cii Student Papers 2022

Karlsruhe Institute of Technology
aifb
CRITICAL INFORMATION INFRASTRUCTURES RESEARCHGROUP

cii Student Papers - 2022

Research Group Critical Information Infrastructures (cii)
Karlsruhe Institute of Technology
Department of Economics and Management
Institute of Applied Informatics and Formal Description Methods

Web: cii.aifb.kit.edu

Corresponding Editor:
Prof. Dr. Ali Sunyaev
Kaiserstr. 89
76133 Karlsruhe, Germany
Phone: +49 721 608-43679
Email: sunyaev@kit.edu

DOI: 10.5445/IR/1000150078

Editorial

Critical information infrastructures are sociotechnical systems comprising essential software components and information systems with a pivotal impact on individuals, organizations, governments, economies, and society. Our research is grounded in the intersection of information systems and computer science, accounting for the multifaceted use contexts of information technology (IT). We employ a variety of interdisciplinary methods and build on theories from computer science, information systems, and related disciplines with research on human behavior affecting information systems and vice versa. This enables us to rigorously generate strong theoretical insights while simultaneously producing research outputs with auspicious value propositions for organizations. Our research group works on practice- and research-driven challenges concerned with the design, development, and evaluation of reliable information systems. Our research features a strong focus on the Internet and healthcare industries as well as on the industry-specific application of trustworthy artificial intelligence (AI) models. The principal goal of our research is theorizing on and designing the applications and methods required for the creation and innovation of sociotechnical systems.

Every year, our research group supervises more than 150 bachelor and master students during their studies at the KIT. To ensure that our students receive a compelling, engaging, and fruitful learning experience, we apply inquiry-based learning methods and actively introduce our research topics to them in various seminars and lectures. There, we give them not only the chance to work on emerging topics in practice and research but also provide them with a look into how we conduct our research and inquiry-based learning. During our courses, students mostly work in groups and deal with problems and issues related to sociotechnical challenges in the realm of (critical) information systems. Topics generally correspond to what we are currently researching. In addition, we allow students to propose their own research topics or even conduct their studies in collaboration with small, medium, or large companies.

Changing from semester to semester topics offered include privacy risks while using disruptive health information systems (Gojka et al., 2021), security concerns regarding data in the cloud, fog and edge services (Lins et al., 2018), conceptual ambiguity concerning gamification and serious games in healthcare (Warsinsky, Schmidt-Kraepelin, Rank, et al., 2021), designing and implementing requirements for distributed ledgers (Kannengießer et al., 2020), and adoption and trust concerns regarding the use of AI in autonomous vehicles (Renner et al., 2021). Our research team assists students throughout the research process, including helping them identify and organize problems, consistently apply appropriate research methods, develop and communicate possible solutions, and write research papers.

Engaging students in everyday academic work and bringing research to students offers many benefits, not just for students but also for our research group, for the research community, and for practice in general. Students get involved in timely practice problems that research is trying to solve. Students can better understand the theoretical foundations learned in previous lectures and apply these through knowledge gained in the seminars and lectures. By offering such scientific courses, students can gain first-hand experience or a deeper understanding of the writing of upcoming bachelor and master theses.

In fact, most students continue their research after completing a seminar or lecture, either in the form of a dissertation, as part of their work as a student assistant in our research group, or even voluntarily in their free time. Students also regularly engage with organizations, which can lead to initial collaborations and even pave the way for upcoming jobs. Outstanding student works often come from our seminars and lectures. Unfortunately, in the past, great student works have often disappeared into drawers despite the disruptive and groundbreaking insights students have come up with.

As a research group, we always appreciate the work of great students, incorporate their insights into our own research projects, and often publish exceptional results in conference proceedings or journals (e.g., Gräbe et al., 2020; Lins et al., 2021; Ohagen et al., 2022; Petry et al., 2021; Renner et al., 2020; Schmidt-Kraepelin, Warsinsky, et al., 2020; Toussaint, Thiebes, et al., 2022).

