, focusing on only subsets of the development process (Morley, et al., 2020), these guidelines have become the primary tools for implementing AI ethics for the time being.

However, though both academic and public discussion in the area of AI ethics has accelerated, the state of practice in the area remains unclear. In a past study, we argued that a gap between research and practice in the area exists, based on quantitative survey data (Vakkuri, et al., 2020). In this paper, we take a closer look at this gap to better understand the issues companies face in implementing AI ethics. Specifically, we study the current industry mindset in relation to AI ethics from the point of view of some of the most common AI ethics principles discussed in AI ethics guidelines, including IEEE's EAD (The IEEE Global Initiative, 2019). The exact research question of this paper is formulated as follows:

RQ: What practices, tools, or methods, if any, do industry professionals utilize to implement ethics into AI design and development?

The rest of this article is structured as follows. In the next section, we discuss the theoretical background of the study. In the third section, we discuss the research design. In the fourth section, we present our results, the implications of which we then discuss in the fifth section. The sixth and final section concludes the paper.

Background

In this section, we discuss the context of this study. In the first subsection, we discuss the current state of AI ethics. In the second subsection, we discuss AI in the context of Autonomous Vehicles (AVs). In the third and final subsection, we discuss commitment, which was used as the research framework for data analysis in this study.

The Current State of Ethics in AI and Ethically Aligned Design

The ethics of AI is a long-standing area of ethical discussion in ICT ethics. This discussion has accelerated notably in the past decade following technological progress in the area. As AI systems become increasingly sophisticated, hypothetical AI ethics scenarios of the past are becoming practical issues.

Indeed, researchers from various disciplines have voiced concerns over ethics in AI systems (Borenstein, et al., 2021). Following various incidents out on the field, public voices of concern have also been heard. The general public is, for example, becoming increasingly aware of data privacy issues and the way their data is handled by companies. The General Data Protection Regulation (GDPR), while not AI-specific, does end up affecting AI systems among others given how reliant most current AI systems are on large masses of data.

Laws and regulations, however, do generally tend to be slow in the face of technological progress. Some companies have already begun to consider AI ethics, publishing their own AI ethics guidelines or statements online (many of which were reviewed by Jobin et al. (2019)). It remains largely unknown to what extent these guidelines are then really employed in practice inside these organizations, but some companies are at least aware of some of the current AI ethics issues. Aside from companies (e.g., Google (Pichai, 2018)), governments (e.g. EU (AI HLEG, 2019)), and standardization institutions have also begun to work on and publish guidelines intended to help organizations implement AI ethics in practice. One such notable initiative has been the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which has since branded the concept of Ethically Aligned Design (EAD) and published a set of guidelines (The IEEE Global Initiative, 2019) featuring various principles for AI ethics, distilling much of the recent academic discussion into another set of guidelines.

These guidelines have been initial attempts at creating tools to help organizations implement AI ethics in practice. As much of the academic research on AI ethics has been conceptual and theoretical, focusing on defining and structuring AI ethics through principles and values, bringing this discussion to industry organizations presents evident challenges. The various guidelines published so far have summarized this discussion into principles, although these principles can still be difficult for developers to implement in practice. Indeed, the entire idea behind using these guidelines to implement AI ethics has been criticized (Mittelstadt, 2019).

On the other hand, methods in the area ethical AI/ML are largely technical, focused mainly on managing machine learning and other subsets of the development process (Morley, et al., 2020). Though this is important, such methods do not help with the big picture in developing ethical AI systems. In the absence of AI ethics methods to direct the development process as a whole, the aforementioned AI ethics guidelines such as EAD (The IEEE Global Initiative, 2019) have become common as tools for implementing AI ethics. Numerous such guidelines exist, and though they discuss different principles, some consensus in the area already exists. (Canca 2020; Jobin, et al., 2019)

Indeed, the ongoing academic discussion on ethics in AI has so far converged on different principles, some of which are also discussed in EAD (The IEEE Global Initiative, 2019). Jobin et al. (2019), based on their analysis of 84 AI ethics guidelines, argued that the following principles were the most common ones, in a descending order of popularity: (1) transparency, (2) justice, fairness and equity, (3) non-maleficence, (4) responsibility and accountability, (5) privacy, (6) beneficence, (7) freedom and autonomy, (8) trust, (9) sustainability, (10) dignity, and (11) solidarity. In our analysis, we utilize transparency, accountability, and responsibility, as well as what we argue can be considered a subset of transparency, predictability, as a framework for the data collection in this study (as we discuss again in the third section).

Transparency is the central AI ethical construct present in most AI ethics guidelines (Jobin, et al., 2019). Turilli and Floridi (2009) argue that it is, in fact, the pro-ethical circumstance that

makes it possible to implement AI ethics in the first place. Very related to transparency is also the idea of *explainable* AI systems, which has recently been discussed extensively both in academia and among practitioners (e.g., Adadi & Berrada, (2018); Rudin, (2019)).

We consider there to be two types of transparency: (1) transparency of algorithms and data (Dignum, 2017) (i.e., the transparency of systems), and (2) transparency of systems development (i.e., decision-making etc.). Predictability can be considered a subset of transparency, as the EAD guidelines do (The IEEE Global Initiative, 2019), and as we thus do in our analysis. As the word implies, it refers to whether the system acts predictably. For example, if an autonomous coffee machine successfully brews coffee 8 times out of 10, we are left wondering what happened the other two times and why.

Accountability and responsibility are in some ways related, though still separate constructs. Accountability focuses on who is accountable or liable for the decisions made by the AI. Dignum (2017), in her work, defines accountability to be the explanation and justification of one's decisions and one's actions to the relevant stakeholders. Transparency is required for accountability, as we must understand why the system acts in a certain fashion, as well as who made what decisions during development in order to establish accountability. Whereas accountability can be considered to be externally motivated, responsibility is internally motivated. Responsibility can be considered to be an attitude or a moral obligation for acting responsibly (The IEEE Global Initiative, 2019). In order to act responsibly, one has to weigh their options and consciously evaluate the effects of their actions and decisions.

These three main constructs (Transparency, Accountability, and Responsibility) and one sub construct (Predictability) are our focus in this study. They are AI ethics principles that have become some of the most prominent ones commonly featured in the numerous AI ethics guidelines currently in existence (Jobin, et al., 2019). We discuss this choice further in the research design section that follows.

To conclude this section, we further position this paper in this area. While ethics in AI has become a prominent topic among the academia, as well as in public discussion, the current state of industrial practice remains unclear. In another study, we argued that there is a gap between research and practice in the area (Vakkuri, et al., 2020). However, the exact nature of this gap is not clear. The focus of this paper is to further explore the situation in the industry and to begin tackling the present lack of tooling for EAD and other AI ethics guidelines. By better understanding the gap in the area, we are able to provide better tools to tackle the issues out on the field.

Artificial Intelligence and Autonomous Vehicles

Currently, AVs are being developed across industries. Though arguably the most media exposure is on cars given their nature as B2C personal vehicles, the possibilities of AI have been explored in relation to drones, cargo ships, buses, trains, and airplanes alike. While the degree of autonomy exhibited by various types of vehicles is steadily increasing, fully autonomous vehicles are still rarely used in practice. Such vehicles are actively being tested in various fields, however.

Safety in these systems is a justified and widely acknowledged concern (Nascimento, et al., 2020). Regardless of software quality in AVs, accidents and dangerous situations are inevitable. Such situations may, for example, result from faulty sensors. However, whereas human actors seldom have time to make a carefully thought-out decision in the face of an impending accident, and may sometimes be too slow to properly react at all, AI systems are capable of making a decision near instantaneously. Thus, such systems are required to make difficult ethical decisions in situations where accidents are inevitable one way or the other (Evans, et al., 2020). This includes dilemmas such as the commonly cited example “Should Your Car Kill You to Save Others?” (Bonneton, 2016; Lo Piano, 2020).

From the point of view of AI ethics, the AI ethics principles discussed in the preceding subsection also apply in the context of AVs. Accountability, for example, can be argued to be even more relevant when material damage is a possibility. Similarly, data and data-related issues are also relevant for AVs.

In practice, ethical issues are ultimately left for the developers to tackle. Though company level policies and guidelines can direct development work, micro-level decisions are nonetheless left to individual developers. Thus, developers working with AI need to be able to implement ethics into the systems they develop. This calls for both awareness of AI ethics among developers, as well as tools to implement it (Vakkuri, et al., 2021). Currently, little is known about how AI ethics is handled in practice in organizations.

Commitment

As the theoretical framework for this study, we approach ethics in AI through the lens of commitment. In industrial psychology and organizational behavior, commitment is a long-standing area of research (Benkhoff, 1997). The idea of commitment has been of interest primarily because of the assumption that the commitment of employees relates to performance. O’Reilly and Chatman (1986) remark that “although the term commitment is broadly used to refer to antecedents and consequences, as well as the process of becoming attached and the

state of attachment itself, it is the psychological attachment that seems to be the construct of common interest." Drawing from this, we consider commitment to be the attachment an individual feels towards an object (organization, ideal etc.).

Aside from behavioral studies from fields such as psychology, commitment has been studied in the past in relation to software process improvement (SPI) (Abrahamsson, 2002). Abrahamsson (2002) proposed a model of commitment nets (Figure 1). The model suggests that drivers, both internal and external, may result in concerns which would then manifest as actions, and those actions would then lead to both intended and potentially unintended outcomes. Commitment, in this model, can be observed when concerns result in actions. We utilize this commitment net model as the theoretical framework of this study, as we discuss in detail in the next section.

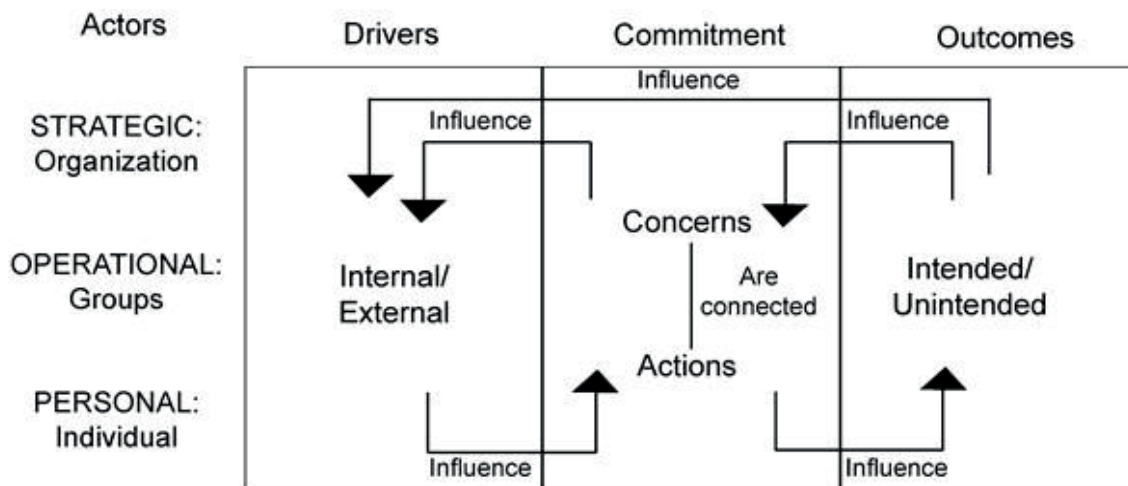


Figure 1 The Commitment Net Model of Abrahamsson (2002)

Research Design and Protocol

This study was carried out as a multiple case study of five cases (Table 1). Each case company develops AI systems, although in different fields, or only as a portion of their operations. Data were collected via semi-structured qualitative interviews. The interview instrument in its entirety can be found in the Appendix 1.

Table 1 Case Company Information

#	Company Description	Respondent [Reference]
1	Large, >400 employees; Software, Generic	Data Scientist [R1]; Senior Data Scientist [R2]
2	SME (Small/Micro), <25 employees; Software, Healthcare	Development Lead [R3]
3	SME (Small/Micro), <25 employees; Software, Process Industry	CTO [R4]
4	Large (Multinational), >100 000 employees; Consulting	Functional Designer [R5]
5	Large (Multinational), >25000 employees; Vehicle Industry	AI Development Lead [R6]

In short, the interview protocol (Appendix 1) was designed to focus on the key constructs discussed in background section: transparency, accountability, responsibility, and predictability. We avoided directly discussing ethics as different individuals have different conceptions of what ethics is in this context. This is underlined by the on-going academic discussion as well (see for example (Friedman et al., 2013)). Instead, we focused on asking practical questions related to these ethical principles.

In devising this interview instrument, we chose to focus on some of the earlier AI ethics themes that have remained prominent. The principles we focused on in the interview instrument were common in the various guidelines reviewed by Jobin et al. (2019) and are present in IEEE's EAD (2019) as well. Thus, though we focus on only some AI ethics principles, the principles utilized here are some of the most central ones. We discuss this research framework behind the interviews in detail in another paper (Vakkuri, et al., 2019).

We utilized the commitment net model of Abrahamsson (2002) (see Section Commitment) as the theoretical framework for the analysis of these cases. We approached commitment through the concerns that the employees might have had towards implementing ethics in AI design, as well as through the actions they might have taken as a result of their concerns.

To analyze the data, we used the grounded theory method inspired by (Corbin et al., 2014). After the interview, data was coded to classify themes and we focused on concerns the respondents had towards AI ethics issues. Then, after identifying concerns, we looked at what actions the respondents (or their organizations) had taken to tackle these concerns - if any at all. By doing so, we sought to understand whether any commitment towards ethics in AI design existed in the case companies. To give a practical example, if one indicates concern towards losing weight, but exhibits no actions such as making dietary changes or exercising, there is no commitment present.

However, the goal of this analysis was not to find out whether the organizations exhibited any commitment towards AI ethics. Rather, we focused on the actions the respondents and their organizations had taken to address their concerns. In this fashion, we wanted to identify any practices, tools, or methods that had been used to address ethical concerns, i.e., to find out how the respondents had implemented AI ethics.

In our analysis of the data, we summarize our findings through what we refer to as Primary Empirical Contributions (PEC). We consider these to be findings that are worth noting despite occasionally being outside the direct scope of our research question. These PECs are then further discussed in the discussion section and provide a framework for it.

Empirical Results

Our interviews of the case companies indicated that the industry is aware of the potential importance of AI ethics. Every respondent agreed that ethics is useful. However, the case companies had highly differing views on how relevant it was in practice, and none of them remarked using development practices that directly supported implementing it. This underlined that the companies did not have clear tools or methods for implementing ethics. This disconnect seemed, in part, to also stem from a lack of consensus on what (AI) ethics actually referred to. As a part of our empirical results, we elaborate some of our findings with relevant quotes from the respondents. However, our findings are not solely based on the quotes, but on our data in general.

"...I actually try to use the word 'ethics' as little as possible because it's the kind of word that everyone understands in their own way, and so they can feel that it's not relevant to what we're doing at all..." [R4]

"...the discussion on AI ethics doesn't really affect most ... excluding maybe Google and some others like that ... the AI really isn't at the level where it would really necessitate in-depth ethical consideration" [R3]

PEC1: Ethics is considered important in principle, but as a construct it is considered detached from the current issues of the field by developers. In other words, the on-going academic discussion on AI ethics has not reached the industry at large.

Only the respondent involved in developing a medical AI system had a more practical view of ethics in relation to their current project. However, the respondent noted that the ethical consideration had already been carried out externally. Indeed, fields such as the field of

medicine inherently have very strict regulations regarding, for example, data management, leaving little leeway for developers to make their own ethical decisions:

"We have in-house quality measurements and these regulation requirements are very strict, so these things pretty much come as a given for us. And, of course, if you think about it the other way, we consequently think about these things [ethics] even less because we already have such clear regulations and requirements for what we do" [R3]

PEC2: Regulations force developers to take into account ethical issues while also raising their awareness of them.

On the other hand, though ethics as a construct was considered impractical and too theoretical, the respondents did all nonetheless concern themselves with various constructs related to AI ethics (in this case: transparency, predictability, accountability, and responsibility). These constructs were considered practical by the respondents, as we discuss in the following subsections.

Transparency

All case companies were concerned with both transparency of systems and transparency of systems development. Furthermore, transparency of systems was considered both from the point of view of developers and users. However, the actions taken to address these concerns (if any) were varied (Table 2) across cases:

"The most important thing is that we can see directly how it works, and that it's trackable, now, and later." [R5]

"...it is typically a little un-transparent how the decisions are made. Of course we can analyze them, but due to the complexity of the neural network architecture, it's a little difficult to accurately explain why it did something." [R6]

Table 2 Commitment Towards Transparency to Developers

Driver	Actor	Concern	Action(s)
Project need	R1	Keeping the system understandable to developers (i.e. transparency to developers)	No recognized actions
Legislation; Regulations	R3		Devoting time to understanding the training data
Company need	R4		Devoting time to understanding the AI used as a template for the system; Building analytics into the system
Company need	R5		No recognized actions; (Planned future action: documentation)
Company need	R6		Devoting time to understanding the training/testing data; Mode verification

Whereas transparency from the point of view of developers was considered in relation to e.g., the algorithms and the neural network architecture, transparency from the point of view of the users was considered on a less technical level (Table 3). The respondents felt that the users had little reason to be able to see inside the system or the so-called black box as such. It was considered more important that the users would be able to understand how it works on the very basic level:

"Our systems are aimed at these... operational personnel, like the paper plant guys down on the factory floor [...] they don't really know what happens inside the system and we don't feel that they really need to know, either [...] they just understand that, okay now all this data goes in, and the suggestions are then based on that data" [R4]

"...the users won't really notice a difference compared to the earlier systems they have used. We just want to offer them better and more timely data. So that's of course one question: how to make it clear for them that there are some uncertainties there so that they don't expect the information to always be perfect. But... I don't really know how much of a problem this is -- I haven't really spoken to our end-users" [R5]

PEC3: Developers have a perception that the end-users are not tech-savvy enough to gain anything out of technical system details.

Table 3 Commitment Towards Transparency to Users

Driver	Actor	Concern	Action(s)
Project need	R1	Keeping the system understandable to the end users (i.e. transparency to users)	No recognized actions
No clear driver	R2		Educating the customer/user
Market edge; Process improvement	R3		No recognized actions
Company need; Professionalism	R4		Educating the customer/user
Company need	R5		Writing helpful system descriptions
Company need; Professionalism	R6		Educating the customer/user; Communication with customer/user

In terms of transparency of systems development, four of the five companies indicated clear concern towards it and had taken actions to address the concern (Table 4). Largely, (code) documentation was considered to be the primary way of producing transparency in the development process by making it apparent who made what changes, why, and when. Additionally, conducting audits was discussed as one tangible practice for producing transparency in the development process. This was one of the few areas where a consensus among the companies could be observed in ethical practices.

PEC4: Documentation and audits are established Software Engineering project practices that form the basis in producing transparency in AI/AS projects.

Table 4 Commitment Towards Transparency of Development

Driver	Actor	Concern	Action(s)
Project need; Customer need	R1	Keeping track of who does and decides what and why (i.e., transparency of development)	Documentation
Project need; Customer need	R2		Documentation; Conducting audits; Distinct roles in development team
Customer need; Market need; Regulations	R3		Documentation; Conducting audits, audit trail
Company need	R5		Documentation
Company need	R6		Launch of new management process

Predictability

One of the main concerns shared by all respondents was the potential unpredictability of the system (Table 5). The respondents discussed clear actions they had taken to either avoid unpredictable behavior, to mitigate it, or to prevent it in the future in case it takes place. An example of such an action can be ML management by means of using different sets of training data or by limiting its utilization.

"...we have even cut some functionalities [...] of the system in order to make it more predictable, which has reduced the amount of unexplained results we have gotten out of it [...] in practice we've been able to explain all of the faulty results so far" [R3]

PEC5: Machine learning is considered to inevitably result in some degree of unpredictability. Developers need to explicitly acknowledge and accept heightened odds of unpredictability.

Table 5 Commitment Towards Preventing Unpredictability

Driver	Actor	Concern	Action(s)
No clear driver	R1	System acts unpredictably (i.e., preventing an incident)	Awareness of unpredictability; Recognizing what errors are acceptable; Preparedness for incidents of unpredictability
Company need	R2		Representative training data; Training for designer
No clear driver	R3		Reduce functionalities and complexity of system; Narrow the scope of use of machine learning
No clear driver	R4		Accept the (minimal) odds of unpredictability; Acknowledging that statistical tools also make mistakes; Root cause analysis
No clear driver	R5		Using the system only in confined spaces
Company need	R6		AI/ML model validation

When discussing steps taken to avoid unpredictability, the respondents also discussed their concerns related to a hypothetical situation in which the system has already acted unpredictably (Table 6). All six respondents and five case companies had outlined some courses of action for such a scenario, although some of the actions pointed towards a lack of commitment (e.g., apologizing and reacting on a case-by-case basis is a very ad hoc plan).

Table 6 Commitment Towards Addressing an Incident of Unpredictability

Driver	Actor	Concern	Action(s)
Customer need; Company need	R1	System makes mistake in production (i.e. hypothetical scenario in which an incident took place)	Accept the (minimal) odds of unpredictability; Be willing to react; Apologize
Company need; Project need; Professionalism	R2		Be willing to react; Apologize; [Planned future action: communication/ action plan]
Customer need; Financial	R3		Feedback options to product development; Using mistake as example in learning data; Accept the unlikely unpredictability; Acknowledging that statistical tools also make mistakes
No clear driver	R4		Piloting before full release; Reacting feedback and fixing issues; Narrowing functionalities in design
Company need; Customer need	R5		Piloting oversight; Cutting system functionalities; Fixing bugs when noticed
Company need; Customer need; Legislation	R6		Backup systems

Finally, in relation to predictability, four of the respondents discussed cyber security threats as a part of unpredictable system occurrences (Table 7), even if they are caused by external actors as opposed to the system itself. Indeed, in the case of especially CPSs, cybersecurity threats can pose life-threatening danger if e.g., an autonomous bus is hijacked digitally. Given that cybersecurity is a longstanding area of research and industry practice, companies generally have established policies and even cybersecurity departments for dealing with cybersecurity issues. Thus, few actionable measures or practices were underlined by the respondents in response to their actions in tackling cybersecurity concerns.

Table 7 Commitment Towards Cybersecurity

Driver	Actor	Concern	Action(s)
Company need; Customer need	R1	Cybersecurity / Data security / Adversary attacks	Follow quality process and corporate policy
Company need; Project need; Professionalism	R2		Recommendations on how to prepare; Awareness of context of use (i.e., who can do what with the system)
Company need; Customer need; Legislation	R3		Follow quality process and corporate policy
Company need; Customer need	R6		Backup systems; Preparing for attacks

Accountability and Responsibility

The consensus among the respondents was that no system could be completely fault-free, with five respondents expressing concern towards potential harm caused by their system(s) (Table 8). Most respondents could also list some actions their organization had taken to either avoid or mitigate harm caused by their system. However, some of the respondents felt that their system(s) had no direct potential for harm even if it did act unpredictably or wrongfully, due to it e.g., being a purely digital business intelligence system.

PEC6: Developers consider the harm potential of a system primarily in terms of physical harm. Potential systemic effects are often ignored.

Additionally, the respondent working on healthcare AI (R3) indicated a more personal approach to responsibility than the other respondents as they felt that they were directly responsible for the well-being of some of their users.

PEC7: Physical harm potential motivates personal drivers for responsibility.

Notably, the respondents ultimately outsourced the responsibility and/or accountability to their users despite exhibiting a commitment to mitigate or prevent harm. They felt that they had taken what measures they could to prevent harm, and that it was then up to the user to stay safe (e.g., doctors should be critical of the suggestions of medical AI):

PEC8: Main responsibility is outsourced to the user, regardless of the degree of responsibility exhibited by the developer.