Therefore, we conceived the idea for this book, offering students the possibility to publish their excellent works in this dedicated miscellany. We are very pleased to present this collection for the second time after last year (Sunyaev et al., 2021), bringing together the best student works from 2021 and the first half of

2022. Contributions in this book come from four different courses that provide students with a broad range of topics related to (critical) information systems:

Emerging Trends in Internet Technologies:

The seminar *Emerging Trends in Internet Technologies* aims at providing students with insights into current topics in the field of information systems while mainly focusing on fundamental and innovative Internet technologies. Students are offered a selection of topics around the lectures and present research of our group, including distributed ledger technology (Kannengießer et al., 2020), cloud, fog, and edge computing (Lins et al., 2021; Renner et al., 2020), AI (Renner et al., 2022; Thiebes, Lins, et al., 2020), security, and privacy (Yari et al., 2021). For example, our research group gave a profound conceptualization of the phenomena of *Artificial Intelligence as a Service* due to the lack of conceptual clarity in research and practice (Lins et al., 2021).

Emerging Trends in Digital Health:

Similarly, the seminar *Emerging Trends in Digital Health* aims to provide insights into current topics in the field of information systems with a focus on innovative digital healthcare systems. Kicking off with a short introduction and corresponding topics, students can choose to work on many different topics around the lectures and research topics of the research group, including genomics (Thiebes, Toussaint, et al., 2020; Toussaint, Renner, et al., 2022), distributed l edger technology (Beyene et al., 2022), patient consent (Beyene et al., 2019), AI (Pandl, Feiland, et al., 2021; Thiebes, Lins, et al., 2020), and gamification in healthcare (Schmidt-Kraepelin, Toussaint, et al., 2020). An example of our interdisciplinary work in this field is a recent scoping review on the use of distributed ledger technology in the field of genomics, where we investigate how blockchain and other distributed ledger systems are currently or could be used to store and share genomic data between multiple entities such as patients, doctors, hospitals, researchers, and many more (Beyene et al., 2022).

Digital Health:

The course *Digital Health* introduces master students to the subject of digitization in healthcare. Students learn about the theoretical foundations and practical implications of various topics surrounding digitization in healthcare, including health information systems, telematics, big healthcare data, and patient-centered healthcare (e.g., Pandl, Thiebes, et al., 2021; Rädsch et al., 2021; Thiebes, Schlesner, et al., 2020; Warsinsky, Schmidt-Kraepelin, Thiebes, et al., 2021). After an introduction to the challenge of digitization in healthcare, the following sessions focus on an in-depth exploration of selected cases that represent current challenges in research and practice. Students work in groups of three to four on specific topics and must write a course paper. One current topic our research group investigates in this field is how and why consumers perceive certain direct-to-consumer genetic testing business models as fair or unfair, respectively (Toussaint, Thiebes, et al., 2022).

Critical Information Infrastructures:

The course *Critical Information Infrastructures* introduces students to the world of complex sociotechnical systems that permeate societies on a global scale. Being offered every winter term, master students learn to handle the complexities involved in the design, development, operation, and evaluation of critical information infrastructures. At the beginning of the course, critical information infrastructures are introduced on a general level (Dehling et al., 2019). The following sessions focus on an in-depth exploration of selected cases that represent current challenges in research and practice. Students work in groups of four on specific topics and must write a course paper. The research group has also published a book chapter providing a discussion on the characteristics and challenges of critical information infrastructures (Dehling et al., 2019).

Selected Issues on Critical Information Infrastructures: Collaborative Development of Innovative Teaching Concepts

The COVID-19 pandemic had and still has a significant impact on the teaching at KIT, as a radical shift from face-to-face to online teaching had to be accomplished at short notice. While online teaching has meanwhile become more and more successful, the pandemic has highlighted an important need: innovative, digital teaching concepts are required that can be used in times of remote but also face-to-face teaching. We at the cii research group have therefore set ourselves the goal of rethinking our teaching concepts and techniques to adapt to changing needs and offer our students an insightful and engaging learning experience. Hence, we decided to offer a groundbreaking course since the winter term of 2021. During the course, students collaboratively developed innovative, digital and scientifically based teaching concepts while being supported by researchers from our research group. Students had the opportunity to contribute their own ideas and conceptions for "Teaching 4.0" at KIT and to integrate them into an innovative teaching concept. Existing teaching units (e.g., basic lectures, seminars, or ILIAS learning modules) were used as examples and reflected prototypically. More modern teaching concepts, such as interactive learning modules, gamification, serious gaming, or flipped classrooms, were examined and critically discussed with regard to their applicability and usefulness. Students were also advised to think of completely new teaching concepts based on their experience or their wishes, thereby disrupting the traditional thinking of teaching at the KIT. Students based their concepts on existing research or scholarly theories to ensure an effective and pedagogically valuable teaching experience. The developed teaching concepts were further tested in a live session during the course, providing a first proof-in-concept and valuable feedback to the students. Not only our students really enjoyed participating in this unique course and were excited to better understand how teaching feels at the KIT, but we also valued learning what students desire from (post)-pandemic teaching experience. More importantly, the course resulted in promising teaching concepts, which the research group is currently trying to apply in teaching practice to improve our teaching for the future. Given the large success of the course, the research group has also decided to offer the selected issues course once again in the summer term of 2022 and winter term 2022/2023.

Out of these courses, we selected the student works that represent the best and most interesting studies. The student works in this book cover a wide range of research problems, including an overview of opportunities and risks for the adoption of privacy-preserving machine learning fairness in healthcare; how to certify AI regarding fairness; security threats for vehicular fog computing; fairness, accountability, transparency, and explainability in intelligent automation while looking at the use case autonomous vehicles; trade-offs in AI as a Service; and innovative teaching methods like continuous learning and flipped classrooms.

- Böck et al. investigate the drivers and inhibitors for the adoption of privacy-preserving machine learning in organizations. While conducting interviews, they provide insights into the current state of developments in research, politics, and industry.
- Özdemir et al. performed a literature review and expert interviews to understand how to certify fairness of AI-embedded systems. They show that three key structural building blocks for fairness certifications include fairness criteria to be assessed (certification content), potential issuers and auditors (certification sources), and auditing methods to evaluate fairness (certification process).
- The study by Erb et al. examines security threats for vehicular fog computing contexts and conducts a ranking-type Delphi study among non-domain experts to identify, select and rank relevant threats. Their results indicate that privacy and safety-related security threats are considered the most dangerous regarding vehicular fog computing.
- Braun et al. conducted interviews to understand which properties are important for autonomous vehicles to generate trust. By building on the FATE properties (Fairness, Accountability, Transparency, and Explainability) by Shin (2020) they show that the trust-building in autonomous vehicles consists of many different properties.
- Jestram et al. reveal in their interview study trade-offs that follow specific design decisions for AI as a Service. They show that these novel cloud-based AI services share similar service design trade-offs of common cloud services but also highlight novel challenges that require further attention.

- The study by Eipper et al. performed a literature research to provide an overview of the current state of research in the emerging research field on fairness of medical AI. While developing AI, they show that a holistic perspective is needed to optimize the bias of data and system problems.
- Appenzeller et al. iteratively developed the new concept "Continuous Learning and Applying (KoLA)" based on a flipped classroom design. It is composed of alternating in-class and out-of-class sessions to engage students in a meaningful way.
- Wiegand et al. present a new teaching concept based on the flipped classroom design while including a combination of online lecture videos with regular interactive "bootcamp" meetings to consolidate what has been learned.

In closing this brief overview, we are grateful that students have once again taken time off to further review and improve their papers to ensure the high quality of this miscellany. In addition to the students who wrote the articles, this book would not have been possible without the dedicated researchers of our research group, who supervised the students during the course. We would like to take this opportunity to express our sincere appreciation for their active support, motivation, and commitment to all student works in the cii research group. We are committed to our mission of teaching excellence and intend to publish each year the best work in a compendium, striving to bring the scientific work closer to the students.

Sincerely,

Ali Sunyaev, Maximilian Renner, Philipp A. Toussaint, Scott Thiebes, Sebastian Lins

Miscellany Team 2022

Prof. Dr. Ali Sunyaev
Editor-in-Chief

Maximilian Renner
Editor

Philipp A. Toussaint
Editor

Scott Thiebes
Editor

Dr. Sebastian Lins
Editor

Supervising Research Associates

Jan Bartsch | Mikael Beyene | Mandy Goram | Dr. Malte Greulich | Shanshan Hu | David Jin | Niclas Kannengießer| Florian Leiser | Dr. Sebastian Lins | Felix Morsbach | Konstantin D. Pandl | Sascha Rank | Maximilian Renner | Manuel Schmidt-Kraepelin | Dr. Benjamin Sturm | Heiner Teigeler | Scott Thiebes | Philipp A. Toussaint | Simon Warsinsky