Table 8 Commitment Towards Responsibility for Potential Harm

Driver	Actor	Concern	Action(s)
Customer need	R1	Responsibility for potential harm caused by the system or a specific algorithm	Adhering to contracts; Responsible project management
Company need; Project need; Personal	R2		No recognized actions
Personal	R3		Accept the (small) odds of harm; Communication with the customer to minimize the risk of harm
No clear driver	R5		Design the system so that even wrong decisions are not harmful.
No clear driver	R6		Minimizing potential harm; Accept small odds of harm; Build a system that produces less harm than humans in the same context

As the respondents discussed having concerned themselves and their project teams very little with direct discussions about ethical matters related to their systems, they did not consider responsibility strongly from an ethical point of view. Instead, they approached responsibility largely from the point of view of delivering a product that fulfilled expectations set for it (Table 9) internally, by various stakeholders, or by regulations. Some of the respondents also felt that delivering a quality product was their responsibility as professionals of the field.

PEC9: Developers typically approach responsibility pragmatically from a financial, customer relations, or legislative point of view rather than an ethical one.

Table 9 Commitment Towards Addressing an Incident of Unpredictability

Driver	Actor	Concern	Action(s)
Company need; Commercial; Professionalism	R1	Delivering a working product / Delivering what was promised	Setting realistic goals for the system
Commercial	R3		No recognized actions
Company need; Customer need; Professionalism	R4		Piloting; Keeping the human in the loop
No clear driver	R5		Discussion inside project team; Communication with customer

Discussion

We have collected the Primary Empirical Contributions (PECs) outlined in the results section into Table 10. They have been split into three categories based on their contribution: (1) empirically validates existing literature, (2) contradicts existing literature, and (3) new knowledge. Overall, the primary contribution of this study is its empirical approach focusing on developers and the state of practice. Existing studies in the area have been largely theoretical.

The most general finding of this study is that it further confirms that there is a gap between research and practice in the field of AI ethics (PEC1). The academic discussion on AI ethics and the values related to it (transparency, etc.) seems to not have affected the industry yet. This is consistent with the findings of McNamara et al. (2018) who concludes that the ACM Code of Ethics (Gotterbarn, et al., 2018) has done little to change the way developers work. Whittlestone et al. (2019), Mittelstadt (2019) and Canca (2020), also argue that guidelines are likely to be difficult to implement in practice out in the field. Moreover, we have also argued that there is indeed such a gap in the area in another paper with a quantitative approach (Vakkuri, et al., 2020). There thus seems to be a clear gap between research and practice in the area. The rest of the findings of this study serve to further our understanding of said gap.

We argue that this gap largely stems from a lack of tooling and methodologies in the area, as has been suggested by Whittlestone et al. (2019) as well. Based on our data, industry professionals currently address ethical issues through various ad hoc practices. While numerous guidelines exist (Jobin, et al., 2019), they are not actionable (Whittlestone, et al., 2019; Canca 2020) and consequently see little use. Tools and methods are needed to make them actionable. Currently, tools and methods in the area offer little help in designing ethical AI

systems and managing the big picture, as they focus on the technical aspects of the development such as managing ML (Morley, et al., 2020).

To help in tackling this gap in practice, we have begun to work on a method to help implement AI ethics in practice. This method, ECCOLA, that builds on existing research, has been developed by researchers and applied in industry projects. We have published this method in another paper (Vakkuri, et al., 2021). It is an on-going initiative, and though ECCOLA is still being developed further, it has reached a state of maturity where we wish to share the method with the scientific community, as well as the industry.

Aside from tooling, one way of addressing this gap would be through changes in legislation and regulations (PEC2). However, legislative changes are slow and may struggle to keep up with the advances in technology. They may also have negative, limiting effects on AI development (e.g., regulations on international waters limit testing maritime AVs). Nonetheless, legislation and regulations are starting to address AI issues, with the General Data Protection Regulation (GDPR) and the upcoming AI Act affecting AI systems in the EU area.

However, it should nonetheless be noted that some companies do seem to utilize these AI ethics guidelines. Nagadivya et al. (2020) studied companies using AI ethics guidelines to guide AI system development and argue that they can be useful in doing so. Arguably, the guidelines certainly do provide a starting point for implementing AI ethics, even if it takes effort from the organization to make them actionable. It would seem, though, that most organizations currently do not wish to devote resources towards doing so.

Table 10 Primary Empirical Conclusions of the Study

#	Theoretical component	Description	Contribution
1	Conceptual	Ethics is considered important in principle, but as a construct it is considered detached from the current issues of the field by developers.	Empirically validates existing literature
2	Conceptual	Regulations force developers to take into account ethical issues while also raising their awareness of them.	Empirically validates existing literature
3	Transparency	Developers have a perception that the end-users are not tech-savvy enough to gain anything out of technical system details.	Contradicts existing literature
4	Transparency	Documentation and audits are established Software Engineering project practices that form the basis in producing transparency in AI/AS projects.	Empirically validates existing literature
5	Transparency	Machine learning is considered to inevitably result in some degree of unpredictability. Developers need to explicitly acknowledge and accept heightened odds of unpredictability	Empirically validates existing literature
6	Responsibility; Accountability	Developers consider the harm potential of a system primarily in terms of physical harm. Potential systemic effects are often ignored	New knowledge
7	Responsibility; Accountability	Physical harm potential motivates personal drivers for responsibility.	Empirically validates existing literature
8	Responsibility; Accountability	Main responsibility is outsourced to the user, regardless of the degree of responsibility exhibited by the developer.	New knowledge
9	Responsibility; Accountability	Developers typically approach responsibility pragmatically from a financial, customer relations, or legislative point of view rather than an ethical one.	New knowledge

Indeed, based on our findings, it seems that developers currently do not approach ethics in a systematic manner and do not utilize any tools or methodologies to implement it. However, ethical values discussed in academic literature are nonetheless taken into account in the industry to some extent. According to the IEEE EAD guidelines (The IEEE Global Initiative, 2019), documentation is a key in producing transparency. This was also acknowledged by all case companies (PEC4), although the sufficiency of their documentation remains unknown. Similarly, the challenges ML poses to system predictability are discussed in existing literature and also acknowledged by industry professionals (PEC5).

On the other hand, while the IEEE EAD guidelines (The IEEE Global Initiative, 2019) and other such guidelines typically encourage transparency in terms of providing users with technical details of the systems as well, developers feel that their users do not possess the technical knowledge to make any use of said information (PEC3). Here the opinions of the developers also notably contradict existing literature in which transparency has been extensively discussed e.g., from the point of view of the users or the general public being able to understand the technical side of the system.

In terms of responsibility, developers do not seem to possess the skills to evaluate the harm potential of AI systems comprehensively. They exhibit a narrow view of the harm potential of such systems, focusing on physical harm (PEC6). This is a topic that has not been extensively studied thus far but practical incidents do point towards this being the case. In other words, either developers are unaware of these issues or they are simply ignored, e.g., in favor of financial gain. While developers exhibit more responsibility if they consider the system to have physical harm potential (PEC7), social and emotional impacts of AI systems are ignored (PEC6). Developers also do not consider the systemic effects of AI systems, which can be important (German Federal Ministry of Transport and Digital Infrastructure, 2017). This further highlights the gap in the area, as AI ethics literature discusses the harm potential of AI systems extensively and takes into account social issues such as racial bias (See for FAccT community focusing fairness, accountability, and transparency in socio-technical systems).

However, we do feel that one cannot expect developers to conduct such comprehensive ethical analysis unassisted and without training. Training developers (or university students who will go on to become developers in the future) to take into account AI ethics and teaching them how to do so is important. Additionally, carrying out such ethical analyses calls for distribution of work in organizations, or even hiring ethical experts to carry out the analysis (Canca 2020). Furthermore, we once more underline the importance of tools and methods in this regard.

Moreover, in relation to responsibility, developers seldom consider responsibility important purely for ethical reasons. Rather than being concerned about being ethical, they are concerned about potential financial losses or bad publicity resulting from the system being unethical (PEC9). This is to some extent similar to how companies have approached environmental issues or business ethics at large, although nonetheless new in the specific context of AI ethics. Companies are more likely to tackle these issues for financial or legislative reasons, as opposed to doing so simply to act responsibly. This should be considered when attempting to raise awareness of AI ethics in the industry.

Regardless of the degree of responsibility exhibited by the developers, the responsibility is ultimately outsourced to the user(s) of the system (PEC8). In other words, the developers feel that the user should always be critical towards the suggestions of the system, whether the user is a doctor or a factory worker, and that how they use the system is their responsibility. Similar lines of argumentation are seen, for example, in relation to firearm legislation, and thus while

this is new in the context of AI ethics, outsourcing responsibility in this sense as a phenomenon is not novel.

Also, outsourcing responsibility in this context is interesting when combined with PEC3, as the developers simultaneously feel that their end-users are not tech savvy enough to benefit from being explained or shown the technical details of the system. Yet, despite the users thus having no in-depth understanding of how the systems work, the developers feel that the users should be able to evaluate the actions of the systems in an informed fashion. This issue has been, in part, acknowledged in existing literature. Scholars have repeatedly voiced their concerns over black boxes and demanded explainable AI systems. (Bryson & Winfield, 2017; Adadi & Berrada, 2018). Recently the demand has even switched beyond explainable AI and ML models to interpretable models (Rudin, 2019).

In terms of future research directions, we recommend any studies seeking to address the evident gap between research and practice in the area. This includes further studies into the state of practice (e.g., further studies on how companies implement AI ethics when using AI ethics guidelines to do so), as well as tools or methods for implementing AI ethics.

Limitations of the Study

The generalizability of the findings is always an issue for qualitative case studies. Given the qualitative approach of this study, we cannot claim that our results would be representative of the current state of the industry at large with 5 case companies involved. However, we would turn to Eisenhardt (1989) who argues that for novel research areas, five cases is an acceptable number.

Empirical studies in AI ethics, including those looking into the current state of the art, are currently still few in number and there seems to be a gap in the area between research and practice (see for example (Vakkuri et al., 2020) or (Morley, et al., 2020), which leads us to argue that this is a novel area of research.

Another limitation is, still related to these case companies, that all the case companies were either Finnish or international companies whose Finnish branch was the only one involved in this study. This is a potentially notable limitation in this context because much of the discussion on AI ethics has been US-based. Therefore, it is possible that especially US companies might be more concerned with AI ethics than companies based in Finland. However, in another study (Vakkuri, et al., 2020), we have taken on a quantitative approach to studying the current state of practice and did not find any notable differences between Finnish and US companies.

Finally, the research framework used in this study presents some limitations as well. In particular, the construct of ethics can impose threat to the validity of this study as ethics and values have tendency to mean different things to different individuals (Friedman et al., 2013). In an attempt to tackle this limitation, the concept of *ethics* was approached through more context related sub-constructs (grounded in existing research) and questions directly mentioning ethics were kept to a minimum. As much of the research so far in AI ethics has focused on defining principles for ethical AI systems, existing research in the area offered various concepts that could be used for this purpose. In this study, we have utilized, but some of these (transparency, accountability, responsibility, and predictability). While these themes are central, with e.g., transparency being the most high-profile one (Jobin, et al., 2019), there are various other principles associated with AI ethics. Our approach, thus, only focused on some aspects of AI ethics. Additionally, while planning the interview protocol and conducting the data collection, we have mostly kept our distance as researchers, maintaining a distinct role and doing our best to only collect data while avoiding advising or leading the participants on into any direction.

Conclusions

In this paper, we have conducted a case study to understand the current state of practice in relation to ethics in AI. The case study featured five case companies, in which the data was gathered through semi-structured, qualitative interviews. We utilized the commitment net mode and grounded theory to analyze the data through the concerns the organizations or individuals exhibited towards various ethical issues, as well as the actions they had taken to address said concerns.

In summary, developers consider ethics important in principle. However, they consider ethics as a construct impractical and distant from the issues they face in their work. There is thus a clear gap between research and practice in the area as the developers are not aware of the academic discourse on the ethics of AI.

The key finding of this study was that none of the case companies utilized any tools or methodologies to implement AI ethics. Based on our data, it seems that developers lack ways to systematically implement AI ethics into practice. They tackle ethical issues separately from other development tasks and in an ad hoc fashion, using highly differing practices across organizations. While various guidelines for AI ethics currently exist, written by both practitioners and scholars alike, these guidelines are not used by industry experts. One reason behind this lack of adoption is likely the fact that these guidelines consist of principles and values rather than actionable practices, which can make them challenging to utilize in practice. At very least, this results in a situation where organizations hoping to utilize these guidelines in practice must devote resources towards first making them actionable for the developers.

We recommend that future studies seek ways to make these guidelines, or AI ethics in general, actionable for the industry. This could be achieved in a number of ways. For example, methods and tools can help organizations implement AI ethics in practice. Alternatively, among other options, a maturity model for AI ethics focusing on processes could also help in this regard. Ultimately, and in any case, it seems that guidelines may not be the way to proceed and that we should look elsewhere when it comes to making AI ethics practical. A large number of AI ethics guidelines already exists and it is unlikely that any new set of guidelines would provide a notable contribution at this point.

References

Abrahamsson, P. (2002) "Commitment nets in software process improvement", *Annals of Software Engineering*, Vol. 14, No. 1, pp. 407-438.

Adadi, A and Berrada, M. (2018) "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)", *IEEE Access*, Vol. 6, pp. 52 138-52 160.

AI HLEG (High-Level Expert Group on Artificial Intelligence) (2019) "Ethics guidelines for trustworthy ai". Available <https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1>

Ananny, M. and Crawford, K. (2018) "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability", *New Media & Society*, Vol. 20, No. 3, pp. 973-989

Balasubramaniam, N. and Kauppinen, M. and Kujala, S. and Hiekkanen, K. (2020) "Ethical guidelines for solving ethical issues and developing ai systems", *Product-Focused Software Process Improvement*, pp. 331-346.

Benkhoff, B. (1997) "Disentangling organizational commitment: The dangers of the ocq for research and policy", *Personnel Review*, Vol. 26, No. 1, pp. 114-131.

Bonnefon, J. F. and Shariff, A., and Rahwan, I. (2016). "The social dilemma of autonomous vehicles. *Science*", Vol. 352(6293), pp. 1573-1576.

Borenstein, J., Grodzinsky, F.S., Howard, A., Miller, K.W., & Wolf, M.J. (2021). "AI ethics: A long history and a recent burst of attention", *Computer*, 54(1), pp. 96-102.

Bostrom, N. and Yudkowsky, E. (2014), "The ethics of artificial intelligence", in Frankish, K. and Ramsey, W.M. (Eds.), *The Cambridge handbook of artificial intelligence*, Cambridge University Press. pp. 316-334.

Bryson, J. and Winfield, A.F. (2017) "Standardizing ethical design for artificial intelligence and autonomous systems", *Computer*, Vol. 50, No. 5, pp. 116-119.

Canca, C. (2020). "Operationalizing AI ethics principles", *Communications of the ACM*, 63(12), pp. 18-21.

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M. Sombetzki, J. and Winfield, A.F. and Yampolskiy, R. (2017). Towards moral autonomous systems. arXiv preprint. Available <https://doi.org/10.48550/arXiv.1703.04741>

Corbin, J. and Strauss, A. (2014). "Basics of qualitative research: Techniques and procedures for developing grounded theory, Sage publications.

Dignum, V. (2017) "Responsible autonomy", arXiv preprint. Available <https://doi.org/10.48550/arXiv.1706.02513>

Eisenhardt, K. M. (1989) "Building theories from case study research", *The Academy of Management Review*, Vol. 14, No. 4, pp. 532-550.

Evans, K. and de Moura, N. and Chauvier, S. and Chatila, R. and Dogan, E. (2020) "Ethical decision making in autonomous vehicles: The av ethics project", *Science and Engineering Ethics*.

Friedman, B. and Kahn, P. H. and Borning, A., and Huldtgren, A. (2013), "Value sensitive design and information systems", in Doorn, N. and Schuurbiens, D. and Van de Poel, I. and Gorman, M. E. (Eds.), *Early engagement and new technologies: Opening up the laboratory*, Springer, Dordrecht. pp. 55-95.

German Federal Ministry of Transport and Digital Infrastructure (2017). "Automated and Connected Driving". Available <https://www.bmvi.de/EN/Topics/Digital-Matters/Automated-Connected-Driving/automated-and-connected-driving.html>

Gotterbarn, D. W. and Brinkman, B. and Flick, C. Kirkpatrick, M. S. and Miller, K. and Vazansky, K. and Wolf, M. J. (2018) "Acm code of ethics and professional conduct", *Association for Computing Machinery*. Available <https://www.acm.org/code-of-ethics>

Jobin, A. and Ienca, M. and Vayena, E. (2019) "The global landscape of ai ethics guidelines", *Nature Machine Intelligence*, Vol. 1, No. 9, pp. 389-399, 2019.

Lo Piano, S. (2020) "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward", *Humanities and Social Sciences Communications*, Vol. 7, No. 9.

McNamara, A. and Smith, J. and Murphy-Hill, E. (2018) "Does ACM's code of ethics change ethical decision making in software development?" *Proceedings of the 2018*

26th ACM ESEC/FSE, pp. 729-733.

Mittelstadt, B. (2019) "Principles alone cannot guarantee ethical ai", Nature Machine Intelligence, pp. 1-7.

Morley, J. and Floridi, L. and Kinsey, L. and Elhalal, A. (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, Vol.26 No.4, pp. 2141-2168.

Nascimento, A. M. and Vismari, L. F. and Molina, C. B. S. T. and Cugnasca, P. S. and Camargo, J. B. and d. Almeida, J. R. and Inam, R. and Fersman, E. and Marquezini, M. V. and Hata, A. Y. (2020) "A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety", IEEE Transactions on Intelligent Transportation Systems, Vol. 21, No. 12, pp. 4928-4946.

O'Reilly, C. A. and Chatman, J. (1986) "Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization on prosocial behavior", Journal of Applied Psychology, Vol. 71, No. 3, pp. 492-499.

Pichai, S. (2018) "Ai at google: our principles". Available <https://www.blog.google/technology/ai/ai-principles/>

Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", Nature Machine Intelligence, Vol. 1(5), pp. 206-215.

The IEEE Global Initiative, (2019) "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, first edition", Available <https://standards.ieee.org/content/ieee-standards/en/industryconnections/ec/autonomous-systems.html>

Turilli, M. and Floridi, L. (2009) "The ethics of information transparency", Ethics and Information Technology, Vol. 11, No. 2, pp. 105-112.

Vakkuri, V. and Kemell, K-K and Kultanen, J and Abrahamsson, P (2020) "The current state of industrial practice in artificial intelligence ethics", IEEE Software, Vol. 37, No. 4, pp. 50-57.

Vakkuri, V. and Kemell, K-K. and Abrahamsson, P. (2019) "Ai ethics in industry: research framework", CEUR Workshop Proceedings. RWTH Aachen University, Vol. 2505. Available <http://ceur-ws.org/Vol-2505/paper06.pdf>

Vakkuri, V. and Kemell, K-K. and Jantunen, M. and Halme, E. and Abrahamsson, P. (2021). "ECCOLA — A method for implementing ethically aligned AI systems", Journal of Systems and Software, Vol. 182, 111067.

Whittlestone, J. and Nyrup, R. Alexandrova, A. and Cave, S. (2019) "The role and limits of principles in ai ethics: Towards a focus on tensions", Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 195-200.

Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, Terah., Manyika, J. Ngo, H., Niebles J.C., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., and Perrault, R. (2022) "The AI Index 2022 Annual Report", AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University Available <https://aiindex.stanford.edu/report/>

APPENDIX 1 – AI Developer Questionnaire

General

1. What kind of software does your organization develop?
2. To whom are they developed to? / Who uses them? (customers / in-house projects)
3. How is AI involved in the software development? (AI / AI based solutions?)
4. What is your own role in the development?

Accountability

5. How much can you personally affect the functionalities of the AI solutions and the decisions made on them?
6. Who makes the final decisions concerning the development? (Such as what functionalities are good and what to choose to use?)
7. If the AI solution causes harm or damage to the user or third parties, who is responsible?
 - a. How much responsibility do you consider to be on you, based on your role in the organization
8. Are there other questions or issues on accountability that you have considered within your organization in relation to the development process or the end-products?

Predictability

9. How well do you consider the behavior of your AI solutions can be predicted beforehand? Could there be or has there been unexpected behavior to be noticed?
10. How do you prepare for this kind of unexpected behavior or possible malfunctions, and how do you react to them if they occur?
11. What is the level of acceptable risk or damage in case of malfunctions to the end-users or third parties?
12. How have you considered possible cases of misuse or abuse of your product? What could they be?

Transparency

13. How well the development process is being documented? For instance, can certain functions or decisions made during the development process be led back to the individuals behind them?
14. Are all the actions made by the AI solution transparent in a sense, that the logic behind the functions can be understood? (For example, the algorithms used and how they perform the reasoning – also during exceptions in functionalities.)
15. How well do the end-users know what the AI solution does and how it does it?

AI Ethics

16. Has your organization already faced some ethical issues or questions regarding AI development, and what have they been?
17. Do your organizational policies consider ethical aspects within AI development, and how?
18. How does the consideration of ethical aspects show in practice in the development process?
19. Do you consider taking ethical aspects into account in AI development would be beneficial to your organization? How?



III

THE CURRENT STATE OF INDUSTRIAL PRACTICE IN ARTIFICIAL INTELLIGENCE

by

Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, & Pekka Abrahamsson
2020

Magazine of IEEE Software vol. 37(4)

DOI 10.1109/MS.2020.2985621

Reproduced with kind permission by IEEE.

The Current State of Industrial Practice in Artificial Intelligence Ethics

Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, and Pekka Abrahamsson

Abstract—As Artificial Intelligence (AI) systems become increasingly widespread, we have begun to witness various failures highlighting issues in these systems. These incidents have sparked public discussion related to AI ethics and further accelerated the on-going academic discussion in the area. High-level guidelines and tools for managing AI ethics have been introduced to help industry organizations make more ethical AI systems, but we currently know little about the state of industrial practice. Have these guidelines been adopted by the software industry for developing AI solutions? Are these failures that make the news just the tip of the iceberg? We provide insights into the current state of practice by presenting the results of a survey of 211 software companies.

Index Terms—Artificial Intelligence, Ethics, Software Development

1 INTRODUCTION

VARIOUS technologies related to Artificial Intelligence (AI) have been at the top of the Gartner Hype Cycle for Emerging Technologies for years. Organizations from across industries are looking for ways to reap benefits from utilizing AI in different ways. During our recent visit to Slush, the world's leading startup and tech event with 25 000 attendees, we saw the booths of the AI startups down-right flooded, with lines forming on occasion. In general, the hype surrounding AI has long since reached a fever pitch.

As AI technologies become increasingly widespread, they start to exert a society-wide influence. Most of us interact with AI systems every day as consumers and customers, mostly without even realizing it. As the number of AI systems grows, so does the number of AI system failures we witness.

Various high-profile incidents that have made the global news have sparked public discussion on AI ethics. A growing number of voices, both from researchers and media, as well as governments, have called for more ethical AI systems in the wake of these failures. Sometimes these incidents are a result of simply not knowing better, as was the case with the Amazon recruitment AI that became biased against women [15]. Having been trained using past recruitment data, the AI saw mostly men hired, and learned thus that they were preferable hires.

On the other hand, sometimes these incidents are simply about intentional misconduct. While it was more of a lesson in relation to data handling in general, the case of Cambridge Analytica is one such example. Cambridge Analytica utilized data from the users of Facebook without

their consent to use for political advertising purposes [18]. Even though they were not the ones misusing the data themselves, it resulted in Facebook taking a publicity hit as well. With AI systems typically handling vast amounts of data, questions of data governance are important. The temptation to gather any and all data that may or may not be useful one day can be high when dealing with AI.