CRITICAL INFORMATION INFRASTRUCTURES RESEARCHGROUP

References

Beyene, M., Thiebes, S., & Sunyaev, A. (2019). *Multi-Stakeholder Consent Management in Genetic Testing: A Blockchain-Based Approach* Proceedings of the Pre-ICIS SIGBPS 2019 Workshop on Blockchain and Smart Contracts, München, Deutschland.

Beyene, M., Toussaint, P. A., Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2022). A scoping review of distributed ledger technology in genomics: thematic analysis and directions for future research. *J Am Med Inform Assoc*, *29*(8), 1433-1444. https://doi.org/10.1093/jamia/ocac077

Dehling, T., Lins, S., & Sunyaev, A. (2019). Security of critical information infrastructures. In *Information Technology for Peace and Security : IT Applications and Infrastructures in Conflicts, Crises, War, and Peace Hrsg.: C. Reuter* (pp. 319-339). Springer Vieweg. https://doi.org/10.1007/978-3-658-25652-4_15

Gojka, E.-E., Kannengießer, N., Sturm, B., Bartsch, J., & Sunyaev, A. (2021). *Security in Distributed Ledger Technology: An Analysis of Vulnerabilities and Attack Vectors* Advances in Intelligent Systems and Computing, London, Vereinigtes Königreich.

Gräbe, F., Kannengießer, N., Lins, S., & Sunyaev, A. (2020). *Do Not Be Fooled: Toward a Holistic Comparison of Distributed Ledger Technology Designs* Hawaii International Conference on System Sciences 2020, Maui, Hawaii, January 7 - 10, 2020, Grand Wailea, Maui, Hawaii. http://dx.doi.org/10.24251/HICSS.2020.770

Kannengießer, N., Pfister, M., Greulich, M., Lins, S., & Sunyaev, A. (2020). *Bridges Between Islands: Cross-Chain Technology for Distributed Ledger Technology* Hawaii International Conference on System Sciences 2020, Maui, Hawaii, January 7 - 10, 2020, Grand Wailea, Maui, Hawaii. http://dx.doi.org/10.24251/HICSS.2020.652

Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2021). Artificial Intelligence as a Service – Classification and Research Directions [journal article]. *Business & information systems engineering*. https://doi.org/10.1007/s12599-021-00708-w

Lins, S., Schneider, S., & Sunyaev, A. (2018). Trust is Good, Control is Better: Creating Secure Clouds by Continuous Auditing. *IEEE Transactions on Cloud Computing*, *6*(3), 890-903. https://doi.org/10.1109/tcc.2016.2522411

Ohagen, P., Lins, S., Thiebes, S., & Sunyaev, A. (2022). *Using ChatOps to Achieve Continuous Certification of Cloud Services* 55th Hawaii International Conference on System Sciences (HICSS), Online.

Pandl, K. D., Feiland, F., Thiebes, S., & Sunyaev, A. (2021). *Trustworthy machine learning for health care: scalable data valuation with the shapley value* CHIL '21: Proceedings of the Conference on Health, Inference, and Learning, April, 2021. Ed.: M. Ghassemi, Online. http://dx.doi.org/10.1145/3450439.3451861

Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2021). How detection ranges and usage stops impact digital contact tracing effectiveness for COVID-19. *Sci Rep*, *11*(1), 9414. https://doi.org/10.1038/s41598-021-88768-6

Petry, L., Lins, S., Thiebes, S., & Sunyaev, A. (2021). Technologieauswahl im DigitalPakt: Wie werden Entscheidungen im Bildungssektor getroffen? = Technology selection in the german digital pact: How are decisions made in the educational system? [journal article]. *HMD*. https://doi.org/10.1365/s40702-021-00751-x

Rädsch, T., Eckhardt, S., Leiser, F., Pandl, K. D., Thiebes, S., & Sunyaev, A. (2021). *What Your Radiologist Might be Missing: Using Machine Learning to Identify Mislabeled Instances of X-ray Images* Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS), Online.