Yet, despite all the talk in the area recently, outside these incidents highlighting failures, we know little of the current state of practice of ethics in AI. Software engineering researchers have recently begun to understand more broadly how artificial intelligence and machine learning are changing the way the software is being developed [13]. Has the public and academic discussion in the area motivated smaller industry players to develop more ethical AI?

To the best of our knowledge, no surveys utilizing data from company respondents on the current state of practice in AI ethics exist. Existing surveys have relied on document data, for example from guidelines or project documents. Such surveys have been conducted on tools and methods [14], AI ethics guidelines [12], Artificial General Intelligence projects [3]. Various such document-based surveys also exist on the technical side of AI development, such as on machine and deep learning techniques and tools. Respondent data have been utilized in surveys on public opinions [8], as well as surveys on evaluating AI ethics guidelines [16], but not the state of practice AI ethics specifically.

To provide needed insight into the current state of practice in the industry, we present survey data from 211 software companies. Our data provides some context for this special issue by helping us understand where we currently are as an industry in terms of AI ethics. For practitioners, the data can also serve as a way to benchmark where your organization stands.

- V. Vakkuri is with the University of Jyväskylä, 40014, Jyväskylä, Finland. Email: ville.vakkuri@jyu.fi.
- K.K. Kemell is with the University of Jyväskylä, 40014, Jyväskylä, Finland. E-mail: kai-kristian.o.kemell@jyu.fi.
- J. Kultanen is with the University of Jyväskylä, 40014, Jyväskylä, Finland. E-mail: joni.kultanen@jyu.fi
- P. Abrahamsson is with the University of Jyväskylä, 40014, Jyväskylä, Finland. E-mail: pekka.abrahamsson@jyu.fi.

2 WHAT IS AI ETHICS?

Much of the research on AI ethics up until now has been predominantly theoretical and conceptual; valuable work aiming to define what is AI ethics (e.g. [5]). This has mostly been done by focusing on key principles [11]. These principles focus on specific categories of practical issues related to AI ethics, such as accountability. In our survey here, we focused on transparency, accountability, and responsibility, as well as predictability as a subset of transparency. Except for predictability, these three principles comprise the so-called ART principles for AI ethics.

Transparency is about understanding how the system works [7]. This is both about transparency of algorithms and data, the technical side of the system, but also transparency related to the development of the system [1], [10]. Transparency in terms of data and algorithms is related to the idea of explainable AI systems. Aside from being able to understand the system, we should also be able to understand who made the system into what it is today, and why.

Predictability can be considered a subset of transparency. It is about having a system that does what we expect it to do [2]. We certainly expect our autonomous thermostat in a smart home to keep the room temperatures in comfortable levels despite what it learns about our habits.

Accountability refers to liability issues related to stakeholders: who is liable to whom, for what, and why? To this end, laws and regulations can also be considered to fall under accountability. [1], [7], [10].

Responsibility is vaguer. It is about acting ethically or doing what we feel is the right thing. It is not tied to any specific idea of morality. [7]

Finally, Fairness, though not touched upon in our survey, is about equality in AI systems. Fairness has been discussed in terms of fairness in data or bias, as well as in terms of who benefits from AI systems [9], [10]. For example, do AI systems widen the societal gap between technologically skilled individuals and those less skilled?

Though these principles have focused on recently, various others have also been discussed. For example, the recently published European Union (EU) Ethics Guidelines for Trustworthy AI [1] considered trustworthiness to be the goal AI systems should aim for. The guidelines treat trustworthiness as a higher-level principle that principles such as transparency are required to achieve. Other principles include, for example, data-related ones such as privacy [11].

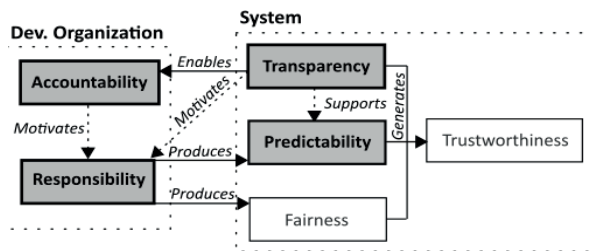


Fig. 1. Relations of key principles in AI ethics [19]

Fig. 1 portrays the relations between the principles we focus on. We have focused on the highlighted themes in our survey. Moreover, transparency is only considered in terms of data and algorithms in the figure.

Bringing this discussion and these principles into practice has been an ongoing challenge in the area [4]. For the most part, attempts at bridging this gap have been made by producing guidelines for AI ethics. The most prominent ones have been IEEE's Ethically Aligned Design (EAD) [10] guidelines. Other notable AI ethical guidelines include the European Union (EU) Ethics Guidelines for Trustworthy AI [1]. Overall, various guidelines have been produced by larger industry players, standardization organizations, academia, and governments alike [11]. These guidelines have various other principles than the ones we have chosen to focus on as well, such as data privacy, non-maleficence, and human well-being [11].

We currently have no knowledge of what impact these guidelines have had in the industry, however. Similarly, the current state of practice of AI ethics in general remains unknown, which is something we now shed some light on in this article.

3 AND WHAT IS ACTUALLY HAPPENING IN THE INDUSTRY?

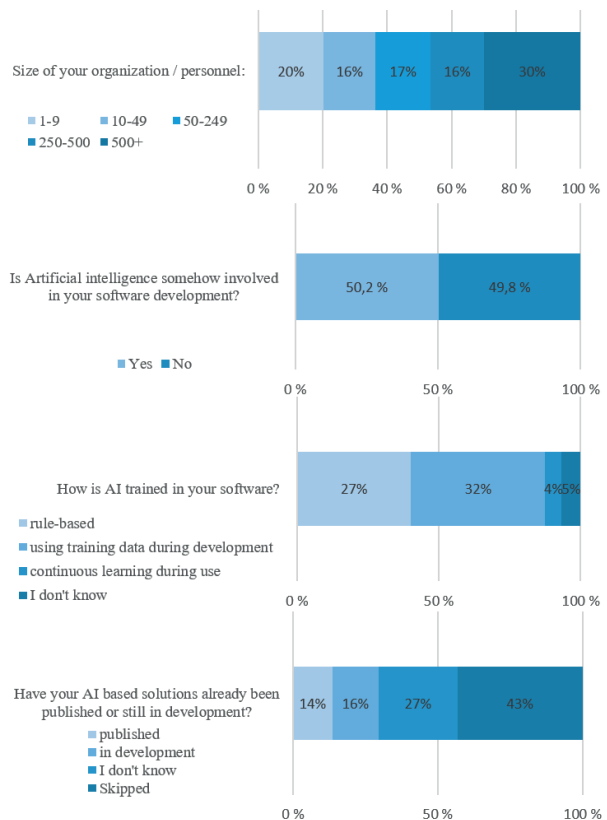


Fig. 2. Demographic description of the companies

Has the public and academic AI ethics discussion had an impact? Have these guidelines been adopted by the industry? To provide insights into the current state of practice in AI ethics, we conducted a survey, gathering responses from 211 software companies. The respondents were largely individuals capable of influencing the development in their companies: 68% of the respondents answered 4 to 7 in response to the question "how much can you personally affect the functionalities of the software developed in your organization and decisions made on them?".

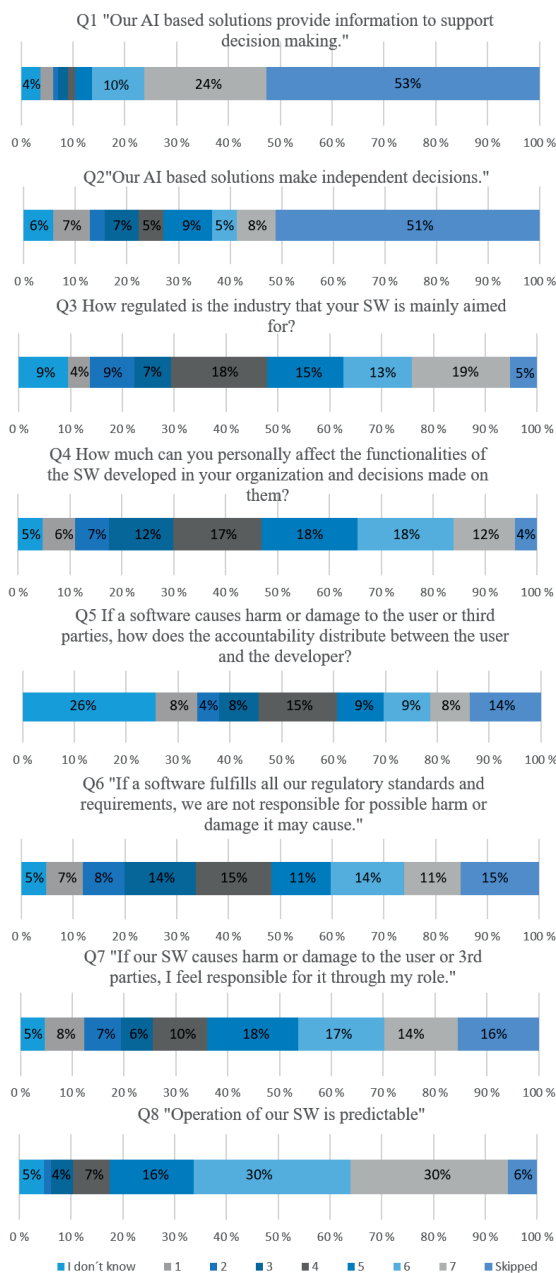


Fig. 3 The developer, liability, and responsibility. Scale from Strongly disagree to Strongly agree. Q3 Scale from Not at all to Very regulated. Q4 Scale from Not at all to Full authority

A little more than half of these companies (Fig. 2) were either developing or deploying an AI system. However, the responses did not notably differ between the companies that did not develop AI and the ones that did. We therefore included all responses. This is an interesting observation in and of itself: AI is currently simply treated as a feature in terms of ethics. This is in line with a study that argues that 90% of the activities we do in AI projects are the same as in any software project [17].

Overall, the responses indicated mixed maturity in implementing AI ethics. Responses to some of the questions directly indicated immaturity in relation to AI ethics, while some indicated some maturity. It would appear that the various AI ethics guidelines have not had a notable impact on practice, as has been suspected to be the case [4].

As many as 39% of the respondents skipped or answered "I don't know" to the liability question (Q5)(Fig. 3). This points to this being an unfamiliar theme, and thus an overlooked issue from an ethical viewpoint. Moreover, the qualitative responses from the companies also indicated that they did not tackle these issues even as well as their responses to the likert scale questions would have made it seem otherwise.

On the other hand, in response to some questions, such as predictability, the companies indicated more concern towards AI ethics related issues. For example, half of the organizations (48%) had a fallback plan for irregularities. Many respondents nonetheless noted that they did not have a fallback plan in place for unexpected system behavior in place, or that they did not know whether they had one (Fig. 4). Interestingly, most organizations (51%), felt their system could not be misused.

The respondents felt that they could influence the development of the system(s) highly, but still outsourced responsibility to the users when asked whether the developer or user was responsible (Fig. 3). 36% of the ones that answered (Fig. 3) considered meeting mandatory regulatory standards sufficient in terms of responsibility; past that it was up to the (end-)user to stay safe. Aside from the responsibility of their company, 49% of the respondents (aside from the 16% who skipped the question) felt personally responsible for any harm caused by their software, even if they largely didn't know that who was ultimately the one responsible.

Meeting the mandatory regulatory standards was also considered sufficient in terms of documentation by 43% of the respondents that answered (Fig. 4). On the other hand, 26% simply reported documentation being scarce or there being no documentation at all. The idea of being able to trace decisions back to individuals which is often discussed with accountability was reportedly achieved by 43% of the companies. However, the qualitative answers of many these companies made us doubt whether they really did address accountability to this extent with their work practices.

Their responses to documentation also somewhat conflicted with how the companies considered transparency important (Fig. 4). Transparency seemed to not be considered in terms of transparency of systems development.

Moreover, transparency in terms of data and algorithms

was mostly considered from the point of view of the development team and to some extent from the point of view of the user. Few companies considered transparency to public authorities, with 19% simply answering “I don’t know” to the question regarding it as well. Transparency to public authorities is one topic of discussion in AI ethics [6].

Despite machine learning being associated with an increased unpredictability, the responses between the AI companies and other software companies did not notably differ. By far the most respondents felt that their systems

were predictable. Yet, 34% of the companies had also faced issues due to unexpected operations in the system, pointing to a possible contradiction.

As most of our respondents were either from Finnish or US companies, we also compared the data between these two locations. There were no notable geographic differences in the data. Primarily, the Finnish companies operated in more regulated industries, and consequently seemed to place more emphasis on adhering to industry regulations.

The Survey

We collected survey data from 249 respondents in 211 software companies, out of which 106 developed AI systems. All responses were included together in the figures, as we noticed during the analysis that the trends were very similar whether the companies developed AI. Indeed, the original idea of the survey was to compare how much more well-versed in AI ethics AI companies were compared to other software companies. Given the increasing ubiquitousness of AI systems, every software company is likely to soon to be involved with AI.

The survey featured three types of questions: (1) demographic questions (organization size, name etc.); (2) Likert scale questions; and finally (3) open-ended questions. In this article, we focus on the Likert questions, which are covered in their entirety.

The survey focused on some of the central principles in AI ethics in the past few years. Namely, we discussed issues related to transparency, accountability, responsibility, and predictability. We have discussed the meaning of these principles in the second section of this article. Furthermore, we discuss the research model in detail in another research article [19].

In the Likert scale questions, we asked the participants to evaluate the importance of principles such as transparency. They were also posed some practical questions, such as whether they had faced issues with unpredictability in their software.

We collected data from both multi-national organizations as well as ones locally based ones. Most companies were either US (53) or Finnish (111). The rest were from 18 other countries. Responses were collected either as F2F structured interviews or via an online survey. US based company responses were obtained by purchasing the SurveyMonkey Audience service. Interviews were conducted when possible in terms of scheduling. Most of the responses were collected F2F.

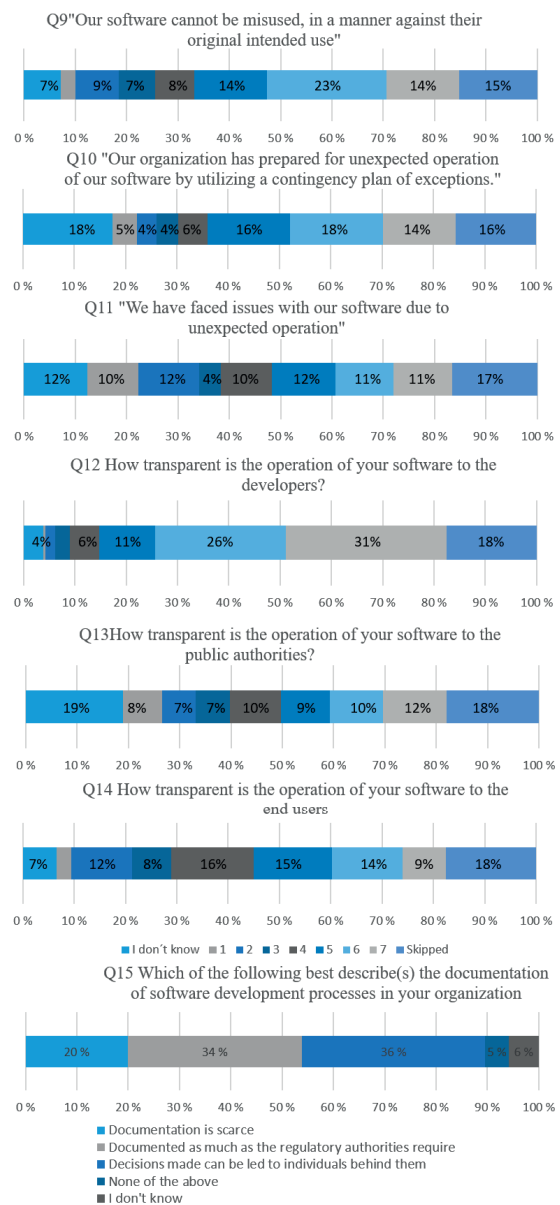


Fig. 4 Unexpected operation and transparency. Q9-Q11 Scale from Strongly disagree to Strongly agree. Q12-Q14 Scale from Not at all transparent to Fully transparent

4 WHAT SHOULD YOUR ORGANIZATION DO?

The data we collected points to AI ethics implementation still being in its infancy. This observation is mostly based on how the companies developing AI had largely similar responses to the survey as the ones not developing AI. AI seems to be considered just another feature, at least as far as the ethical side of things is considered.

AI ethics is closely tied to other emerging ethical mega trends. Ecological issues such as data center electricity consumption are tied to the larger trend of being environmentally conscious. Similarly, data privacy issues are highly related to AI systems as AI systems typically handle vast amounts of data [1]. Regulations such as the General Data Protection Regulation (GDPR) are already forcing industry organizations to act in terms of data handling and have highlighted the interest of governments to tackle AI ethical issues. As your users become increasingly conscious about privacy issues, being ethical in relation to data privacy for example can become a selling point.

If you wish to implement AI ethics, guidelines such as the IEEE EAD [10] ones, among others [11], can provide a starting point. However, utilizing these guidelines requires additional work from your organization as they do not come in the form of an off-the-shelf method. You need to first make them more practical for your developers, project teams, and product owners and customers.

On the other hand, various tools for implementing AI ethics also exist [14]. However, unlike guidelines, which focus on the bigger questions in the design and development of the system, the currently available tools focus on small portions of the development process. For example, various tools to manage unpredictability in machine learning exist, but they only cover a small subset of AI ethics. Project-level methods for software development do not yet exist for AI ethics [14]. This is something research in the area is currently working to tackle [12]. As a starting point, we recommend focusing on certain key practices rather than relying solely on values and principles.

Ultimately, AI projects are, at least currently, like any other software project. According to a study [17], 90% of what is done in AI projects is the same as in any software project. AI development is still software development, and for that reason, developers play an important role in AI ethics as well. Product owners' responsibility is to make sure that sprint backlog items have ethical user stories included. From the software development viewpoint, ethics in AI could be viewed as a non-functional requirement of an AI-based software system. When it becomes tangible, it becomes more manageable.

Finally, in implementing ethics in AI, there are some antipatterns to avoid:

- Outsourcing ethics, for example to a high-level ethics committee. Quality in software development cannot be outsourced and neither can ethics.
- Assuming ethics can be successfully implemented without doing so systematically. Leaving ethical issues for the developers to tackle is unlikely to work. With no methods to help them, developers are left to rely on their own capabilities.

- Appointing one individual to implement ethics. No one person can or should do it. AI ethics is a strategic matter. For example, the whole development team should be involved, going back to what we mentioned in the previous paragraph.

Currently, few laws and regulations that force the industry to implement AI ethics exist. However, with regulations such as the GDPR being drafted globally, preparing to tackle AI ethics issues already is insurance for the future. Much like how adding pipes to an already finished house is far more expensive than adding them while it is being built, ethical issues are much cheaper to tackle during design or even development than deployment.

Even without being forced to do so, devoting resources towards tackling ethical issues such as transparency can already be beneficial for your organization. When you increase the level of documentation in the name of transparency, you also support stakeholder communication. In this fashion, AI ethics can produce benefits. From the point of view of AI ethics, stakeholder communication is important particularly in relation to the general public and regulatory authorities. You can also learn valuable lessons from past incidents such as the two mentioned in the introduction.

As AI systems continue to become even more widespread, the number of such incidents, large and small, will only grow. The software industry is in a key position in preventing this from happening. Acting on AI ethics today will quickly pay back.

REFERENCES

- [1] AI HLEG (High-Level Expert Group on Artificial Intelligence), "Ethics guidelines for trustworthy AI," <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>. 2019.
- [2] M. Ananny and K. Crawford, "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media & Society*, vol. 20, no. 3, pp. 973–989. 2018.
- [3] S. Baum, "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy (November 12, 2017)". Global Catastrophic Risk Institute Working Paper 17-1. <http://dx.doi.org/10.2139/ssrn.3070741>. 2017.
- [4] M. Brent, "Principles Alone Cannot Guarantee Ethical AI". *Nature Machine Intelligence*. 2019.
- [5] J. Bryson and A. Winfield, "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems," *Computer*, vol. 50, no. 5, pp. 116–119. 2017.
- [6] V. Charisi, L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Loh, A. F. T. Winfield and R. Yampolskiy, "Towards Moral Autonomous Systems," Preprint arXiv:1703.04741. 2017.
- [7] V. Dignum, "Responsible Autonomy," Preprint arXiv:1706.02513. 2017.
- [8] European Commission, "Autonomous Systems - Report". Special Eurobarometer 427 / Wave EB82.4 - TNS Opinion & Social, https://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_427_en.pdf. 2015.
- [9] A. W. Flores, K. Bechtel and C. T. Lowenkamp, "False positives, false negatives, and false analyses: a rejoinder to Machine bias: there's software used across the country to predict future criminals, and it's biased against blacks," *Federal Probation*, vol. 80, no. 1, pp. 14–20. 2016.

- no. 2, 38-46. 2016.
- [10] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition," IEEE. Available at <<https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>>. 2019.
- [11] A. Jobin, M. Ienca and E. Vayena (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, no. 1, pp. 389-399. 2019.
- [12] J. Leikas, R. Koivisto and N. Gotcheva, "Ethical framework for designing autonomous intelligent systems". *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 5, no. 1. 2019.
- [13] L. E. Lwakatare, A. Raj, J. Bosch, H. H. Olsson and I. Crnkovic, "A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation," In *Proceedings of the International Conference on Agile Software Development*, pp. 227-243. Springer, Cham. 2019, May.
- [14] J. Morley, L. Floridi, L. Kinsey and A. Elhalal, "From what to how. an overview of AI ethics tools, methods and research to translate principles into practices," Preprint arXiv:1905.06876. 2019.
- [15] Reuters, "Amazon scraps secret AI recruiting tool that showed bias against women," <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>. 2017.
- [16] L. Rothenberger, B. Fabian and E. Arunov, "Relevance of Ethical Guidelines for Artificial Intelligence - A Survey and Evaluation". In *Proceedings of the 2019 European Conference on Information Systems (ECIS)*. 2019.
- [17] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Philips, D. Ebner, V. Chaudhary, M. Young, J. F. Crespo and D. Dennison, "Hidden technical debt in machine learning systems," *Advances in Neural Information Processing Systems*. 2015.
- [18] The New York Times, "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far". <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>. 2018.
- [19] V. Vakkuri, K-K. Kemell, J. Kultanen, M. Siponen and P. Abrahamsson, "Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study," Preprint arXiv:1906.07946. 2019

Ville Vakkuri is a PhD student in Information Systems at the University of Jyväskylä. Contact him at ville.vakkuri@jyu.fi.

Kai-Kristian Kemell is a PhD student in Information Systems at the University of Jyväskylä. Contact him at kai-kristian.o.kemell@jyu.fi.

Joni Kultanen is a MSc student in Information Systems Science at the University of Jyväskylä. Contact him at joni.m.kultanen@jyu.fi.

Pekka Abrahamsson is a full professor of Information Systems and Software Engineering in the Faculty of Information Technology at the University of Jyväskylä. Contact him at pekka.abrahamsson@jyu.fi.



IV

IMPLEMENTING AI ETHICS IN PRACTICE: AN EMPIRICAL EVALUATION OF THE RESOLVEDD STRATEGY

by

Ville Vakkuri & Kai-Kristian Kemell, 2019

International Conference on Software Business

DOI 10.1007/978-3-030-33742-1_21

Reproduced with kind permission by Springer.

Ethically Aligned Design: An empirical evaluation of the RESOLVEDD-strategy in Software and Systems development context

Ville Vakkuri, Kai-Kristian Kemell, Pekka Abrahamsson

Faculty of Information Technology, University of Jyväskylä
Jyväskylä, Finland

ville.vakkuri@jyu.fi, kai-kristian.o.kemell@jyu.fi, pekka.abrahamsson@jyu.fi

Abstract—Use of artificial intelligence (AI) in human contexts calls for ethical considerations for the design and development of AI-based systems. However, little knowledge currently exists on how to provide useful and tangible tools that could help software developers and designers implement ethical considerations into practice. In this paper, we empirically evaluate a method that enables ethically aligned design in a decision-making process. Though this method, titled the RESOLVEDD strategy, originates from the field of business ethics, it is being applied in other fields as well. We tested the RESOLVEDD strategy in a multiple case study of five student projects where the use of ethical tools was given as one of the design requirements. A key finding from the study indicates that simply the presence of an ethical tool has an effect on ethical consideration, creating more responsibility even in instances where the use of the tool is not intrinsically motivated.

Keywords—*artificial intelligence, ethics, design methods, ethical tool, RESOLVEDD, developer commitment*

1. INTRODUCTION

Artificial Intelligence and Autonomous Systems (AI/AS) are becoming increasingly ubiquitous. No longer are robots only found in factories, working highly repetitive conveyor belt tasks in closed environments. With autonomous vehicles entering the roads and AI systems filtering job applications out on the field, AI/AS are growing increasingly influential on a societal scale. It is practically impossible to opt out of using AI systems, with e.g. AI-based surveillance systems tracking you regardless of your consent. Similarly, due to the cyber-physical nature of many AI systems, their damage potential is not as narrow or predictable as that of conventional, purely digital software systems.

The pervasiveness of AI/AS systems forces us to analyze more profoundly under what type of ethical norms, rules and regulations AI systems should operate, and what kind of ethical standards should designers and developers hold when building these systems. As software engineers, developers are constantly making decisions when building systems. In doing so, they build their own values into the systems, which end up reflecting their views [1]. It is known that developers are not well-

informed and aware of ethics[2]. Combined with the current lack of tools to support ethical AI development, this results in a situation where developers do not have the necessary means to tackle potential ethical issues, or even recognize them during development. Ethical issues are often simplified or simply neglected, only to be re-discovered later during the operational life of these systems once the damage has already been done.

One solution to this problem is to offer the developers an ethical instrument or tool to support ethical considerations in design and value alignment. However, our understanding of what kind of methods should be used in introducing developers to ethics and how these proposed methods work in practice is lacking. Developers prefer simple and practical methods if they use methods at all [3]. Ultimately, ethics are currently not considered important by developers, and therefore tools for supporting ethical consideration should not be resource-intensive to adopt, lest developers potentially see them as a nuisance.

To begin tackling this issue, we tested an ethical tool from business ethics, the RESOLVEDD strategy, in the context of AI/AS design. We conducted a multiple case study of five different prototype projects where the use of ethical tool was given as one of the design requirements for the teams. The goal of this study is to better understand how the introduction of an ethical tool affects developers' ethical consideration in the design process and how the RESOLVEDD-strategy works in the given context.

A. Ethically Aligned Design

Ethically Aligned Design [4] refers to the involvement of decision-making in practice and ethical consideration in a the practice and design AI and autonomous systems and technologies. Involving ethical consideration into the context of software and interactive systems design has a history of more than 30 years. For example, Computer Ethics pioneer Bynum [5] introduced adapting human values in design before the rise of human values emphasizing the role of computer ethics. In response to ethical issues related to software and interactive systems development, Friedman [6] introduced a theoretically grounded Value Sensitive Design (VSD) approach and a

method for the design of technology that accounts for human values in a principled, structured, and comprehensive manner throughout the design process [6,7]. Over the years, VSD has been tailored into various different branches of methods. For example, Davis and Nathan [8] further developed VSD by reinforcing its philosophical foundations. Wynsberghe [9] presented the Care Centered Value-Sensitive Design (CCVDS) for care robotics. Miller, Friedman, and Jancke [10] proposed Value Dams and Flows method to address values-oriented design tradeoffs. As a result, VSD has become a domain-agnostic general model for consideration of human values in the design, implementation, use, and evaluation of interactive systems [8].

To better incorporate human values into the design process of AI systems, some AI-specific values have been proposed. For example, the importance of transparency in AI systems was emphasized by Bryson and Winfield [11]. Dignum [12] presented two more values in addition to transparency by presenting the ART principles (Accountability, Responsibility, Transparency) to guide ethical development of AI systems [12]. Finally, fairness of AI systems and freedom from machine bias have also gained a significant role as core values expected from AI systems [13].

To direct the discussion on aligning ethics with system design, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems was launched. The initiative was branded under a concept titled Ethically Aligned Design (EAD), a construct we discussed at the start of this section. The initiative aims to encourage practitioners to consider and prioritize ethics in the development of AI/AS. So far, the initiative has defined values and ethical principles that prioritize human well-being in a given cultural context. These guidelines and values have been published online, first in two versions for comments 2016-2018 and full release in 2019 EAD First Edition [4].

Arguably, the key audience of the EAD thinking should be the developers of the AI systems. AI development, much like conventional software development, is a cognitive activity [14] where humans play a significant role in deciding how the system behaves. Extant research has established that developers' interests are driven by work related concerns [15]. Concerns are the foundation of developer commitment development in his/her work. *Commitment* (discussed in detail in the next section) is important as it directs attention and helps in maintaining the chosen course of action [3, 15]. Should EAD practices become used by the developer, it should contribute to his work related concerns and help the developer to accomplish his or her tasks.

B. The RESOLVEDD- strategy

The step-by-step decision-making tool titled the RESOLVEDD strategy was first introduced by Pfeiffer and Forsberg [16]. Originally, the RESOLVEDD strategy was intended for teaching practical ethics to bachelor students. The method helps those who do not have prior knowledge of ethics or philosophy to evaluate ethical principles in practice. This aspect of the RESOLVEDD

strategy makes it particularly appealing for the field of Software Engineering (SE) where few curriculums have traditionally included studies in ethics or philosophy.

The RESOLVEDD strategy is based on professional ethics and approaches ethics from the point of view of personal ethical problems in work contexts. It is not connected to any particular ethics theory and it does not enforce any set of values on its would-be users. Instead, RESOLVEDD is intended to support its users in taking into account ethical issues and tackling them through their own set of values or through an ethics theory of their choice. [16]

The strategy is presented as a series of nine concrete steps portraying the rational ethical decision-making process. By using the method, one is able to justify and explain the decision-making process leading up to whatever actions were ultimately taken. It is intended to help its users understand the ethical issues present in their work and encourages them to address them in their way of choosing, though nonetheless without compromising ethical principles. Though it originates from the field of business ethics, the method can also be utilized for tackling ethical issues outside the field of business. [17]

The nine steps of the RESOLVEDD strategy can be seen as a process depiction in Figure 1. While utilizing resolved, however, these nine steps can be freely and flexibly modified to better suit each use context [16].

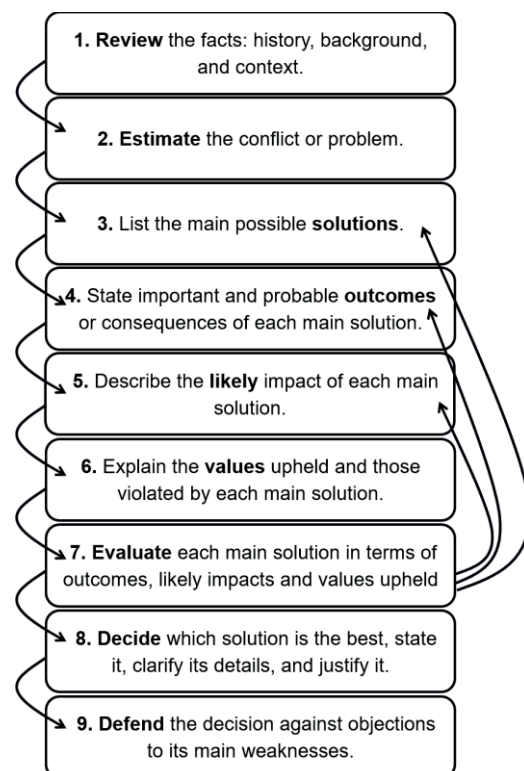


Figure 1. The Nine Steps of the RESOLVEDD Strategy

2. RESEARCH FRAMEWORK AND STUDY DESIGN

A. Research Framework

In addressing ethical principles in AI/AS design, *accountability*, *responsibility*, and *transparency* (the ART principles) have recently been considered to be key constructs [12]. This study uses these three constructs as a basis and attempts to identify their possible relations, as well as relations of other constructs that may be involved in the process (Figure 2). The ART constructs have a central role in determining design protocols that take into consideration the designer, the product, and the end-users [12]. While other principles have been proposed for the ethical design of AI systems (see e.g. [4]), we consider the ART constructs a good starting point for understanding the involvement of ethics in ICT projects.

Developers' interests are driven by work-related concerns [15]. From the point of view of the developers, an important question to pose is: why would the developer act responsibly and take into account ethical issues? To begin tackling this question, meaningfulness of taken actions has been shown to be important in explaining work-related behavior [18]. For this reason, we need to understand the relationship between meaningfulness and the meaning of an activity, as we argue next. We have established that in order for an action to become meaningful for a developer, they must understand the meaning of the task. Therefore, a task that may be perceived as time consuming, boring, or otherwise lacking in motivational elements, will still be executed because it plays a role in the developer's commitment behavior [15].

Commitment, accountability, responsibility and transparency can therefore be seen as a cycle with links (Figure 2). These links are explorative as little empirical data is currently available. We can hypothesize that by strengthening commitment to the RESOLVEDD strategy action, ethics will become implemented in the system. Ethics, as defined by EAD, is evidenced by increased in responsibility in design and clarity of accountability in order to help create more transparent culture in development of AI/AS. Transparent culture can likewise influence commitment, responsibility and accountability in design. In order to achieve this goal, the RESOLVEDD strategy should (1) support responsibility, responsible culture, (2) help people to make more meaningful decisions in their own work, and (3) take into consideration ethical principles such as accountability, privacy, autonomy, and fairness.

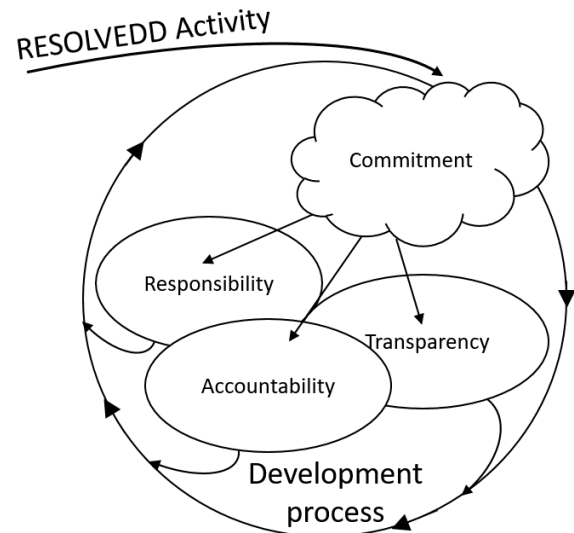


Figure 2. Framework for Ethically Aligned Design

1) Commitment

Commitment is the psychological bond between a person and an object (of the commitment) [19]. This bond is characterized by focus, strength, and type. The focus of the commitment can be work-related or personal. At least four types of commitment can be found in extant literature: affective, normative, continuance and instrumental commitment.

Affective commitment refers to a situation where a person truly believes in the focus of her commitment. This is indicated with phrases such as "I really want to do this". Affective commitment is the type of commitment that we typically refer to when we think of the construct. It is by definition a strong bond and thus difficult to influence from the outside.

Normative commitment refers to a situation where a person feels obliged to do something because of internal or external pressure. For this reason, in many cases, promises made in public are more binding than those that are kept to oneself.

Continuance commitment is the third type of commitment form. It is also known as escalation of commitment in the field of management. Continuance commitment refers to a situation where you have continued some activity for so long that the costs of aborting it are higher than those of completing the effort.

Finally, *instrumental* commitment is the most typical form of commitment and is often utilizing when motivating people to perform at a work place. The intent of the incentives is to tie the person to the commitment object (e.g. the objective of a project). [15]

Understanding how a person may be committed to a certain object is related to understanding what key concerns in that individual's work life. This can be modeled with a commitment net. A commitment net is a web of concerns and their corresponding actions. It is a tool for making sense of what the priorities of an organization, a project, and an individual are. [15]

Literature [15] has established that a concern drives the behavior. In this study, we seek to understand the commitment of the developers when they were using the RESOLVEDD-strategy to better understand the results of their designs.

2) *Transparency*

In the ART model, Dignum [12] presents a rather narrow view of transparency, focusing on the transparency of the algorithms and data used, as well as their provenance and their dynamics. We argue that transparency has a more significant role in determining ethical design. As Turilli & Floridi [20] state, transparency acts as a pro-ethical circumstance that makes it possible to implement ethical principles into the design process.

The construct of transparency is used when referring to the visibility of information from the design and development process, as well as from the product itself. There are thus two types of transparency: transparency of systems, and transparency of systems development. The former refers to understanding how the systems are designed and why they act in certain ways in certain situations. The latter, on the other hand, refers to understanding what decisions were made during the development process, and why.

Transparency has been considered to be crucial for the ethical design and use of AI/AS since it provides a simple and objective way of understanding what an AI/AS is doing and why. Processes, products, values as well as design practices should be transparent in order to help to enhance human well-being and acceptance of technology [4, 12]. Without transparency in the actions of oneself or the system being developed, it is impossible to assess the justifications for the actions or the ethical principles behind them. E.g. if an autonomous vehicle crashes and we cannot understand why, ethical assessment of the incident and the decisions leading up to it is impossible as well. Systems need to be transparent so that the reasons behind unwanted results can be understood [4].

3) *Accountability*

To prevent misuse and to support EAD, accountability structures are needed [4]. In the ART model, accountability is seen as demand for the derivability of who is accountable for the decisions made by system and its algorithms. In their more recent work, Dignum [12] defines accountability to refer to the explanation and justification of one's decisions and one's actions to the relevant stakeholders.

In order to consider someone accountable, there needs to be transparency in information, data, and design as discussed in the preceding sub-section. Therefore, transparency is required for accountability to be achievable. To achieve accountability, developers should be aware of the accountable matters that they are involved with and that are present in their systems.

In context of this study, accountability is used not only in the context of systems, but also in a more general sense. We consider, for example, how various accountability issues (legal, social) were taken into consideration during the design process.

4) *Responsibility*

Whereas accountability is related to the connection between one's decisions or actions and the stakeholders of the system, responsibility is an internal process. In order to act responsibly, one needs to understand the meaning of their action. In the ART model, responsibility is related to the idea of the chain of responsibility, even when there is no human agent as a direct cause of action there must be a linking chain to the responsible stakeholder. Therefore, artificial intelligence is an actor with a role in the chain of responsibility.

Responsibility in the context of this study connects the designer to the outside world, to others as stakeholders for example. In order to be responsible, one has to make weigh their own actions and to consciously evaluate their choices. E.g. one very simple way of considering responsibility would be to ask oneself "would I be fine with using my own system?".

B. Study design

The RESOLVEDD strategy was empirically evaluated using a case study research method [21]. More specifically, we conducted case studies of five student projects that all utilized the RESOLVEDD strategy. Yin [30] explains that the use of multiple case study makes it possible to have multiple data sources with rich in-depth investigations that would not be possible with a survey. This method also allowed the analysis within each case and across the cases to validate the observations by cross-referencing [21].

The study was conducted in an Information Systems (IS) course at the University of Jyväskylä. Bachelor level students were introduced to the RESOLVEDD strategy as a part of the system design and development methods. In the course, the students were given the task of developing a concept and prototype of a futuristic innovation that could be possible in the near future, but which was not considered plausible with current technologies. The projects were carried out as a group work in five groups of 4-5 students. Choosing from a list., the students had to decide which technology they would want to utilize as part of their solution. For example, the students could make solutions that utilized Augmented Reality (AR), AI, or more specific technologies such as the Raspberry Pi computer.

3. FINDINGS

The findings from the analysis of the empirical data are reported here as topic-related Primary Empirical Conclusions (PEC). In total 5 PECs were formulated in the analysis. This section is structured into four sub-sections according to the research framework discussed in the preceding section.

A. Commitment to Ethically Aligned Design

All five teams had rather critical sentiments towards dealing with ethical issues or using ethical tool as a part of their product design. Using an ethical tool was perceived as something completely novel to them, and they did not seemingly place value on considering the ethical aspects

on their project. This was despite of the fact that the employed method is focused on helping its users detect ethical issues. When considering commitment to EAD, it is important to understand what the true concerns of the developers are. In this case, the teams were more concerned about the usefulness and viability of their product than its ethical aspects.

PEC 1: While normative commitment to the use of Ethically Aligned Design brings immediate results, it will cease to exist when the external pressure is taken away. The RESOLVEDD strategy needs adaptation in application context. In practice, group discussions were seen effective in addressing the ethical issues.

B. Transparency in design

Even though the teams were not affectively committed to using the ethical tool in their design process, they were required to follow the steps of the RESOLVEDD strategy and to produce documents that increased the transparency and the visibility to the teams' decision-making process. Teams adapted the RESOLVEDD strategy to fit their needs in order to carry out ethical thinking. The external pressure to use a specific method did not please the teams. Nonetheless, the necessitated use of the RESOLVEDD strategy method did increase transparency and ensured that the ethical considerations of the teams were documented for later use. The teams remained skeptical, however, whether their documentation would be beneficial.

PEC2: When the RESOLVEDD-strategy is followed step-by-step a paper trail is born where each decisions made and the respective justification can be found. This produces transparency in the design process, but it does not promote transparency at the product layer.

C. Accountability in design

The question of accountability divided the teams. It was not clear to the teams who can be held accountable for the design. Teams defended their position (not being accountable) by arguing that the systems are only concepts and prototypes. They outsourced the issue of accountability to the end user, or they were not able to explain how it is managed from the legal or social viewpoints. The teams' lack of knowledge on accountability issues plays an important role.

PEC3: The RESOLVEDD-strategy does not deliver accountability.

D. Responsibility in design

Expecting the teams to engage in EAD and supporting their engagement in EAD by introducing an ethical tool made it possible to talk about the ethical issues related to their current projects. Our introduction to the RESOLVEDD strategy could have been improved.

PEC 4: Requiring Ethically Aligned Design activated reflections on the developers' own sense of responsibility

We also found that the teams were not keen on using the method, nor were they satisfied with the results they obtained by doing so. External pressure for the use of the tool nonetheless created tangible results, promoted EAD,

and even supported the developers' sense of responsibility

PEC 5: The mere presence of an ethical tool has an effect on ethical consideration creating more responsibility even when it the use of the method is not voluntary.

4. DISCUSSION AND CONCLUSIONS

In this study, we have evaluated empirically the RESOLVEDD strategy for ethical decision-making through an exploratory, multiple case-study of five AI/AS projects. The study subjects were students and thus formed a limitation of the study that needs to be considered. We find that the limitation is not so relevant since Höst et al [22] finds that the differences between students and professionals is minor and not statistically significant. In fact, he recommends the use of students in software engineering studies. Runeson [23] finds similar improvement trends between undergraduate, graduate and professional study groups. For a novel topic in the field (such as EAD in our case), the students provide an excellent platform for an empirical evaluation, method development and experimentation. Future studies should consider case studies in industrial settings.

We found that while normative pressure to the use of Ethically Aligned Design brings immediate results, it will cease to exist when the external pressure is taken away (PEC1). RESOLVEDD increased transparency in the design process (PEC2) but it does not deliver accountability (PEC3). Requiring Ethically Aligned Design from the developers increased their sense of responsibility (PEC4). As a concluding finding it can be stated that the mere presence of an ethical tool has an effect on the ethical consideration exerted by developers, creating more responsibility even when the use of the method is not voluntary (PEC5).

The research framework formed in this study also has practical implications by making the level of ethically aligned design evaluable. We have shown, initially, that while it is possible to introduce EAD by force, results will not sustain over time. The RESOLVEDD strategy needs to be adjusted in practice. One important adjustment done by our case teams was the introduction of group discussions as the primary means to do EAD in practice. Thus, a possible avenue for tailoring is to identify what are the practices that actually lead to favorable outcomes increasing transparency, responsibility and accountability.

REFERENCES

- [1] C. Allen, W. Wallach and I. Smit, "Why Machine Ethics?" IEEE Intelligent Systems, vol. 21, (4), pp. 12-17, 2006 doi: 10.1109/MIS.2006.83.
- [2] A. McNamara, J. Smith and E. Murphy-Hill, "Does ACM's code of ethics change ethical decision making in software development?", Proceedings of the 2018 26th ACM ESEC/FSE, pp. 729-733, 2018. doi:10.1145/3236024.3264833
- [3] P. Abrahamsson and N. Iivari, "Commitment in software process improvement - in search of the process," Proceedings of the 35th HICSS, pp. 3239-3248, 2002. doi: 10.1109/HICSS.2002.994403.
- [4] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human

Well-being with Autonomous and Intelligent Systems, First Edition. IEEE. 2019.

- [5] T. Bynum, "Flourishing Ethics," *Ethics and Information Technology*, vol. 8, (4), pp. 157-173, 2006. doi: 10.1007/s10676-006-9107-1
- [6] B. Friedman, "Value-sensitive Design," *Interactions*, vol. 3, (6), pp. 16-23, 1996. doi: 10.1145/242485.242493.
- [7] B. Friedman, P. H. Kahn, A. Borning and A. Huldtgren, "Value Sensitive Design and Information Systems," in *Early engagement and new technologies: Opening up the laboratory. Philosophy of Engineering and Technology*, vol 16, N. Doorn et al. Eds. Dordrecht, Springer 2013. doi: 10.1007/978-94-007-7844-3
- [8] J. Davis and L. P. Nathan, "Value sensitive design: Applications, adaptations, and critiques," in *Handbook of Ethics, Values, and Technological Design*, J. van den Hoven et al. Eds. Dordrecht, Springer 2015, pp. 11-40 doi: 10.1007/978-94-007-6970-0_3
- [9] A. Wynsberghe, "Designing Robots for Care: Care Centered Value-Sensitive Design," *Sci. Eng. Ethics*, vol. 19, (2), pp. 407-433, 2013. doi: 10.1007/s11948-011-9343-6
- [10] J. Miller, B. Friedman and G. Jancke, "Value tensions in design: The value sensitive design, development, and appropriation of a corporation's," *Proceedings of the ACM Group 2007*, pp. 281-290, 2007. doi: 10.1145/1316624.1316668.
- [11] J. Bryson and A. Winfield, "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems," *Computer*, vol. 50, (5), pp. 116-119, 2017. doi: 10.1109/MC.2017.154
- [12] V. Dignum, "Responsible autonomy," *arXiv Preprint arXiv:1706.02513*, 2017.
- [13] A. W. Flores, K. Bechtel and C. T. Lowenkamp, "False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals, and it's biased against blacks"," *Federal Probation*, vol. 80, (2), pp. 38, 2016.
- [14] D. Graziotin, X. Wang and P. Abrahamsson, "Are happy developers more productive?" in *Product-Focused Software Process Improvement*, pp. 50-64, 2013. doi: 10.1007/978-3-642-39259-7_7
- [15] P. Abrahamsson, "Commitment Nets in Software Process Improvement," *Annals of Software Engineering*, vol. 14, (1), pp. 407-438, 2002. doi: 10.20526329708".
- [16] R. S. Pfeiffer and R. P. Forsberg, *Ethics on the Job: Cases and Strategies*. Wadsworth Publishing Company, 1993.
- [17] C. Johansen, "Teaching the ethics of biology," *The American Biology Teacher*, vol. 62, (5), pp. 352-358, 2000.
- [18] N. E. Bowie, "A Kantian Theory of Meaningful Work," *J. Bus. Ethics*, vol. 17, (9), pp. 1083-1092, 1998.
- [19] P. Abrahamsson, "Rethinking the Concept of Commitment in Software Process Improvement," *Scandinavian Journal of Information Systems*, vol. 13, (1), 2001.
- [20] M. Turilli and L. Floridi, "The ethics of information transparency," *Ethics and Information Technology*, vol. 11, (2), pp. 105-112, 2009. doi: 10.1007/s10676-009-9187-9
- [21] R. K. Yin, *Qualitative Research from Start to Finish*, Second edition. New York, Guilford Press, 2016.
- [22] M. Höst, B. Regnell and C. Wohlin, "Using Students as Subjects A Comparative Study of Students and Professionals in Lead-Time Impact Assessment," *Empirical Software Engineering*, vol. 5, (3), pp. 201-214, 2000. doi 10.26586415054".
- [23] P. Runeson, "Using students as experiment subjects – an analysis on graduate and freshmen student data," *Proceedings of the 7th International Conference on EASE*. pp. 95-102 2003.