Renner, M., Lins, S., Söllner, M., Thiebes, S., & Sunyaev, A. (2021). *Achieving Trustworthy Artificial Intelligence: Multi-Source Trust Transfer in Artificial Intelligence-capable Technology* Building Sustainability and Resilience with IS: A Call fpr Action : ICIS 2021 proceedings ; 42nd International Conference on Information Systems (ICIS), Austin, TX, USA.

Renner, M., Lins, S., Söllner, M., Thiebes, S., & Sunyaev, A. (2022). *Understanding the Necessary Conditions of Multi-Source Trust Transfer in Artificial Intelligence* 55th Hawaii International Conference on System Sciences (HICSS), Online.

Renner, M., Münzenberger, N., Hammerstein, J. v., Lins, S., & Sunyaev, A. (2020). *Challenges of Vehicle-to-Everything Communication. Interviews among Industry Experts* 15th International Conference on Wirtschaftsinformatik (WI2020), Potsdam, 08.-11. März 2020, Potsdam, Deutschland. http://dx.doi.org/10.30844/wi_2020_r12-renner

Schmidt-Kraepelin, M., Toussaint, P. A., Thiebes, S., Hamari, J., & Sunyaev, A. (2020). Archetypes of Gamification: Analysis of mHealth Apps. *JMIR Mhealth Uhealth*, *8*(10), e19280. https://doi.org/10.2196/19280

Schmidt-Kraepelin, M., Warsinsky, S., Thiebes, S., & Sunyaev, A. (2020). *The Role of Gamification in Health Behavior Change: A Review of Theory-driven Studies* Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS 2020), Maui, Hawaii, Grand Wailea, Maui, Hawaii.

Shin, D. (2020). User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541-565. https://doi.org/10.1080/08838151.2020.1843357

Sunyaev, A., Renner, M., Toussaint, P. A., Thiebes, S., & Lins, S. (2021). Editorial on cii Student Papers - 2021. In A. Sunyaev, M. Renner, P. A. Toussaint, S. Thiebes, & S. Lins (Eds.), *cii Student Papers - 2021*. Karlsruher Institut für Technologie (KIT). https://doi.org/10.5445/IR/1000138902

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence [journal article]. *Electronic markets*. https://doi.org/10.1007/s12525-020-00441-4

Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2020). Distributed Ledger Technology in genomics: a call for Europe [journal article]. *European journal of human genetics*, *28*, 139-140. https://doi.org/10.1038/s41431-019-0512-4

Thiebes, S., Toussaint, P. A., Ju, J., Ahn, J. H., Lyytinen, K., & Sunyaev, A. (2020). Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing. *J Med Internet Res*, *22*(1), e14890. https://doi.org/10.2196/14890

Toussaint, P. A., Renner, M., Lins, S., Thiebes, S., & Sunyaev, A. (2022). Direct-to-Consumer Genetic Testing in Social Media: Analysis of YouTube Users' Comments. *JMIR Preprints*. https://doi.org/10.2196/preprints.38749

Toussaint, P. A., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2022). Perceived fairness of direct-to-consumer genetic testing business models. *Electronic markets*. https://doi.org/10.1007/s12525-022-00571-x

Warsinsky, S., Schmidt-Kraepelin, M., Rank, S., Thiebes, S., & Sunyaev, A. (2021). Conceptual Ambiguity Surrounding Gamification and Serious Games in Health Care: Literature Review and Development of Game-Based Intervention Reporting Guidelines (GAMING). *J Med Internet Res*, *23*(9), e30390. https://doi.org/10.2196/30390

Warsinsky, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2021). *Are Gamification Projects Different? An Exploratory Study on Software Project Risks for Gamified Health Behavior Change Support Systems* Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, Online. http://dx.doi.org/10.24251/HICSS.2021.159

Yari, I. A., Dehling, T., Kluge, F., Eskofier, B., & Sunyaev, A. (2021). *Online at Will: A Novel Protocol for Mutual Authentication in Peer-to-Peer Networks for Patient-Centered Health Care Information Systems* Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS 2021), Online.