V

**ECCOLA—A METHOD FOR IMPLEMENTING ETHICALLY
ALIGNED AI SYSTEMS**

by

Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen., Erika Halme &
Pekka Abrahamsson 2021

Journal of Systems and Software vol. 182

DOI 10.1016/j.jss.2021.111067

Reproduced with kind permission by Elsevier.



Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jssECCOLA – A method for implementing ethically aligned AI systems[☆]Ville Vakkuri^{*}, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, Pekka Abrahamsson

University of Jyväskylä, PO Box 35, FI 40014, Jyväskylä, Finland



ARTICLE INFO

Article history:

Received 10 January 2021
 Received in revised form 4 May 2021
 Accepted 17 August 2021
 Available online 2 September 2021

Keywords:

Artificial intelligence
 AI ethics
 Ethics
 Implementing
 Method

ABSTRACT

Artificial Intelligence (AI) systems are becoming increasingly widespread and exert a growing influence on society at large. The growing impact of these systems has also highlighted potential issues that may arise from their utilization, such as data privacy issues, resulting in calls for ethical AI systems. Yet, *how* to develop ethical AI systems remains an important question in the area. How should the principles and values be converted into requirements for these systems, and what should developers and the organizations developing these systems *do*? To further bridge this gap in the area, in this paper, we present a method for implementing AI ethics: ECCOLA. Following a cyclical action research approach, ECCOLA has been iteratively developed over the course of multiple years, in collaboration with both researchers and practitioners.

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As Artificial Intelligence (AI) technology is developed with speeding progress, these systems become increasingly widespread and exert a growing impact on society. This has led to us witnessing a number of AI system failures, many of which have made global headlines and resulted in public backlash. Occasionally, these failures have served to highlight some of the various potential ethical issues associated with AI systems, in cases where these systems are found to, for example, exercise unfair bias or act in socially unacceptable ways. Some such famous incidents occurred when AI-based systems have endorsed or exercised unethical behavior such as gender discrimination¹ or racism.² Especially issues related to privacy, in cases like facial recognition technology, have become a prominent topic among the general public, as well as for policymakers.³

Though these incidents have resulted in collective learning experiences, the systems we developed are still far from being problem-free. Ethical issues persist, and more arise as the level of

sophistication of AI-related technologies rises. Aside from the obvious physical damage potential of systems such as autonomous vehicles, many areas of AI systems and their development are ripe with ethical issues without universal answers, starting from well-known topics such as data handling and extending to complex societal impacts of future systems (advanced general AI, etc.) currently still unattainable without further progress in the area.

The discussion on the field of AI ethics has soared in activity in the past decade following AI-related technological progress, resulting in the birth of some key principles that are now widely acknowledged as central issues in AI ethics. These principles cover a wide range of subjects, such as a demand for AI systems to be explainable (Rudin, 2019) and aligned with human rights and well-being (IEEE Global Initiative, 2019). The problem thus far has been transferring this discussion into practice, i.e., how to actually influence the development of these systems.

So far, this has mostly been carried out either via guidelines or laws and regulations. Guidelines have been devised by various parties, such as companies (e.g., Google (Pichai, 2018)), governments (e.g., EU (HLEG, 2019)) and standardization organizations (e.g., IEEE (IEEE Global Initiative, 2019)). Despite their ubiquity, guidelines alone have been lacking in actionability. Developers struggle to implement abstract ethical guidelines into the development process (Vakkuri et al., 2020; McNamara et al., 2018). There may be no consequences for deviating from codes of ethics or using them mainly as a marketing strategy, and there is no guarantee that ethics guidelines will affect the actual decision-making of developers (Hagendorff, 2020).

Methods and practices in the area remain highly technical, focusing on, e.g., specific machine learning issues (Morley et al.,

[☆] Editor: Raffaella Mirandola.

^{*} Corresponding author.

E-mail addresses: ville.vakkuri@jyu.fi (V. Vakkuri),

kai-kristian.o.kemell@jyu.fi (K.-K. Kemell), marianna.s.p.jantunen@jyu.fi (M. Jantunen), erika.a.halme@jyu.fi (E. Halme), pekka.abrahamsson@jyu.fi (P. Abrahamsson).

¹ <https://www.nytimes.com/2019/11/10/business/apple-credit-card-investigation.html>.

² <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>.

³ <https://www.bbc.com/news/technology-48276660>.

2019). While certainly useful in their specific contexts, these types of tools do not help companies in the design and development process as a whole. For example, tools for machine learning, though key in AI systems, do not help companies make decisions regarding the system and its future usage context in the big picture. Thus, other approaches such as development methods for ethical AI are still required to bridge this gap between research and practice in the area.

In this paper, we present our work on an AI ethics method: ECCOLA. ECCOLA is a sprint-by-sprint process designed to facilitate ethical thinking in AI and autonomous systems development, and designed to be used together with existing methods. It takes on the form of a deck of 21 cards, split into 8 AI ethics themes (e.g. transparency). While designing ECCOLA, we had three goals for it: (1) to help create awareness of AI ethics and its importance, (2) to make a modular method suitable for a wide variety of SE contexts, and (3) to make ECCOLA suitable for agile development, while also helping make ethics a part of agile development in general. Overall, ECCOLA is intended to help organizations implement AI ethics in practice, in an actionable manner.

ECCOLA has been developed iteratively over the past three years through empirical use and data resulting from it, with each iteration improving the method. In doing so, we have followed a Cyclical Action Research approach (based on [Susman and Evered \(1978\)](#) and [Davison et al. \(2004\)](#)). So far, there have been 6 stages in this process. ECCOLA has been used and evaluated in student, industry, and academic contexts (e.g. conference workshops), with the evaluation and usage shifting towards the industry over time. This article extends an existing paper presenting an earlier version of ECCOLA published in the proceedings of DSD/SEAA 2020 ([Vakkuri et al., 2020](#)). Since then, we have focused on seeing how companies utilize ECCOLA in practice while continuing to develop ECCOLA in collaboration with other researchers.

The rest of this paper is structured as follows. The second section discusses the theoretical background of ECCOLA. The third section presents the ECCOLA method itself. In the fourth section we introduce our research approach. In the fifth section we discuss how ECCOLA was iteratively developed. In the sixth section we discuss the implications of ECCOLA. In the seventh section we discuss threats to validity. The eighth and final conclusions section concludes the paper.

2. Theoretical background

This section is split into four subsections. In the first one, we provide an overview of the current state of AI ethics in research. In the second one, we focus on the state of the practical implementation of AI ethics, discussing the methods and other tools that currently exist to help practitioners implement it. In the third we discuss Value Sensitive Design to further position this method using existing literature. In the fourth and final one, we discuss the Essence Theory of Software Engineering, and specifically the idea of essentializing software engineering practices, as this is an approach we have utilized in devising ECCOLA.

2.1. AI ethics

AI ethics is a long-standing area of research. In the past, much of the debate has focused on hypothetical future scenarios that would result from technological progress. However, as these hypothetical future scenarios start to become reality following said progress, which to many has been faster than anticipated, the field has become increasingly active.

Much of the research in the area has focused on theory, and specifically on defining AI ethics by highlighting key ethical issues in AI systems. This discussion has focused on principles.

Many have been proposed and discussed, and by now, some have become largely agreed-upon ([Jobin et al., 2019](#)). Based on an analysis of the numerous AI ethics guidelines that now exist, Jobin et al. ([Morley et al., 2019](#)) listed the key principles that could be considered central based on how often they appear in these guidelines: “transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity”.

To provide an example of the type of research that has been conducted on these principles, we can look at transparency. Transparency ([Dignum, 2017](#)) is widely considered one of the central AI ethical principles. Transparency is about understanding AI systems, how they work, and how they were developed ([Dignum, 2017](#); [Ananny and Crawford, 2018](#)). It has been argued to be the very foundation of AI ethics: If we cannot understand how the systems work, we cannot make them ethical either ([Turilli and Floridi, 2009](#)). The discussion on transparency has, aside from defining what it is, focused on how to achieve it. For example, [Ananny and Crawford \(2018\)](#) discussed the limitations of the idea of transparency in relation to the complexity brought on by machine learning. Is being able to see inside the system really enough or even helpful? For example, transparency is featured as a key principle in the high-profile guidelines of EU ([HLEG, 2019](#)) and IEEE ([IEEE Global Initiative, 2019](#)).

Principles are but one way of categorizing the discussion in the area. The discussion in the area is ultimately about bringing attention to potential ethical issues in AI, with or without pinning them under a specific principle. Privacy issues, for example, have been one prominent topic of discussion both in academia and the media following various practical examples of (ethical) AI system failures. For example, privacy issues have been discussed in relation to data handling, and technologies such as facial recognition. Privacy issues are hardly a topic of discussion unique to the field of AI ethics either. Data issues such as bad data have also been discussed in relation to racial bias, which falls under the principle of fairness.

Guidelines have been utilized as a way of bridging the gap between research and practice, with the purpose to distill the discussion in the area into tools in the form of guidelines. However, past research has shown that guidelines are rarely effective in software engineering. [McNamara et al. \(2018\)](#) studied the impact the ACM Code of Ethics⁴ had had on practice in the area, finding little to none. This seems to also be the case in AI ethics: in a recent paper ([Vakkuri et al., 2020](#)), we studied the current state of practice in AI ethics and found that the principles present in literature are not actively tackled out on the field. Moreover, we found that AI development endeavors did not differ from generic development endeavors in this regard, with companies developing AI no more focused on tackling them differently than any other software company. This gap, and the issues with guidelines, are also acknowledged by Johnson & Smith in their gap analysis ([Johnson and Smith, 2021](#)).

The state of affairs as presented here, underlines a need for more actionable tools for implementing AI ethics in practice. In the context of software engineering, we therefore turn to methods; ways of taking action that direct how work is carried out ([Jacobson et al., 2012](#)). As software engineering in any mature organization is carried out using some method, out-of-the-box ones or in-house ones, incorporating AI ethics as a part of these methods would be a goal to strive for. In this next subsection, we look at methods in the area.

⁴ <https://www.acm.org/code-of-ethics>.

2.2. Methods in AI ethics

There are already various methods and tools for implementing AI ethics, as highlighted by Morley et al. (2019) in their systematic review of the field. The study consists of largely tools for the technical side of AI system development, such as tools for machine learning. The study by Morley et al. reviews a collection of tools or methods that are utilized by various companies and organizations for implementing ethics in AI development, and a typology based on ethical principles is used to analyze the results.

The review by Morley et al. brought certain challenges to light regarding AI ethics tools; the study showed that some of the researched tools are immature, and there is an "uneven distribution of effort across the 'Applied AI Ethics' typology" (Morley et al., 2019). Morley et al. believe that creating ethical machine learning technologies is realistically possible, but efforts have so far been focused on the "what", and not the "how" of AI ethics (Morley et al., 2019). The debate has been focusing on the topic on ethical principles, instead of applying them in practice. They suggest that turning ethical principles into design protocols will require increased coordination, and patience to tolerate a slow progression of turning theory into practice, with mistakes along the way (Morley et al., 2019).

On the other hand, we are not currently aware of any method focusing on the higher-level design and development decisions surrounding AI systems. Guidelines have been devised for this purpose but seem to remain impractical given their seeming lack of adoption out on the field (Vakkuri et al., 2020). The field remains active, for example, Leikas et al. (2019) recently proposed an "Ethical Framework for Designing Autonomous Intelligent Systems" and an AI ethics MOOC at the Helsinki University has devoted a chapter to AI ethics in practice (Rusanen et al., 2021).

Aside from AI ethics methods and tools, some ethical tools from other fields do exist that could potentially be used to design ethical AI systems. One example of such a tool is the RESOLVEDD method from the field of business ethics. We have studied the suitability of this particular method for the AI ethics context in the past, with our results suggesting that dedicated methods specifically devised for implementing AI ethics would be more beneficial (Vakkuri and Kemell, 2019). Additionally, we feel that Value Sensitive Design (VSD) is another approach worth mentioning in this context, even though it is not specific to AI ethics. Due to its prominence in existing research (specifically in Information Systems (IS)), we discuss it separately in the following subsection.

2.3. Values in value sensitive design

In addition to looking at the field of AI ethics from the point of view of SE, we feel that a brief look at ethics and value consideration discussion from IS is in order as well to better position ECCOLA. In particular, Value Sensitive Design (VSD) is a prominent approach that has been utilized out on the field. However, as VSD is not specific to AI ethics, we have separated it from the preceding subsection.

VSD can be traced back to the 1990s when the HCI (Human-Computer Interaction) community took a stand on value-oriented design in IS research (Shilton, 2018). The context-specific nature of ethical issues has been acknowledged in VSD as well, with Friedman remarking that different individuals and people have different ideas of ethics and values (Friedman et al., 2013). In the context of Information Systems Design (ISD), Friedman et al. (2008) proposed 13 values: Human Welfare, Ownership and Property, Privacy, Freedom from Bias, Universal Usability, Trust, Autonomy, Informed Consent, Accountability, Courtesy, Identity, Calmness, and Environmental Sustainability. Looking at this list of

values, there is a reasonable amount of overlap with the common AI ethics principles summarized by Jobin et al. (2019) that we discussed in Section 2.1 above.

Even outside the context of AI ethics, integrating ethical considerations into practice in software engineering (SE) is a recurring challenge. For example, the ACM/IEEE Software Engineering Code of Ethics and Professional Practice, while in many ways useful according to Biffi et al. (2006), has also been difficult to integrate into traditional SE. Indeed, a more recent study (McNamara et al., 2018) has also argued that the ACM Ethical Guidelines (Gottbarn et al., 2018) have not changed the way developers work.

Value Sensitive Design (VSD) is a methodology meant to encourage designers to consider ethics and values in the design process, and is "primarily concerned with values that center on human well-being, human dignity, justice, welfare, and human rights". VSD Lab (2021). VSD is at the cross-section of four fields closely related to HCI, namely Computer Ethics, Social Informatics, Participatory Design, and Computer-Supported Cooperative Work. Friedman and Kahn set up a seven principle composite that the VSD is based on, and one of the main principles is that VSD is a proactive methodology (Friedman et al., 2002). VSD encompasses 14 methods for incorporating value consideration into the design process (Davis and Nathan, 2015).

VSD has seen some success out on the field as well, with multinationals such as Intel and Microsoft utilizing it in some projects (Manders-Huits, 2011). Overall, its use has been documented in a wide variety of projects. Perhaps the most notable VSD method in terms of industry utilization has been the Tripartite Method, which is used to involve value consideration into the design process (Winkler and Spiekermann, 2018). Envisioning Cards⁵ can be utilized in deploying the method. Physical tools are commonly used to deploy methods in practice, be it cards or other approaches. We have also chosen to focus on a physical presentation for ECCOLA by making it a card deck.

VSD has, however, also been argued to have its shortcomings. In particular, it has been criticized for lacking in pragmatism and methodological guidance (van der Duin, 2019; Winkler and Spiekermann, 2018). Nonetheless, it has seen some success out on the field, which has been a recurring challenge for any method or tool involving ethics. We have also looked at VSD for some inspiration while designing ECCOLA, as we discuss further in the discussion section.

2.4. Essentializing to create methods from practices

In this final subsection of this section, we discuss a background theory that was utilized especially early on in the development of ECCOLA. The Essence Theory of Software Engineering (Jacobson et al. (2012)) is a method engineering tool. It comprises of two parts: (1) what its authors refer to as a kernel, and (2) a language. In short, the kernel offers premade building blocks for constructing methods using the language, and the language itself is used to model practices and methods.

More specifically, the kernel contains, as its authors argue (Jacobson et al., 2012), all the essential elements found in any SE project. The theory posits that every SE project, at bare minimum, has these elements in it, in addition to any additional project-specific elements. These elements are split into three types of items: alphas (i.e., things to work with), activities (i.e., things to do), and competencies (i.e., the skills required to carry out the project). Moreover, these elements are split into three areas of concern (i.e., categories): customer, solution, and endeavor.

The heart of the kernel consists of the aforementioned alphas, of which there are seven. In the customer area of concern, there

⁵ <https://www.envisioningcards.com/>.

are two alphas: (1) opportunity, and (2) stakeholders. There are also two alphas in the solution area: (3) requirements, and (4) software system. Finally, the endeavor area of concern contains the three final alphas: (5) work, (6) team and (7) way-of-working. Aside from helping the users of the tool structure methods, alphas are used to track progress on a project. Each alpha has alpha *states* that denote progress on that part of the project (e.g. requirements).

Originally, we intended to use the Essence language to describe the ECCOLA method. Essence was chosen due to its method-agnostic approach and modular philosophy on methods. From the get-go, ECCOLA was never intended to be a stand-alone method, but rather, a modular extension to existing software development methods that would bring in AI ethics into the process. Our plan was to devise alphas for AI ethics and to use the language to portray practices used to progress on them.

However, as we discuss in detail the following sections, we ultimately ended up giving up on the idea of using Essence to describe ECCOLA. Briefly put, utilizing Essence to describe ECCOLA made the method too heavy. Not only would the users of ECCOLA have to learn to use ECCOLA itself, they would also have to learn to use, or at least understand, Essence.

On the other hand, though ECCOLA is no longer described using the Essence language, we utilized the idea of *essentializing* practices in ECCOLA. Essentializing practices is described as a process by Jacobson (Jacobson et al., 2019) as follows:

“- Identifying the elements – this is primarily identifying a list of elements that make up a practice. The output is essentially a diagram [...]

- Drafting the relationships between the elements and the outline of each element – At this point, the cards are created.

- Providing further details – Usually, the cards will be supplemented with additional guidelines, hints and tips, examples, and references to other resources, such as articles and books”

As the above quote highlights, Essence utilizes cards to describe methods. This is also an approach we have utilized in ECCOLA. The ECCOLA method is utilized via a physical (or digital) set of cards. The cards are also created in a similar manner, although with some extra steps as ECCOLA cards have more (and different) content than traditional Essence practice cards. Although Essence is no longer used to describe the method itself, we still utilize the idea of essentializing practices to draft the cards for ECCOLA.

3. ECCOLA - A method for Implementing Ethically Aligned AI systems

As we have discussed in Section 2, AI ethics is currently an area with a prominent gap between research and practice. Much of the research has been theoretical and conceptual, focusing on defining key principles for AI ethics and how to tackle them. The numerous guidelines for AI ethics that currently exist (Morley et al., 2019) have tried to bridge this gap to bring these principles to the developers, but seem to not have had much success. Indeed, ethical guidelines tend to not have much impact in the context of SE (McNamara et al., 2018). To bridge this gap with another approach, we propose a method for implementing AI ethics: ECCOLA.

ECCOLA (Fig. 1) is intended to provide developers an actionable tool for implementing AI ethics. To utilize the various AI ethics guidelines in practice, the organization seeking to do so has to somehow make them practical first. ECCOLA, on the other hand, is intended to be practical as is, and ready to be incorporated into any existing method. ECCOLA does not provide any

Table 1
ECCOLA card themes.

Card themes (8)	Card number	Card amount (total 21)
Analyze	#0	1
Transparency	#1–6	6
Safety & Security	#7–9	3
Fairness	#10–11	2
Data	#12–13	2
Agency & Oversight	#14–15	2
Wellbeing	#16–17	2
Accountability	#18–20	3

direct answers to ethical problems, as arguably correct answers are a rare breed in ethics in general, but rather asks questions in order to make the organization consider the various ethical issues present in AI systems. Though how these questions are ultimately tackled is up to the users of ECCOLA, ECCOLA does encourage them to take into account the potential ethical issues it highlights.

In developing ECCOLA, we have had three main goals for the method:

1. To help create awareness of AI ethics and its importance,
2. To make a modular method suitable for a wide variety of SE contexts, and
3. To make ECCOLA suitable for agile development, while also helping make ethics a part of agile development in general.

ECCOLA is built on AI ethics research. It utilizes both existing theoretical and conceptual research, as well as AI ethics guidelines that have been devised based on existing research as well. In terms of guidelines, the cards are based primarily on the IEEE Ethically Aligned Design guidelines (IEEE Global Initiative, 2019) and the EU Trustworthy AI guidelines (HLEG, 2019). As these guidelines have already distilled much of the existing research on the topic under various principles, these principles have been utilized in ECCOLA as well. Existing AI ethics research has then been utilized to expand the way these principles are covered in ECCOLA.

In practice, ECCOLA takes on the form of a deck of cards. This approach was based on the Essence Theory of Software Engineering (Jacobson et al., 2012), which was used to describe the first versions of the method. Methods described using the Essence language are utilized through cards. However, using cards in the context of software engineering methods is not a novel idea, nor one originally proposed by Essence. E.g., Planning Poker in Agile uses cards. Moreover, various SE methods encourage the use of physical tools in general while using the method. The idea of Kanban, for example, is founded around using sticky notes on a signboard.

There are 21 cards in total in ECCOLA. These cards are split into 8 themes, with each theme consisting of 1 to 6 cards. These themes are AI ethics ones found in various ethical guidelines, such as transparency or data. Each individual card deals with a more atomic aspect of that theme, such as data privacy and data quality in the case of data. Aside from the main set of cards, ECCOLA also features an A5-sized game sheet that describes how the method is used (see Table 1).

Each card (see Fig. 2) in ECCOLA is split into three parts: (1) motivation (i.e. why this is important), (2) what to do (to tackle this issue), and (3) a practical example of the topic (to make the issues more tangible). Each card also comes with a note-making space. As the cards are generally utilized as physical cards, the card is split into two with the left half of each card containing the textual contents and the right half containing white space for making notes. This note-making space has been included to make using the cards more convenient in practice.

#0 Stakeholder Analysis

Motivation: In order to understand the big picture, it is important to first understand who the system can affect and how. Try to also think past the obvious, direct stakeholders such as your end users.

What to Do: Identify stakeholders.

- Who does the system affect, and how? Stakeholders are not simple users, developers and customers.
- How are the various stakeholders linked together?
- Can those different stakeholders influence the development of the system?
- Remember that a user is often an organization and the end user is an individual. Similarly, AI systems can treat people as objects for data collection.

Practical Example: Autonomous cars don't just affect their passengers. Anyone nearby is affected; some even change the way they drive. If at one point half of the traffic consists of self-driving cars, what are the societal impacts of such systems? E.g., regulators arising from such systems also affect users.

#1 Types of Transparency

Motivation: When considering transparency, it is important to understand who you are being transparent towards, and what you are being transparent about.

What to Do: Consider the following.

- Are you trying to understand something? (Internal transparency)
- Are you trying to explain something? (External transparency)
- Are you trying to understand or explain how the system works? (Transparency of algorithms and data)
- Are you trying to understand or explain why the system was made in the way it is now? (Transparency of system development)
- External stakeholders to consider, among others: (end-users, safety certification agencies, accident investigators, lawyers or expert witnesses, and society at large for disruptive technologies)

#2 Explainability

Motivation: If we cannot understand the reasons behind the actions of the AI, it is difficult to trust it.

What to Do: Ask yourself:

- Is explainability a goal for your system? How do you plan to ensure it?
- How well can each decision of the system be understood by both developers and end users?
- Did you make trade-offs between explainability and model performance for the context?
- How familiar are you with training or testing data? Can you change it when needed?
- How do you collect the data components in the system, how well do you understand them?

Practical Example: When interacting with a robot, users could clearly ask the robot "why did you do that?" and receive an understandable response. The would make it much easier for them to trust a system.

#3 Communication

Motivation: In practice, communication is a big part of being transparent with your stakeholders. Being transparent in communication can generate trust.

What to Do: Ask yourself:

- What is the goal of the system? Why is this particular system developed in this specific area?
- What do you communicate about the system to its users and end users? Is it enough for them to understand how the system works?
- If relevant to your system, do you sometimes tell your (end) users that they are interacting with an AI system and not with another human being?
- Do you collect user feedback? How is it used to change/improve the system?
- Are communication and transparency towards other audiences, such as the general public, relevant?

Practical Example: Clearly stating what data you collect and why can make you more trustworthy. Compare this to a mobile application that just states it needs to access your camera and storage.

#4 Documenting Trade-offs

Motivation: One important part of transparent system development is the documentation of trade-offs. Whenever you make a decision, you choose one option over other alternatives. However, documenting why and what the alternatives were is important.

What to Do: Ask yourself:

- How have you documented the development of the system and potential trade-offs between them identified and documented?
- Who decides on such trade-offs (e.g., between two competing solutions) and how? Did you ensure that the trade-off decision and the reasoning behind it were documented?

Practical Example: E.g., choosing machine learning algorithms for a trade-off between accuracy and explainability. Documenting trade-offs can improve your customer or end-user's ability to better explain why certain decisions were made over others. Moreover, it can reduce the responsibility placed on the individual developer(s) from an ethical point of view.

#5 Traceability

Motivation: Traceability supports explainability. It helps us understand why the AI did the way it did.

What to Do: Document different types of documentation (code, project etc.) are typically key in producing transparency.

- How have you documented the development of the system, both in terms of code and decision-making? How was the model built or the AI trained?
- How have you documented the testing and validation processes in terms of data and scenarios used and the results?
- What about different actions in mostly similar scenarios (e.g., if the user was different the situation otherwise the same)?

Practical Example: When the system starts making mistakes, being able to trace back to the underlying reasons from an ethical point of view. Consequently, it will be easier to find out the cause. Consequently, it will also be faster and possibly easier to train the underlying system from an ethical point of view.

#6 System Reliability

Motivation: Transparency makes ethical development possible in the first place. To make it ethical, we must understand how the system works and why it makes certain decisions.

What to Do: Ask yourself:

- How do you test if the system fulfills its goals?
- How well has the system comprehensively, including in unlikely scenarios? Have the tests been documented?
- When the system fails in a certain scenario, will you be able to tell why? Can you replicate the failure?
- How do you ensure the (end-user of the system's) reliability?

Practical Example: An autonomous coffee machine successfully brews coffee 8 times out of 10. While this is a decent success rate, we are still wondering what happened the 2 times it failed to do so, and why. Errors are inevitable, but we must understand the causes behind them and be able to replicate them to fix them.

#7 Privacy and Data

Motivation: Privacy is a rising trend in the wake of various recent data misuse events. People are now increasingly conscious about handing out personal data. Similarly, regulations such as the General Data Protection Regulation (GDPR) now affect data handling.

What to Do: Ask yourself:

- How data are used by the system?
- Does the system use or collect personal data? Why? How is the personal data used?
- Do you clearly inform your (end) users about any personal data you collect (e.g., ask for consent, provide an opportunity to revoke it, etc.)
- How do you ensure measures to enhance (end-user) privacy, such as?
- Who makes the decisions regarding data use and collection? Do you have organizational policies for it?

Practical Example: Rather than collecting and selling data, supporting privacy can also be profitable. Regulations are making it increasingly difficult to collect lots of personal data for profit. Privacy can be an alternate selling point in today's climate.

#8 Data Quality

Motivation: As AI are trained using data, the data used directly affects how the system operates. The nature, the quality, and integrity of the data used have to align with the goals of the system.

What to Do: Ask yourself:

- What are good or poor quality data in the context of your system?
- How do you evaluate the quality and integrity of your own data? Are there alternative ways to how you collect your data?
- Do you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for data management and governance?
- How can you tell if your data sets have been compromised? (e.g., data pollution).
- How do you handle the data collection, storage, and use?

Practical Example: In 2017, Amazon scrapped its recruitment data to teach the AI, as they had mostly hired men, the AI began to consider women undesirable based on the data.

#9 Access to Data

Motivation: Aside from carefully planning what data you collect and how, it is also important to plan how it can or will be used and by whom.

What to Do: Ask yourself:

- Who can access the user's data, and under what circumstances?
- How do you ensure that the people who access the data (1) have a valid reason to do so, and (2) adhere to the regulations and policies related to the data?
- Do you keep logs of who accesses the data and when? Do the logs tell why?
- Do you use existing data governance frameworks or protocols? Does your organization have its own?

Practical Example: Third parties you give access to the data can misuse it. A prominent example of this is the case of Cambridge Analytica and Facebook, in which data from Facebook was used irresponsibly. However, such data can be used to target your organization in a bad light even if you do not use the data in the intended way.

#10 Human Agency

Motivation: Trade interesting with the system or using it should be able to understand it sufficiently, users should be able to make informed decisions about whether or not to change its suggestions. AI systems should be humans make independent decisions.

What to Do: Ask yourself:

- Does the system interact with decisions by human actors, and if so, how? (e.g., recommending users actions or decisions, or presenting options)?
- Does the system encourage you to (end) user that a decision, context or action is the result of an algorithmic decision? How much detail does it go?
- In the system or context, what has been done by the system and what tasks are done by humans?
- How do you ensure measures to prevent overreliance or overconfidence on the system?

Practical Example: A medical system recommends diagnoses. How does the system communicate to doctors who it makes a recommendation? How should the doctors know when to challenge the system? Does the system even change how patients and doctors interact?

#11 Human Oversight

Motivation: AI systems should support human decision-making. They should not undermine human autonomy by making decisions for us, meaning they should be subject to human oversight.

What to Do: Ask yourself:

- Who can control the system and how? What situation?
- What would be the appropriate level of human control for the particular system and in use cases?
- Related to the Safety and Security cards: how do you detect and respond if something goes wrong? Does the system then stop entirely, partially, or would control be delegated to a human? Why?

Practical Example: Assuming control is especially related to cyber-physical systems such as drones or other vehicles. For purely digital systems, the focus should be on supporting human decision-making instead of directing it.

#12 System Security

Motivation: While cybersecurity is important in any system, systems prevent new challenges. Cyber-physical systems can even cause fatalities in the hands of malicious actors.

What to Do: Ask yourself:

- Did you assess potential forms of attacks to which the system is exposed or more relevant to AI systems?
- Did you consider different types of vulnerabilities, such as data pollution and physical infrastructure?
- Have you verified how your system behaves in unexpected situations and environments?
- Does your organization have cybersecurity personnel? Are they trained in this system?

Practical Example: The autonomous nature of AI systems makes new vectors of attack possible. A white line drawn across a road can confuse a self-driving vehicle. The case of Microsoft's Tay Twitter bot, who began to exhibit extreme views after being bombarded with such, is one example of a new type of attack.

#13 System Safety

Motivation: AI systems exert notable influence on the physical world whether they are cyber-physical or not. Various risks and their consequences should be considered, striving aimed to the operation of the system.

What to Do: Ask yourself:

- What kind of risks does the system involve? What kind of damage could it cause?
- How do you measure and assess risks and safety?
- What kind of risks does your system pose? Have they been analyzed?
- In what conditions are the failures your system might be autonomous or do they require human input?
- Is there a plan to mitigate or reduce technological errors, accidents, or malfunctions? What if the system provides wrong results, becomes unavailable, or provides completely uninterpretable results?
- What liability and consumer protection laws apply to your system?

Practical Example: AI systems can also accurately predict organizational risks, making it possible to reduce personnel. However, if a customer satisfaction becomes reliant on your system to handle a portion of its operations, what happens if that AI stops functioning for even a few days? What could you do to improve the impact?

#14 Accessibility

Motivation: Technology can be discriminating in various ways. Given the enormous impact AI systems can have, ensuring equal access to their benefits is ethically important.

What to Do: Ask yourself:

- Does the system consider a wide range of individual performance and abilities? If not, why?
- Is the system usable by those with special needs or disabilities, those at risk of accidents, or those using assistive technologies?
- How does your system represent various groups throughout the development of the system?
- How is the potential user audience taken into account?
- Is the system inclusive in building the system representative of your target user audience? Is it representative of the general population?
- Did you assess whether there could be impacts of people who may be disproportionately affected by the negative implications of the system?

Practical Example: AI tends to benefit to those who are already advantagedly capable, resulting in increased inequality.

#15 Stakeholder Participation

Motivation: As AI systems have notable impacts, their stakeholders have a right to be involved in the development. Yet, e.g. when using a decision-making system, its users have to trust the system while also being critical of it.

What to Do: Check your stakeholder analysis (Card #0).

- Which stakeholders are involved in system development?
- How are the different stakeholders of the system involved in the development of the system? If they aren't, why?
- How do you inform your external and internal stakeholders of the system's development?

Practical Example: Often the people an AI system is used on are individuals who are directly affected by the system. For example, a medical system (a developer for hospitals, used by doctors, but ultimately used on patients, who not talk to the patients too?

#16 Environmental Impact

Motivation: Part of the general wellbeing implications, ecological consciousness is a current trend. Being ecological can be a selling point for your organization.

What to Do: Ask yourself:

- Did you assess the environmental impact of the system's development, deployment, and use? E.g., the type of energy used by the data center.
- Did you consider the environmental impact when selecting specific technical solutions?
- Did you ensure measures to reduce the environmental impact of your system's life cycle?

Practical Example: If you are hosting on a third party cloud, try to ascertain the sustainability of the service provider's services. If you are using hardware, are you processing the data in each physical device or your own or are you processing it in the cloud?

#17 Societal Effects

Motivation: The impact of a system goes beyond its user-base. A system may affect negatively even those who do not use it or wish to use it.

What to Do: Ask yourself:

- Did you assess the broader societal impact of the AI system's use beyond the individual (end-user's) customer stakeholders who might be indirectly affected by the system?
- How will the system affect society when in use?
- What kind of societal effects could the system have?

Practical Example: Surveillance technology offering facial recognition AI has long-reaching impacts. People may wish to avoid areas that utilize such technology, negatively affecting businesses in said area. People may become stressed at the mere thought of such surveillance. Some may even engage in a revolt.

#18 Auditability

Motivation: Regulations affecting AI and data may necessitate audits of systems in the future. Similarly, if the system causes damage, an audit might be requested. It is good to have mechanisms in place beforehand.

What to Do: Ask yourself:

- Is the system auditable?
- Can an audit be conducted independently?
- Is the system available for inspection?
- What mechanisms facilitate the system's auditability? How is traceability and logging of the system's processes and outcomes ensured?

Practical Example: In heavily regulated fields such as medicine, audits are typically required before a system can be utilized in the first place.

#19 Ability to Redress

Motivation: Making sure people know they can be compensated in some way in the event something goes wrong with the system is important in generating trust. Such scenarios should be planned in advance to what extent possible.

What to Do: Ask yourself:

- What is your developer/organizational responsibility if the system causes damage or otherwise has a negative impact?
- In the event of negative impact, can the ones affected seek redress?
- How do you inform users and other third parties about opportunities for redress?

Practical Example: AI systems can incorporate users in unforeseen, unpredictable ways. Depending on the situation, the company may or may not be legally responsible for the consequences. Nonetheless, by offering a digital platform for seeking redress, your company can seem more trustworthy while also offering additional value to your users.

#20 Minimizing Negative Impacts

Motivation: Minimizing negative impacts of the system is financially important for any developer/organization. Individuals are often costly.

What to Do:

- First, consider:
 - Is your stakeholder analysis up-to-date (Card #0)?
 - Have you discussed risks? (Card #13)
 - Have you discussed auditability?
 - Have you discussed redress issues?
- Are the people involved with the development of the system also involved with it during operational life? If not, they may not feel as accountable.
- Are you aware of how related to the system?
- Can users of the system somehow report vulnerabilities, risks, and other issues to the system?
- What have you done to discuss accountability and other ethical issues related to the system, including grey areas?

Card Themes

Analyze
Transparency
Safety & Security
Fairness

Data
Agency & Oversight
Wellbeing
Accountability

Vile Vakkuri JYU
ville.vakkuri@jyu.fi

Ka-Kristian Kemell JYU
ka-kristian.o.kemell@jyu.fi

Pekka Abrahamsson JYU
pekka.abrahamsson@jyu.fi

Fig. 1. ECCOLA - a method for implementing ethically aligned AI systems.

5

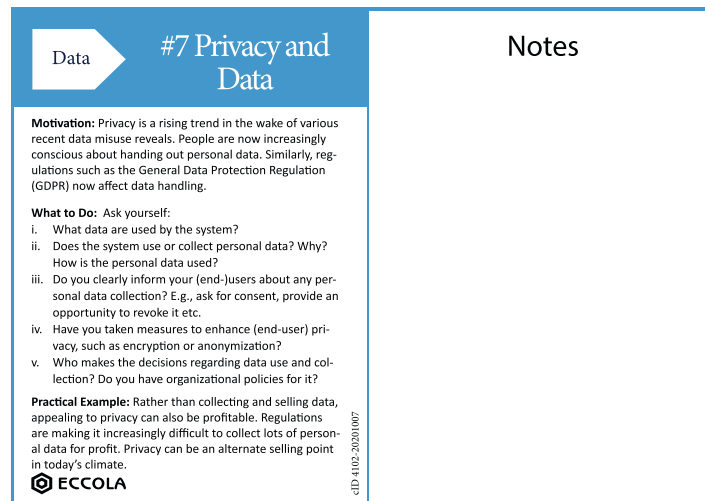


Fig. 2. Card example from ECCOLA, Card #7 privacy and data.

ECCOLA supports iterative development. During each iteration, the team is to choose which cards, or themes, are relevant for that particular iteration. ECCOLA is also method-agnostic, making it possible to utilize it with any existing or in-house SE method. In the following subsection, we discuss how to use ECCOLA in practice.

3.1. How to use ECCOLA in practice?

Expanding on what we already discussed in this section, i.e. what ECCOLA is, this section describes how to implement the ECCOLA method in practice. It includes descriptions of how ECCOLA has been used for different purposes, and our recommendations on how to proceed with using the ECCOLA cards in software development projects.

ECCOLA is a modular, sprint-by-sprint process that has been designed to facilitate ethical thinking in AI/S (Artificial Intelligence/Autonomous System) development. While using ECCOLA, you choose the cards you feel are relevant for your work currently and then evaluate the situation again after each sprint. Using ECCOLA results in a paper trail of choices and trade-offs that documents the ethical consideration conducted during development. This documentation provides a way of evaluating the trustworthiness of the system.

ECCOLA is intended to be used during the entire design and development process in a three step process that is repeated in every iteration. (1) Prepare: Choose the relevant cards for the current sprint. (2) Review: Keep the selected cards on hand during work tasks. Write down on the cards the actions you have taken and (ethical) discussions you have had. (3) Evaluate: Review to ensure that all the planned actions were taken. Revise the card deck as needed, and repeat the process. Remember to do a retrospective afterwards.

Everyone involved with using the cards should read the cards thoroughly at least once before the sorting process in order to familiarize themselves with the topics of the cards as well as their contents. This is recommended not only to make the decision process easier, but also to save time when selecting cards for each sprint.

ECCOLA cards are designed to offer a variety of viewpoints to prompt thoughts during the development process, and the idea is to utilize different cards in different stages of development

- and to not necessarily use all cards in every project either. Each software development endeavor is unique, e.g. in relation to the requirements and the scope of the project. ECCOLA cards should therefore also be selected based on the project and tasks at hand. Cards irrelevant to the current situation can be discarded during the sorting process. The sorting should preferably be conducted before the development process starts, so that the prompts presented by the cards can be utilized from the beginning. The sorting process should include everyone who will be using the cards, and possibly other members of the project who are involved with the product's development.

Before starting to use the cards in a development process, we recommend sorting the cards into piles based on which stage of the development they will be used in. Cards that are deemed irrelevant for the project can simply not be used during that project. This selection process should be documented by briefly explaining why some cards were selected and why some were considered irrelevant in each iteration, to support transparency in the context of systems development. Documenting ethical choices in general is encouraged while using the method. Our recommendation for sorting the ECCOLA cards is to create three piles of cards.

Pile 1 for the early stages and planning stages in a project. Pile 2 for any other parts of the project, throughout development. These should be adjusted on a sprint-by-sprint basis as well. The chosen cards, or specific parts of each card, can then be considered in relation to the activities in that sprint. Finally, Pile 3, if needed, towards the end of the project if there is a need to evaluate a decisions, or if there have been any unexpected occurrences.

When introducing ECCOLA to new organizations and people interested in using it, we have typically held an introductory workshop, which we discuss in the subsection below.

3.1.1. Getting acquainted with the cards/tutorial sessions

To introduce new users to ECCOLA, we have held tutorial sessions in the form of workshops. Similar sessions could also be held in organizations looking to start using ECCOLA. Below is a brief outline of these sessions.

The following outline has been used for ECCOLA tutorials:

1. A presentation on ECCOLA (and AI ethics if necessary).

2. Introducing the hypothetical product and planning its features and requirements.
3. Sprints 1, 2 and 3 where new features or requirements are introduced for each sprint. Each sprint lasts e.g., 15–20 min.
4. Discussion and feedback.

The introduction should familiarize the participants with the method, and can contain a brief introduction to AI ethics as well, focusing on why it is important and what it is, with a focus on practical issues. After the introductory presentation, the participants are given a task to work on. For example, during the COVID-19 pandemic, we had workshop participants design an AI-based mobile application for tracking and limiting its spread. The participants then split into groups (e.g., 5 per group) and design such a system according to the given requirements while using the ECCOLA cards.

This work is carried out in three sprints of e.g., 15–25 min. Each sprint can contain pre-selected cards, or the participants can be instructed to choose the cards themselves for each sprint. If the participants are to select their own cards, the sprints should also be longer in duration. Between sprints you can have a brief discussion session, or you can go through the sprints in quick succession and have a longer one afterwards.

4. Research method

In this section, we discuss the Cyclical Action Research approach we have utilized to develop ECCOLA. Our approach was based on that discussed by [Susman and Evered \(1978\)](#) and, in further detail, by [Davison et al. \(2004\)](#). We chose this approach as we wanted to iteratively develop the method over time, testing it in different contexts in the process. Moreover, Action Research (AR) is well-suited for using different data collection methods in different contexts ([Susman and Evered, 1978](#)).

Thus far, we have completed 7 Action Research (AR) cycles and are currently conducting an eighth one. These have been split into 6 research stages, with most research stages featuring one cycle, aside from stage 2 that consisted of three cycles. These are shown in [Fig. 4](#) and [Table 2](#), and each stage is further discussed in the following data analysis section. In this current section, we discuss the cyclical research approach of this study more generally from a methodological point of view.

Past the very first AR cycle that focused on testing an existing tool, each cycle has proceeded in the same general manner. In each cycle, we have tested a version of ECCOLA in practice in some context, collected data from its use, and then used the data to improve the method. After this, we have started a new cycle. In the diagnosis phase of each cycle, we have looked at literature on the topic to determine whether ECCOLA should be further modified based on literature before a new test in a different context.

The initial cycles (Stages 1–2) focused on student testing. We used student projects early on as we wished to make the method more mature before industry testing. In Stage 3, we started to also include industry testing in the form of a small-scale blockchain project. In addition to this, in Stage 3, we began to host academic workshops at conferences, as well as privately organized academic workshops, to collect feedback from the scientific community (using the Tutorial Session outline in [Section 3.1.1](#)). Finally, we shifted our focus further towards industry testing in Stages 5 and 6, and we are currently cooperating with multiple companies using ECCOLA. The way we have progressed from student testing to industry testing in this fashion is also inspired by the continuous co-experimentation approach described by [Mikkonen et al. \(2018\)](#).

In our industry testing, we have utilized an approach that has been referred to as industry-as-a-lab by [Potts \(1993\)](#). This approach focuses on “what people actually do or can do in practice”. As many of the current problems in the area resulting in the gap between research and practice seem to stem from a lack of practical tools, we have focused on making ECCOLA practical. To achieve this, we have focused on receiving continuous feedback primarily through formal data collection and throughout the process improving the method based on the feedback before then testing it again. A more recent example of this approach is the study of [Mikkonen et al. \(2018\)](#).

Finally, perhaps worth noting is that the research team behind this endeavor has past experience in developing methods as well. Namely, one of the authors proposed the Mobile-D approach for developing mobile applications in an Agile manner when Agile was still emerging ([Abrahamsson et al., 2004](#)).

In the subsections below, we discuss each phase of the Cyclical Action Research model discussed by [Susman and Evered \(1978\)](#) (and [Davison et al. \(2004\)](#)). [Susman and Evered \(1978\)](#) highlight five phases ([Fig. 3](#)) in this cyclical process that they posit are all necessary. We describe our process according to these phases in the subsections of this section.

4.1. Diagnosis

In the initial cycle, diagnosing the problem was focused on understanding the gap in AI ethics in general. We have published papers about this in the past, with [Vakkuri et al. \(2020\)](#) looking at this gap quantitatively and e.g. [Vakkuri et al. \(2020\)](#) looking at it qualitatively. While collecting data for these papers, we began to see that there is indeed a gap between research and practice in the area, and started to also look for ways to bridge the gap.

In Stages 2 and up, when we were already developing ECCOLA, the diagnosis phases focused on better understanding *what* is AI ethics and, to this end, what exactly is the problem ECCOLA should help solve. In addition to improving ECCOLA based on our data from each preceding cycle, in the diagnosis phase of each cycle, we looked at motivation behind ECCOLA. Whereas Action Research traditionally focuses on solving problems an organization has, in this case, it was largely up to us to define the problem and then convince organizations that it was a real problem. However, towards the latest stage, we have noticed that AI ethics has become much more topical out on the field to the point where we have had organizations volunteering to work with us on developing ECCOLA.

The main question in the diagnosis phase of each cycle was always whether our idea of AI ethics was still up-to-date. Was ECCOLA still in line with the current discussion on AI ethics? For example, the EU guidelines on AI ethics ([HLEG, 2019](#)) were published after Stage 2 ([Fig. 4](#)), and in our minds presented a major contribution to the field, which we felt should also influence ECCOLA.

4.2. Action planning

In the first stage ([Section 5.1](#)) where we ultimately tested the RESOLVEDD strategy, we considered alternative courses of action. Having identified a gap in the area, we looked at different alternatives for solving the problem. Using the existing AI ethics guidelines to bridge the gap was one option. However, existing papers argued that ethical guidelines alone were unlikely to work in AI ethics ([Mittelstadt, 2019](#)) or SE engineering in general ([McNamara et al., 2018](#)).

We therefore turned to methods that could help us tackle it. First, we looked at existing methods for implementing ethics. As a result, in Stage 1 of our study ([Section 5.1](#)), we studied

Table 2
Cyclical action research stages.

Stage	Version in action	Primary background theories	Study setting	Timing	Participant
1	n/a	RESOLVEDD, EAD, Essence	Class	Q1-Q2 2018	5 teams of 4-5 students
2	1	RESOLVEDD, EAD, Essence	Class	Q2 2018 - Q2 2019	27 teams of 3-5 students
2	2	RESOLVEDD, EAD, Essence	Class	Q2 2018 - Q2 2019	27 teams of 3-5 students
2	3	RESOLVEDD, EAD, Essence	Class	Q2 2018 - Q2 2019	27 teams of 3-5 students
3	4	EU AI HLEG, EAD	Blockchain Project	Q2-Q3 2019	2 sw development team members
4	5	EU AI HLEG, EAD	Conference Workshop	Q4 2019	8 researchers
5	6	EU AI HLEG, EAD	Industrial & Conference Workshops	Q1-Q3 2020	2 Company cases & 10+ ICT researchers
6	7	EU AI HLEG, EAD	Industrial	Ongoing	

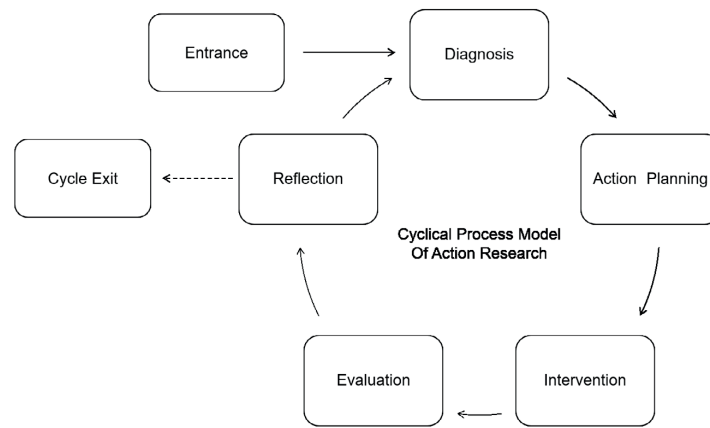


Fig. 3. Based on Davison et al. (2004) and Susman and Evered (1978).

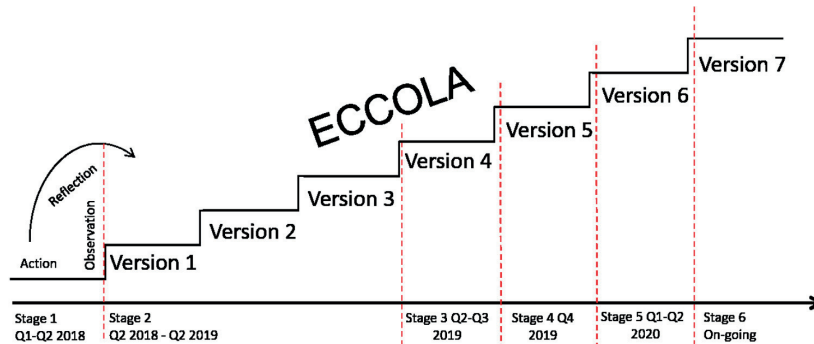


Fig. 4. Cyclical action research process on ECCOLA. Including cycle of action, observation, reflection on each iteration.

an existing ethical tool from the field of business ethics, the RESOLVEDD strategy, in the context of AI ethics, and argued based on our findings that methods and tools specific to AI ethics are required (Vakkuri and Kemell, 2019). As a result, in the absence of existing AI ethics methods, we began to work on ECCOLA.

In the stages past Stage 1, Action Planning was focused on determining how to test each version of ECCOLA. This included deciding on what type of data to collect and how. As we had already committed to developing ECCOLA, we no longer actively considered other ways of tackling the gap.

4.3. Intervention (or action taking)

The main intervention in all the stages of this study past the first one has been the introduction of ECCOLA. In the student and

industry contexts, the project would have existed and been carried out with or without ECCOLA. ECCOLA was simply introduced as a framework for conceptualizing a problem (i.e. various ethical issues). This can be likened to the way Susman (1976) describe surprise in interventions: "the element of surprise evoked by an intervention results when the change agent offers members of the target organization a new way to conceptualize an old problem and offers it in a language or framework that differs from that by which members of the organization define their present situation". On the other hand, the academic workshops were created for the sole purpose of having the participants use ECCOLA, even though the mini-projects of the workshops could have been carried out without ECCOLA as a framework.

The introduction of ECCOLA has been accompanied by other actions taken to facilitate its adoption and use. These have varied between the research stages, but each stage has generally included 1) an introductory lecture or a workshop on ECCOLA, and 2) various check-ups to discuss the use of ECCOLA and any problems faced while using it. These have been used for data collection purposes as well, with especially the check-ups serving as a way of generating important data in the form of feedback for the evaluation phase of each Action Research (AR) cycle.

In student contexts, the use of ECCOLA continued for a set amount of weeks during a course project. In academic contexts, i.e. workshops, the use of ECCOLA lasted some hours. In industry contexts, the use of ECCOLA lasted for a duration of a project (Stage 3) or is still on-going (Stage 6).

4.4. Evaluation

Evaluation was conducted both during and after the use of ECCOLA in each stage. The focus of the evaluation was to understand what effect ECCOLA had had on the way its users worked, i.e. how it had changed existing practices and whether it had added new work practices. In doing so, we wished to also understand how the users of ECCOLA had felt about ECCOLA while using it.

We collected different types of data in different stages of the study (Fig. 4, Table 2). Across these stages, we have used work products (sheets, notes, text etc.), ECCOLA cards with notes on them, observation, unstructured interviews, and informal discussions as sources of data. In the next section (Section 5), we discuss what types of data were used in each stage in the respective subsections. The data collected in each stage is also summarized in Table 3.

4.5. Reflection (or specifying learning)

As we have developed ECCOLA iteratively in this process, the reflection phases have primarily focused on improving ECCOLA based on the data collected in each research stage. Indeed, the evaluation of ECCOLA has also been the focus of the data collection. In each reflection phase, we looked at ECCOLA from two points of view.

First, we looked at how ECCOLA had worked as a method in that stage. Had the method itself been clear to its users? Had the users managed to follow the process suggested by ECCOLA? To determine this, we looked at the notes on the ECCOLA cards and other work products to see how (or if) the cards had been utilized, or discussed their use with the subjects for example.

Secondly, we looked at the theory behind ECCOLA, i.e. AI ethics. Were we presenting the principles in an understandable way and were the users of ECCOLA grasping the concepts? Was something missing based on the data, or did something need to be further emphasized? For example, sometimes we would receive direct feedback regarding the wording on some of the cards.

Additionally, we critically evaluated our research process and choices regarding it. We looked at shortcomings in our data collection methods and how we introduced ECCOLA into the research context in each cycle. For example, the introductory session we have hosted at workshops and for companies (see Section 3.1.1) has been improved over time as well.

5. ECCOLA development stages and data

ECCOLA has been developed iteratively through multiple stages. In each stage, we have collected empirical data, which has then been used to iteratively improve the method. The current version of ECCOLA is its seventh version. The subsections of this section each present one development stage in the iterative development process of ECCOLA. At the end of each section is a brief summary of what changes were made in each stage. This process is also summarized in Table 2 below, as well as in Fig. 4.

5.1. Stage 1 (Q1-Q2 2018)

In early 2018, prior to starting our work on ECCOLA, we searched for existing methods for AI ethics, ultimately finding none. Thus, we expanded our horizons and looked at ethical tools from other fields instead to see if anything would seem applicable in the context of AI ethics as well. This led us to eventually test an existing ethical tool from the field of business ethics, the RESOLVEDD strategy (Jacobson et al., 2012), in the context of AI ethics. Our aim was to see if existing ethical tools, even if they were not specifically created for AI ethics, could be suitable for that context.

We conducted a scientific study on RESOLVEDD in the context of AI ethics. These findings have been published in-depth elsewhere (see Vakkuri and Kemell (2019)). In short, we discovered that forcing developers to utilize RESOLVEDD did have some positive effects. Namely, it produced transparency in the development process, and the presence of an ethical tool made the developers aware of the potential importance of ethics, resulting in ethics-related discussions within the teams. However, the tool itself was not considered well-suited for the context by the developers, and they felt that using the tool was detached from the rest of the processes. Moreover, when forcing developers to utilize such a tool, the commitment towards it quickly vanished when the tool was no longer compulsive.

Stage 1 actions: The development of ECCOLA was initiated

5.2. Stage 2 (Q2 2018 - Q2 2019)

5.2.1. Creating Version 1 (Q2 2018 - Q1 2019)

Based on the results of this study, we began to develop a method of our own, ECCOLA, during the latter half of 2018. This initial version of the method was based on three primary theories: (1) RESOLVEDD strategy (Pfeiffer and Forsberg, 1993), (2) The Essence Theory of Software Engineering (Jacobson et al., 2012), and (3) The IEEE Ethically Aligned Design guidelines (IEEE Global Initiative, 2019).

We utilized some of the general ideas of RESOLVEDD, which were deemed useful based on the data we collected. Namely, we (1) looked at RESOLVEDD for ideas on how to make the tool function in conjunction with iterative SE methods, and (2) for ideas on how to conduct comprehensive stakeholder analyses as the basis of the ethical analysis. We also included some of the aspects of RESOLVEDD which were shown (Vakkuri and Kemell, 2019) to support transparency of systems development (e.g. the idea of producing formal text documents while using the method).

We began to describe the method using the Essence language (see Section 2.4). Methods described using Essence are visualized through cards, and thus, ECCOLA took on the form of a card deck as well. This also meant that we included the various elements of Essence into the cards. For example, we made some of the key AI ethics principles, namely transparency, accountability, and responsibility, into alphas in the context of Essence (i.e., measurable things to work on). The cards also included various activities that were to be performed in order to progress on these alphas, as well as patterns and other Essence elements.

The AI ethics contents of the method, at this stage, were based primarily on the IEEE Ethically Aligned Design guidelines (IEEE Global Initiative, 2019). The field in general was still less formulated than it currently is, and thus the main AI ethics principles were still under more discussion than they currently are (e.g., Jobin et al. (2019) show that the field has since reached some consensus). We included key principles from the guidelines such as transparency and accountability, which have been prominent topics of discussion in AI ethics. Additionally, we utilized various research articles. For example, to expand on transparency, we

Table 3
Research stage and data collection.

Research stage	Data collection tools
1	Semi structured interviews for users
2	Note taking, mentor meetings, work-product (course), ECCOLA cards (user notes)
3	User interview, note taking, ECCOLA cards (user notes)
4	Note taking during workshop, unstructured participant interview, workshops recording
5	Note taking during workshop, unstructured participant interview, workshops recording, ECCOLA cards (user notes)
6	Note taking during tutorial, works, recurring project meetings, workshops recording, unstructured developer interview, ECCOLA cards (user notes), project documentation

utilized the studies of [Dignum \(2017\)](#) and [Ananny and Crawford \(2018\)](#), among others.

Much like how while using RESOLVEDD one produces text answering some questions posed by the tool, we incorporated the same idea of producing text while using ECCOLA into the initial version of the method. The theoretical background of this early version was based primarily on the IEEE EAD guidelines and academic articles discussing some individual principles.

5.2.2. Testing Version 1 (Q1 2019)

This first version of ECCOLA was tested in a large-scale project-based course on systems development at the University of Jyväskylä (Q1 2019). In the course, 27 student teams of 4–5 students worked on a real-world case related to autonomous maritime traffic. Each team was tasked with coming up with an innovation that would help make autonomous maritime traffic possible. The teams were not required to actually develop these innovations into functional products, given the time and capability constraints in a course setting, but rather, to refine the ideas as far as they could in the context of the course. The results of these projects have been published in an educational book⁶

The teams were introduced to ECCOLA during a course lecture and were handed a physical card deck. Each team was then told to utilize the card deck in whatever way they saw fit, while writing down notes on the cards as – or if – they used them. After the students had utilized the cards for a week, they were collected and the written notes on them analyzed. Additionally, unstructured interview data was collected from the teams through their weekly meetings with their assigned mentor and this feedback was taken into account in developing the method.

Prior to the course, the students had been tasked with reading a book on Essence, Software Engineering Essentialized ([Jacobson et al., 2019](#)), which explains the tool. Though the educational goal of this was elsewhere, this also served to make sure the students would not be overtly confused with this version of ECCOLA being described using the Essence language.

Based on the data collected, the language on the cards was considered difficult to understand and overall they were considered too academic by the teams. The cards were considered impractical, with the teams having difficulties applying their contents into practice. The students were also confused by the Essence notation.

Actions based on Iteration 1 of Stage 2, for Version 2: (1) Alpha states were added to the alphas in order to make tracking progress on them easier. (2) Practical examples were added to the cards to make it easier to understand the practical implications of the ethical issues in the cards. (3) Reduced the amount of academic jargon on the cards, focusing on practice over theory. (4) Removed list of academic references from each card.

⁶ <http://urn.fi/URN:ISBN:978-951-39-7689-7>.

5.2.3. Testing Version 2 (Q1 2019)

This iteration took place during the course described above and was carried out in the same manner as the previous one. The same student teams utilized this newer version of ECCOLA again while writing down notes on the cards as they did. Additional data was again collected in the weekly mentor meetings. Overall, this was, in terms of time elapsed, a brief iteration carried out during the course.

After another week, ECCOLA was once evaluated using the data we collected. The teams still found the method confusing. In particular, they found it difficult to understand how the cards tied together, and how they should be utilized. Even if the individual cards were made more practical, the language was still considered difficult to understand. Thus, the following changes were made to the method based on the data.

Actions based on Iteration 2 of Stage 2, for Version 3: (1) Added a game sheet describing how the cards (and the method) should be used. We realized that the method, in this version, required teaching to be understood. (2) Added numbering to the cards. (3) Further reduced the amount of academic jargon on the cards.

5.2.4. Testing Version 3 (Q1 2019)

The third version of ECCOLA was also tested in the same course as the previous two. However, as this was towards the end of the course, there were no further iterations to be tested in the same setting. Thus, we took our time to analyze the feedback from all three versions, reflect on it, and study new publications in the area to improve the method.

In analyzing the data from the teams, we focused on evaluating the level of utilization. This was done by analyzing the notes the teams made on the cards. The notes were evaluated on a scale of 0 = no notes or markings, 1 = single words or markings, 2 = sentences or more.

Also, we evaluated the cards independently based on the notes. The cards that were utilized the most and affected the projects the most were either cards with practical themes (e.g. data handling), or cards focusing on the big picture of the project at hand (e.g. cards focusing on 'what' and 'how' questions). On the other hand, the cards that were utilized the least, were the ones focused on accountability and other AI ethics specific issues. It seemed that many of the AI ethics principles, even with practical examples, were considered difficult (or irrelevant) by the teams. The cards describing AI ethics principles were utilized by 53% of the teams, whereas the other cards had a utilization level of 75% on average.

This resulted in a lengthier creation process for the subsequent version of ECCOLA. Based on the data and our reflection we made substantial changes to the method. We discuss these in the following subsection.

5.2.5. Creating Version 4 (Q2 2019)

The earlier versions of ECCOLA were cumbersome to use based on initial tests (see above). Utilizing these versions did result in ethical analyses and had an impact on the projects. However, the method was difficult to understand and especially the AI ethics

principles in particular were difficult to grasp for the teams utilizing the tools. After the course in which the first three versions of the method were tested, we made larger improvements based on the data.

First, we changed the way the method was described. We opted to lessen the role of Essence in ECCOLA. The Essence language used to describe the method seemed to make the method even more difficult to learn, as its users had to learn to use the method *and* to learn to understand the Essence language (and Essence in general). We stopped using the Essence elements in the cards and instead split the cards into different AI ethics themes. However, the general approach of making the method a card deck seemed to work and thus this approach was kept.

Secondly, the method seemed to be too heavy to use. ECCOLA was initially designed to be a linear process that was iteratively repeated. The idea was that its users could modify this process based on the context at hand to adjust the method to their projects. Nonetheless, this approach was considered too rigid, and the respondents felt, that it was just another process tacked onto their other work processes. Thus, we made the method modular, with the cards being more stand-alone on average, though some cards were still linked together in some ways. The users of ECCOLA could, following this approach, choose which cards to utilize in each situation (e.g., sprint) based on the context. The intent behind this was to make ECCOLA more suited for use with Agile methods.

During this time period, before the next empirical test, we also expanded the theoretical basis of the method. The initial version of the EU Guidelines for Trustworthy AI (HLEG, 2019) were published in early 2019, some aspects of which we chose to incorporate into ECCOLA. Other novel literature was also included to expand on theoretical basis of the method.

Changes made based on Stage 2 overall: (1) The use of Essence to describe the method was discontinued. (2) Contents of the cards reformatted and reformulated. (3) Method made modular rather than one linear, iterative process. (4) Expanded the AI ethics theoretical basis of the method.

5.3. Stage 3 (Q2-Q3 2019)

As the primary concern with the versions 1–3 had been the way ECCOLA was used as a method in practice rather than its AI ethical contents, we chose to focus on making a method, which is easier and more practical to use. For this purpose, we made a spin-off of ECCOLA for the context of blockchain ethics. Many of the AI ethical themes such as transparency and data issues could be translated into this context, even if the contents of the cards had to be modified to be better suited for it. Additional blockchain specific issues were also added into these cards.

In this stage, ECCOLA was utilized in a real-world blockchain project by two of the project team members. Data was collected through observation and various unstructured interviews. The team was free to utilize the cards as they wished, and was encouraged to reflect on how the method would best suit their SE development method of choice. However, the team could also receive consultation from one of the researchers where needed on how to use the cards, as well for clarification on their contents, if needed. As a result, we gained a better understanding of how the method was utilized in practice (e.g., how many cards were used per iteration on average, which was 6) in a real-world SE context.

Based on the data gathered from the blockchain project, the main ECCOLA card deck was iteratively improved. The lessons learned from studying the use of the blockchain ethics version of ECCOLA were incorporated into 5th version of ECCOLA.

Changes made based on Stage 3: (1) A note-making space was added to each card. (2) Added new cards. (3) Split the

cards into themes, such as transparency or data. (4) Added more contextual content into each card, as opposed to focusing largely on instructions on what to do. This resulted in revamping the “motivation” and “practical example” section of many of the cards. (5) Added new content focusing on stakeholder analysis and requirements, in order to help the users of the method gain an understanding of the big picture at hand.

5.4. Stage 4 (Q4 2019)

After improving ECCOLA based on the lessons learned from the blockchain project, ECCOLA was presented in a workshop in a scientific conference (ICSOB2019). In this workshop the participants utilized ECCOLA to discover potential ethical issues in a hypothetical AI development scenario. The participants of the workshop were split into two groups for the task.

The first group was tasked with developing an idea for an AI-based drone that would help farmers improve their harvests. The second group was tasked with developing an AI-based system that would filter and evaluate immigration applications. During the workshop, the groups worked on the ideas in timed iterations. Each group had a customer stakeholder that progressively presented them with more requirements at the end of each iteration. For every iteration, the groups would select the ECCOLA cards they felt were the most relevant for the requirements of that iteration.

At the end of the workshop, verbal feedback from the participants was collected. This was done in the form of a discussion where the participants talked about their experiences with each other and between the two groups. These group interviews were recorded and later transcribed for analysis. The feedback was then utilized to develop the 6th version of ECCOLA.

Changes made based on Stage 4: (1) The themes in the cards were color coded for clarity. (2) The practical examples in the cards were improved.

5.5. Stage 5 (Q1-Q3 2020)

A paper presenting the early 2020 (i.e., 6th) version of ECCOLA was published at DSD/SEAA2020 (Vakkuri et al., 2020). This paper extends said DSD/SEAA paper.

In the first half of 2020, ECCOLA was presented at the XP2020 conference in a workshop. The workshop was organized in a similar manner as the one at ICSOB2019 described in the previous subsection, with some modifications. The participants were split into three groups and tasked with working on a hypothetical AI/S project where they were to design a system for COVID-19 spread monitoring, while using ECCOLA to dwell on the potential ethical issues. This time, as the conference was held remotely, the participants communicated online, utilized a digital version of ECCOLA, and produced work products online. The work products (written documents) produced by the teams were collected for later analysis of the use of ECCOLA.

Additionally, we have held three privately organized ECCOLA workshops not associated with any scientific conference. These have been workshops for researchers active in the field, for the purposes of various research projects. These have been organized in a similar manner to the conference workshops, with the participants utilizing ECCOLA to work on a hypothetical project after a brief introduction to the method.

During 2020, ECCOLA was also adopted by three companies. One of these companies began using ECCOLA as early as late Q1 2020. In preparation for further company adoption, we utilized the workshop data, preliminary feedback from this one case (unstructured), and the other data collected in earlier stages, to create the current (7th) version of ECCOLA.

Changes made based on Stage 5 (resulting in the current version of ECCOLA): (1) Improved card layout based on company feedback (numbered card contents for easier referencing). (2) Improved individual card readability and textual content based on early company feedback with a focus on reducing the chance of any of the content being misunderstood. (3) Made changes based on current academic discussion. (4) Improved some of the practical examples on the cards with a focus on making them less tied to any current real events. (5) Fine-tuned the visual appearance of the cards.

5.6. Stage 6 (on-going)

Currently, we are cooperating with three companies to collect industry use data on ECCOLA. These companies are detailed in Table 4. With each company, we have held a workshop similar to the ones we have held at conferences to introduce them to the method. After this, we have kept in touch with the companies regarding the utilization of the method through recurring meetings. While we have collected data from these meetings as notes and discussed their experiences using ECCOLA during the meetings, these cases are still pending formal data collection.

So far, in our discussions with the participants, the companies have indicated that they have successfully utilized ECCOLA in conjunction with their existing methods. They feel that ECCOLA has successfully been modular. To this end, ECCOLA also seems to work in conjunction with agile methods, as all the companies consider themselves agile. However, we have not yet collected any work products or ECCOLA cards with notes from the companies. The projects are also still on-going, and thus we have not yet been able to conduct formal interviews discussing their ECCOLA use experiences in more detail. As a result, this stage is still on-going as well.

Additionally, ECCOLA has been accepted for presentation in another scientific conference workshop at ICSE2021. This workshop will be held in a similar manner in hopes of further improving the method where needed. Though the development of ECCOLA continues, we feel that we have reached a stage where we wish to share ECCOLA with the scientific community and the industry at large. Given the current lack of methods for AI ethics, with the industry largely reliant on guidelines to implement AI ethics, ECCOLA can serve as a starting point in the area, as we discuss next.

6. Discussion

The ECCOLA method was created to help us bridge the gap between research and practice in the area of AI ethics. Despite the increasing activity in the area, the academic discussion on AI ethics has not reached the industry (Vakkuri et al., 2020). Through ECCOLA, we have attempted to make some of the contents of the IEEE EAD guidelines (IEEE Global Initiative, 2019) and the EU Trustworthy AI guidelines (HLEG, 2019) actionable, alongside other research in the area.

We use the three goals we had for ECCOLA, which we discussed in the Introduction and Section 3, to structure the discussion in this section. These goals were (1) to help create awareness of AI ethics and its importance, (2) to make a modular method suitable for a wide variety of SE contexts, and (3) to make ECCOLA suitable for agile development, while also helping make ethics a part of agile development in general.

In relation to the first goal, there is currently no way of benchmarking what is, so to say, sufficiently ethical in the context of AI ethics. This is arguably a limitation for any such method in the context currently. Benchmarking ethics is difficult and thus it is equally difficult for a method to have a proven effect in a

quantitative manner. Moreover, ethical issues are often context-specific and require situational reflection. This has also been why we have, for now, chosen to focus on raising awareness and highlighting (potential) issues rather than trying to provide direct solutions for ethical questions. Raising awareness has also been a goal of the IEEE EAD initiative (IEEE Global Initiative, 2019). In general, raising awareness is important as AI ethics is a new topic for the industry.

On the other hand, it would be possible to select a specific set of AI ethics guidelines, such as the EU ones (HLEG, 2019), and study whether a tool or a method would help organizations implement those. While ECCOLA is not based on any *one* set of guidelines, the EU guidelines have heavily influenced it, and this is something future studies on ECCOLA should tackle. So far, as ECCOLA is still being iteratively developed further, we have not yet conducted such a study, focusing instead on improving the method before looking to further confirm its usefulness past what we have presented here.

Currently, ECCOLA provides a starting point for implementing ethics in AI. Based on our lessons learned thus far, we argue that ECCOLA facilitates the implementation of AI ethics in two confirmable ways: (1) ECCOLA raises awareness of AI ethics. It makes its users aware of various ethical issues and facilitates ethical discussion within the team. This could be seen on the notes made on the cards we collected from the users of ECCOLA during the different stages of its development, as well as in the discussions and interviews we had with its users. (2) ECCOLA produces transparency of systems development. In utilizing the method, a project team produces documentation of their ethical decision-making by means of e.g., making notes on the note-making space in the cards and non-functional requirements in the product backlog. This could be seen in the notes made on the ECCOLA cards we analyzed while developing ECCOLA.

Transparency is one key issue in AI systems, both in terms of systems and in terms of systems development (Dignum, 2017). These documents, as we have done while testing the method, can also be analyzed to understand how the method was used, aside from seeking to understand the reasoning behind the ethical decisions made during development. Using ECCOLA produces a paper trail of decisions and choices as notes on the cards, alongside other types of written documents such as meeting notes.

So far, we have not utilized control groups while developing ECCOLA, focusing instead on improving the method before aiming to further quantify its effectiveness. We cannot thus argue, based on our data on ECCOLA so far, that ECCOLA would have increased ethical consideration over a baseline of no ethical tool being utilized. On the other hand, we did study the use of the RESOLVEDD strategy in a past paper, which we also briefly discussed here due to its relevance in motivating the development of ECCOLA, and argued that the presence of an ethical tool in general seems to increase ethical consideration (in a student setting). Moreover, out on the field, the baseline largely seems to be that ethical aspects are currently ignored (Vakkuri et al., 2020; Vakkuri et al., 2020). With these studies in mind, we consider it likely that ECCOLA does increase ethical consideration over a baseline of no tool being utilized. However, the effects of ECCOLA on ethical consideration should be further looked into in future studies. This could be done by e.g. studying whether ECCOLA helps fulfill the requirements of one particular set of guidelines, as we have discussed above.

Compared to a baseline where no ethical methods are used, ECCOLA can thus already be argued to increase ethical consideration during development based on this data. This was also the case when we studied student teams using the RESOLVEDD strategy in an existing paper: it increased ethical consideration over the baseline of no ethical tool being used (Vakkuri and Kemell,

Table 4
Participant companies.

Company	Stage	Company description	ECCOLA users
Company A	5&6	small (<30 employees) SW company focusing in Maritime logistic	1 Project owner 2 developers
Company B	5&6	Micro (<10 employees) SW company focusing in data-driven solutions	1 Project owner, 2 developers, 2 consultants
Company C	6	Medium Multinational (>250 employees) SW consulting company	1 Project owner, 2 developers

2019) in a student setting. Out on the field, the baseline largely seems to be that ethical aspects are currently ignored (Vakkuri et al., 2020; Vakkuri et al., 2020). However, the effects of ECCOLA on ethical consideration should be further looked into in future studies. This could be done by e.g. studying whether ECCOLA helps fulfill the requirements of one particular set of guidelines, as we have discussed above.

The second goal has been based on the method-agnostic philosophy of the Essence Theory of Software Engineering (Jacobson et al., 2012). Industry organizations use a wide variety of methods, from out-of-the-box ones to, more commonly, tailored in-house ones (Ghanbari, 2017). ECCOLA is not intended to replace any of these. Rather, ECCOLA is a modular tool that can be added to existing methods and used in conjunction with them, lessening the barrier to its adoption. Though ECCOLA is still being studied in industry settings and we are still collecting data from these cases, so far none of the companies have discussed any issues incorporating ECCOLA into their existing ways-of-working.

This, in turn, leads us to the third goal. As agile development is currently the trend, ECCOLA has been designed to be an iterative process from the get-go. However, during its iterative development, we noticed that a strict iterative process was not a suitable approach due to being too heavy. The users of the method opted out of adhering to the process and used the cards in a modular fashion despite the instructions asking them to repeat the full process every time. Now, ECCOLA is a modular tool by design. Being a card deck, this means that its users are able to select the cards they feel are relevant for each of their iterations, as opposed to having to go through the same process every time. Based on our data, the users of the method prefer this approach, and it seems to work in Agile development as the companies utilizing it are all Agile and have had no issue incorporating it into their way-of-working.

On the other hand, we do not know whether this is detrimental from the point of view of implementing ethics. Do the users of the tool make informed decisions about which cards to exclude? Would advising them to go through a full process (or e.g. all the cards in each iteration in this case) result in more ethical consideration? However, as this is a question of whether ECCOLA helps implement ethics (and to what extent), this is more related to the first goal discussed above.

In designing ECCOLA, we have also turned to VSD (Section 2.3) for some inspiration. First, as already mentioned, we have also chosen a gamified approach in the form of a card deck for ECCOLA. Secondly, both VSD and ECCOLA are iterative methods that can be used in conjunction with SE methods. Thirdly, both methods take on a proactive perspective to ethical consideration in the design or development process. Fourthly, there is some overlap in ethical themes in the methods (e.g., privacy, stakeholder analysis, etc.). On the other hand, they differ in their theoretical backgrounds (SE vs. IS), how ECCOLA is far more focused on the perspective of SE and developers, and how ECCOLA is an AI/S-specific method as opposed to a general design method.

Overall, ECCOLA is intended to become a part of the agile development process in general. Ethics should not be merely an afterthought. Ethics should be another set of non-functional requirements, as well as a part of the user stories for the system. ECCOLA is a tool for developers and product owners. Ethics cannot be outsourced, nor can ethics be implemented by hiring

an ethics expert (Vakkuri et al., 2020). AI ethics should be in the requirements, formulated in a manner also understood by the developers working on the system.

As governments and policy-makers have already begun to regulate AI systems in various ways (e.g., bans on facial recognition for surveillance purposes,⁷ this trend is likely to only accelerate. With more and more regulations imposed on AI systems, organizations will need to tackle various AI ethics issues while developing their systems. This will consequently result in an increasing demand for methods in the area. While this will also inevitably result in the birth of various new methods, developed by companies, scholars, and standardization organizations alike in the future, for the time being ECCOLA can serve as one initial option where there currently are next to none. For the time being, only some commercial methods have already been proposed for AI ethics (e.g.,^{8,9}).

7. Threats to validity

In this section, we discuss the limitations of the study through validity threats. These threats are split into four categories as follows: reliability, construct validity, internal validity and external validity.

7.1. Reliability

First, reliability. The research approach chosen here, action research, on its own already presents threats to reliability. As the research approach influences the research target (organization), changing it and producing unreliability, it is not possible for subsequent studies to carry out the same study in the same context.

We have had separate plans for data collection in each stage. The types of data collected are detailed in Table 3. Most of the data used to develop ECCOLA has either been user notes on ECCOLA cards or unstructured interview data. However, in the later stages while working with companies, we have collected increasing amounts of informal discussion data as e.g. meeting notes.

While collecting data, we have mostly kept our distance as researchers, maintaining a distinct role and doing our best to only collect data while avoiding advising or leading the participants on into any direction. However, in the workshops, academic and company ones, we have occasionally involved ourselves in the group work as facilitators while trying to not provide any answers to the workshop participants. In analyzing our data, we have had multiple researchers (two or three) involved in the analysis process in an attempt to limit researcher error and bias.

Additionally, in action research, an audit trail is recommended by some authors. We would highlight our past publications in the area as one type of audit trail in this regard. We published our results from testing the RESOLVEDD method in the context of AI ethics (Vakkuri et al., 2019), we published an earlier version of ECCOLA in another paper (Vakkuri et al., 2020), and we have studied the gap in the area in existing studies (e.g. Vakkuri et al. (2020) among others).

⁷ <https://www.bbc.com/news/technology-51148501>.

⁸ <https://www.ideo.com/post/ai-ethics-collaborative-activities-for-designers>.

⁹ <https://www.33a.ai/ethics>.

7.2. Construct validity

The construct validity of this study has three primary threats as we see them: 1) the research strategy, 2) the construct of method, and 3) the construct of ethics. Cyclical action research is a typical SE research approach. Additionally, in designing our research strategy, we have utilized existing studies that have proposed methods in SE in designing our strategy in more detail (e.g. Fagerholm et al. (2017)). In terms of data collection and use, we looked at another study that has proposed an Agile method in the past (Abrahamsson et al., 2004). We have described our research strategy in detail in Section 4.

As mentioned in the background section, ethics and values can mean different things to different individuals (Friedman et al., 2013), and different cultures may have different ethical theories. To tackle this potential threat to validity, ECCOLA tries to be agnostic in terms of ethical theories and the definition of ethics. ECCOLA presents potential issues that should be tackled, but leaves it up to the users of the tool to decide on how to tackle them. It asks questions but does not provide the answers directly. Admitted, values such as privacy are not equally important to everyone, and as such ECCOLA does take on a stand to some extent in terms of which AI principles it includes. However, these principles are grounded in existing research and white and gray literature in the area.

Another threat to construct validity is related to the construct of method. Methods in SE describe ways of working. They consist of techniques (IS) (Tolvanen, 1998) or practices (SE) (Jacobson et al., 2012) which together describe how work should be carried out by an organization. Past studies have argued that developers prefer simple and practical methods, if they use any at all (Abrahamsson and Iivari, 2002). Moreover, organizations tend to tailor methods into in-house ones better suited for their specific context (Ghanbari, 2017), which is also something Essence encourages (Jacobson et al., 2012). To make ECCOLA desirable to the industry, we have 1) made it modular to let organizations tailor it, 2) designed it to be used on conjunction with existing SE methods, and 3) to make it more practical. The industry-as-a-lab approach (Potts, 1993) we have used in the later stages of ECCOLA's development is intended to ensure that ECCOLA is practical.

7.3. Internal threats to validity

The main threat to internal validity so far is that we cannot ascertain that ECCOLA produces ethical AI systems, and thus we do not claim that it does. This is not only a challenge in the data we have utilized, but also on a more general level: there are, as far as we know, no benchmarks or measures for ethical AI. On the other hand, we have argued that ECCOLA helps implement AI ethics and produces more ethical consideration during development, compared to a situation where no ethical method is used. Our data indicates that using ECCOLA results in ethical consideration. However, what actions are taken as a result of the ethical consideration is ultimately up to the developers and the organizations.

The wide variety of data we have utilized here presents both internal and external (discussed next) threats to validity, having been collected from different contexts and using different data collection methods. Most of the data we now have on ECCOLA has been collected after influencing the subjects in some way (as opposed to having both before and after data). We wanted to avoid asking questions beforehand so as to not direct the subjects into any particular line of thinking in relation to AI ethics. Instead, we wanted to have our subjects work as usual while additionally utilizing ECCOLA to be able to see how they use the tool. This has,

however, made it difficult to measure any changes in attitudes in the subjects, or any other such changes that could be measured based on data collected both before and after utilizing ECCOLA. To this end, wanting to primarily focus on improving the method based on user experiences, we have not utilized control groups in the earlier stages to further ascertain its impacts.

Aside from what we can say based on our data on the use of ECCOLA, we would also again highlight other ethical tools discussed earlier in this paper, namely the RESOLVEDD strategy (Pfeiffer and Forsberg, 1993) and the Tripartite Method and the associated Envisioning Cards (discussed in Section 2.3). In designing ECCOLA, we have studied these existing approaches for involving ethics in broader business and development contexts, which have been argued to increase ethical consideration, and adopted similar elements as a part of ECCOLA. We would argue that ECCOLA, being founded on these approaches, should have retained some of their effectiveness in increasing ethical consideration when used.

7.4. External threats to validity

As we have utilized a wide variety of data while working on ECCOLA (data from students, companies, conference workshops, and interviews, notes, observation, etc.), these different data collection and analysis approaches present an equally wide variety of potential threats. We have, especially early on, utilized student data from classroom settings. We felt that having students utilize the method in its early stages would still provide us with data on, e.g., whether the AI ethics principles in the method were understandable and whether the process suggested by the method made sense. This let us make even large changes to the method without inconveniencing any industry organization using it, as it was still confined to a student setting. We had a large number of students use the method, giving us ample data to work with early on. However, in this case, the student setting is quite different from an industrial one (e.g. in a student project, the shortcomings of an immature ECCOLA would not result in a project manager getting into trouble).

On the other hand, when working with companies, we have thus far relied on a low number of cases, e.g. 1–3 case projects at a time. Moving forward, we wish to widen the industrial testing (and use) of ECCOLA, but while developing the method, we wanted to get more in-depth feedback from fewer cases to improve the method while working in closer cooperation with the involved parties. This presents a threat to validity as data from a low number of companies makes it less generalizable. We would turn to Eisenhardt (1989) who argues that for novel research areas (in case study research), such a low number of cases can be an acceptable number. While Eisenhardt speaks of case studies in particular, the issue of generalizability is still present in other research approaches as well. Empirical studies in AI ethics are currently few in number, and there seems to be a gap in the area (Vakkuri et al., 2020). In particular, studies on methods such as ECCOLA in the area hardly exist. In this light, we would argue that even a few cases is better than none in moving forward in this novel research area.

8. Conclusions

In this paper, we have presented a method for implementing AI ethics: ECCOLA. It is an approach intended to make AI ethics more practical for developers and organizations. Whereas guidelines can seem abstract to developers, methods are a typical approach to software engineering. To this end, ECCOLA is intended to help organizations develop more ethical AI systems by making AI ethics issues a part of the development process.

The method takes on the form of a card deck, as we discussed in more detail in Section 3. These cards from a modular method which can be tailored according to the use context. For example, one sprint may only feature a handful of cards. The method supports iterative development and can be used in conjunction with existing SE methods. Indeed, ECCOLA is not a novel approach to SE but a tool for better involving AI ethics into the development process, to be used with existing methods.

ECCOLA has been developed iteratively using the Cyclical Action Research approach (Susman and Evered, 1978) and continuous experimentation (Mikkonen et al., 2018). During its development thus far, we have gone through a number of stages, discussed in Sections 4 and 5. In each stage, we have collected data, with a focus on empirical data on the use of ECCOLA. In the process, we utilized both student data and project data from industry projects, as well as feedback from academic workshops. Though ECCOLA is still being developed further, we have reached a state of maturity where we wish to share the method with the scientific community, as well as the industry.

The use of ECCOLA in practice is discussed in Section 3.1 of this paper. The materials for using the method (cards, instructions) can be downloaded from (<https://doi.org/10.6084/m9.figshare.12136308>).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is partially funded by Business Finland (business-finland.fi) research projects: Sea4Value-Fairway & Stroke-Data and ITEA4 (itea4.org/project/mad-work.html) research project: Mad@Work. The authors are grateful for the founders for their support.

References

- Abrahamsson, P., Hanhineva, A., Hulkko, H., Ihme, T., Jaälinoja, J., Korkala, M., Koskela, J., Kyllönen, P., Salo, O., 2004. Mobile-d: An agile approach for mobile application development. In: Companion To the 19th Annual ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages, and Applications. Association for Computing Machinery, New York, NY, USA, pp. 174–175. <http://dx.doi.org/10.1145/1028664.1028736>.
- Abrahamsson, P., Iivari, N., 2002. Commitment in software process improvement - in search of the process. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences. pp. 3239–3248. <http://dx.doi.org/10.1109/HICSS.2002.994403>.
- Ananny, M., Crawford, K., 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* 20 (3), 973–989. <http://dx.doi.org/10.1177/1461444816676645>.
- Biffi, S., Aurum, A., Boehm, B., Erdogmus, H., Grünbacher, P., 2006. *Value-Based Software Engineering*. Springer Science & Business Media.
- Davis, J., Nathan, L.P., 2015. Value sensitive design: Applications, adaptations, and critiques. In: *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. Dordrecht: Springer Netherlands, pp. 11–40.
- Davison, R., Martinsons, M.G., Kock, N., 2004. Principles of canonical action research. *Inf. Syst. J.* 14 (1), 65–86. <http://dx.doi.org/10.1111/j.1365-2575.2004.00162.x>.
- Dignum, V., 2017. Responsible autonomy. arXiv preprint [arXiv:1706.02513](https://arxiv.org/abs/1706.02513).
- van der Duin, P., 2019. Toward “responsible foresight”: Developing futures that enable matching future technologies with societal demands. *World Futur. Rev.* 11 (1), 69–79.
- Eisenhardt, K.M., 1989. Building theories from case study research. *Acad. Manag. Rev.* 14 (4), 532–550. <http://dx.doi.org/10.2307/258557>.
- Fagerholm, F., Sanchez Guinea, A., Mäenpää, H., Münch, J., 2017. The RIGHT model for continuous experimentation. *J. Syst. Softw.* 123, 292–305. <http://dx.doi.org/10.1016/j.jss.2016.03.034>.
- Friedman, B., Kahn, P., Borning, A., 2002. *Value Sensitive Design: Theory and Methods*. Tech. rep., University of Washington technical report.
- Friedman, B., Kahn, P.H., Borning, A., 2008. Value sensitive design and information systems. In: *The Handbook of Information and Computer Ethics*. Wiley Online Library, pp. 69–101.
- Friedman, B., Kahn, P.H., Borning, A., Hultgren, A., 2013. Value sensitive design and information systems. In: *Early Engagement and New Technologies: Opening Up the Laboratory*. Springer, pp. 55–95.
- Ghanbari, H., 2017. *Investigating the Causal Mechanisms Underlying the Customization of Software Development Methods* (Ph.D. thesis). University of Jyväskylä.
- Gotterbarn, D., Brinkman, B., Flick, C., Kirkpatrick, M.S., Miller, K., Vazansky, K., Wolf, M.J., 2018. *AcM code of ethics and professional conduct*. 2019, (18.3.). <https://www.acm.org/code-of-ethics>.
- Hagendorff, T., 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* 1–22.
- HLEG, 2019. *Ethics Guidelines for Trustworthy AI*. EU, <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
- IEEE Global Initiative, 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>.
- Jacobson, I., Ng, P.-W., McMahon, P.E., Goedicke, M., et al., 2019. *The Essentials of Modern Software Engineering: Free Practices from the Method Prisons!*. Morgan & Claypool.
- Jacobson, I., Ng, P.-W., McMahon, P.E., Spence, I., Lidman, S., 2012. The essence of software engineering: the SEMAT kernel. *Commun. ACM* 55 (12), 42–49.
- Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1 (9), 389–399.
- Johnson, B., Smith, J., 2021. Towards ethical data-driven software: filling the gaps in ethics research & practice. In: *2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics)*. pp. 18–25. <http://dx.doi.org/10.1109/SEthics52569.2021.00011>.
- Leikas, J., Koivisto, R., Gotcheva, N., 2019. Ethical framework for designing autonomous intelligent systems. *J. Open Innov. Technol. Mark. Complex.* 5 (1), 18.
- Manders-Huits, N., 2011. What values in design? The challenge of incorporating moral values into design. *Sci. Eng. Ethic.* 17 (2), 271–287.
- McNamara, A., Smith, J., Murphy-Hill, E., 2018. Does acm’s code of ethics change ethical decision making in software development? In: *Proceedings of the 2018 26th ACM ESEC/FSE*. In: *ESEC/FSE 2018*, ACM, New York, NY, USA, pp. 729–733. <http://dx.doi.org/10.1145/3236024.3264833>.
- Mikkonen, T., Lassenius, C., Männistö, T., Oivo, M., Järvinen, J., 2018. Continuous and collaborative technology transfer: Software engineering research with real-time industry impact. *Inf. Softw. Technol.* 95, 34–45. <http://dx.doi.org/10.1016/j.infsof.2017.10.013>.
- Mittelstadt, B., 2019. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* 1–7.
- Morley, J., Floridi, L., Kinsey, L., Elhalal, A., 2019. From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. arXiv preprint [arXiv:1905.06876](https://arxiv.org/abs/1905.06876).
- Pfeiffer, R.S., Forsberg, R.P., 1993. *Ethics on the Job: Cases and Strategies*. Wadsworth Publishing Company.
- Pichai, S., 2018. AI at google: our principles. Accessed: 2021-04-30, <https://www.blog.google/technology/ai/ai-principles/>.
- Potts, C., 1993. Software-engineering research revisited. *IEEE Softw.* 10 (5), 19–28. <http://dx.doi.org/10.1109/52.232392>.
- Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1 (5), 206–215.
- Rusanen, A.-M., Nurminen, J., Raisanen, S., Tarkoma, S., Halmetoja, S., 2021. *Ethics-of-ai mooc*. Accessed: 2021-04-30, <https://ethics-of-ai.mooc.fi/>.
- Shilton, K., 2018. Values and ethics in human-computer interaction. *Found. Trends Human Comput. Int.* 12 (2).
- Susman, G.I., 1976. *Autonomy At Work: A Sociotechnical Analysis of Participative Management*. Praeger, New York.
- Susman, G.I., Evered, R.D., 1978. An assessment of the scientific merits of action research. *Adm. Sci. Q.* 582–603.
- Tolvanen, J.-P., 1998. In: Tolvanen, J.-P. (Ed.), *Incremental Method Engineering with Modeling Tools: Theoretical Principles and Empirical Evidence* (Ph.D. thesis). In: *Jyväskylä studies in computer science, economics and statistics*, University of Jyväskylä.
- Turilli, M., Floridi, L., 2009. The ethics of information transparency. *Ethics Inf. Technol.* 11 (2), 105–112. <http://dx.doi.org/10.1007/s10676-009-9187-9>.
- Vakkuri, V., Kemell, K.-K., 2019. Implementing AI ethics in practice: An empirical evaluation of the RESOLVEDD strategy. In: *International Conference on Software Business*. Springer, pp. 260–275.
- Vakkuri, V., Kemell, K.-K., Abrahamsson, P., 2019. 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)*. pp. 46–50. <http://dx.doi.org/10.1109/SEAA.2019.00015>.

- Vakkuri, V., Kemell, K.K., Abrahamsson, P., 2020. Eccola - a method for implementing ethically aligned AI systems. In: 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). pp. 195–204. <http://dx.doi.org/10.1109/SEAA51224.2020.00043>.
- Vakkuri, V., Kemell, K.-K., Jantunen, M., Abrahamsson, P., 2020. “This is just a prototype”: How ethics are ignored in software startup-like environments. In: Stray, V., Hoda, R., Paasivaara, M., Kruchten, P. (Eds.), *Agile Processes in Software Engineering and Extreme Programming*. Springer International Publishing, Cham, pp. 195–210.
- Vakkuri, V., Kemell, K., Kultanen, J., Abrahamsson, P., 2020. The current state of industrial practice in artificial intelligence ethics. *IEEE Softw.* 37 (4), 50–57.
- VSD Lab, 2021. Value sensitive design lab. Accessed: 2021-01-05, <https://vsdesign.org/>.
- Winkler, T., Spiekermann, S., 2018. Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics Inf. Technol.* 1–5.