Table of Contents

Fairness, Accountability, Transparency, and Explainability in Intelligent Automation

Emerging Trends in Internet Technologies, Summer Term 2021

Bastian Franz Braun
Master Student
Karlsruhe Institute of Technology
bastian.franz.braun@gmail.com

Maren Cordts
Bachelor Student
Karlsruhe Institute of Technology
maren.cordts@hotmail.de

David Siebeneich
Master Student
Karlsruhe Institute of Technology
david.siebeneich@gmail.com

Abstract

***Background*:** *The focus of this thesis is to find out which properties are important for autonomous vehicles to generate trust. Reference is made to the concept of FATE properties (Fairness, Accountability, Transparency, and Explainability) by Shin (2020).*

Objective: *The aim is to identify the extent to which such a FATE framework helps build trust, the extent to which the FATE attributes are already being applied in practice, and the measures that can be used to promote the FATE attributes.*

Methods: *As part of the research, six experts from research and industry were interviewed in qualitative, semi-structured interviews. For the qualitative analysis of the interviews, the Inductive and the Deductive-Inductive Hybrid Approach of Mayring is used.*

Results: *The results show that individual FATE properties are considered in the development, but that the framework in this constellation is not known to the experts and is not applied in this composition. Rather, the terms are often used in an undifferentiated manner. In principle, the experts see the potential for application in the FATE framework. Especially since compliance with each aspect promotes trust.*

Conclusion: *It is important here not only to pay attention to the AI and the algorithms behind it but to apply the framework to the entire automotive development. In addition, through the analysis of the interviews, several measures could be identified that promote compliance with the individual FATE properties and thus foster trust in the autonomous vehicle.*

Keywords: Autonomous driving, FATE, Fairness, Accountability, Transparency, Explainability, Intelligent automation

Introduction

In Germany, approximately 48.2 million passenger cars are registered (as of 01.01.2021). Assuming an average operating time of 45 minutes per passenger car, this results in 36.15 million hours of driving time

every day (Nobis & Kuhnimhof, 2018). In addition, persons with mentally or physically limiting illnesses are not suitable to drive a motor vehicle (BGBl, 2010).

The application of autonomous vehicle pilots can, firstly, make the driving time freely available to the occupants and, secondly, give people who are not fit to drive a motor vehicle the opportunity to drive a passenger car without another person who is fit to drive. Moreover, autonomous systems are believed to have a positive impact on road safety (Bonnefon et al., 2016) and environmental footprint (Nastjuk, 2020). Furthermore, the application of AI, in general, represents a billion-dollar business opportunity (Oxborough et al., 2018). To capture this business opportunity, new machine learning approaches are being used to greatly advance commercialization (Tian et al., 2018).

One of these approaches is exemplified by Tesla Autopilot. A driver assistance system built into electric cars manufactured by Tesla Inc. The system provides numerous autonomous functions (Tesla, 2021). For this, Tesla Autopilot relies on computing hardware that processes sensor data using machine learning approaches. In addition to onboard processing, Tesla vehicles constantly transmit data resulting from driving and using the Autopilot function to Tesla, where it is used to train and improve the algorithms (Fehrenbacher, 2015). The current generation of vehicles forms a network where cars learn from each other, and each driver contributes to the training of an AI that will enable future generations of cars to drive fully autonomously. This is just one example of a vehicle that is designed for level five automation and thus transports its occupants without a driver.

Problem Definition

Despite these prospects, autonomous driving vehicles are currently not publicly accepted. In addition, there are considerable uncertainties regarding the technical feasibility of road safety and open legal questions, such as liability and responsibility in the event of a road accident (Schreurs & Steuwer, 2015; Beiker 2012; Awad et al., 2018). According to over 60% of decision-makers in surveyed companies, the use of AI will hurt stakeholder trust towards the company (Oxborough et al., 2018). In an accident situation, there is a critical moment for the driver to transfer his decision-making power to the autonomous system. Since life may be at stake in the accident situation, this is a sore point through which the trustworthiness of the autonomous system is highly questioned. The algorithm can act according to two different priorities. Either it decides for the greater good or to protect the driver. The majority favors letting the algorithm act for the greater good (Bonnefon et al., 2016). However, the majority would not choose to purchase such a vehicle, making it controversially difficult to market an autonomous vehicle.

Finally, the practical problem is that there is no social acceptance of autonomous vehicles. As a result, the advantages of autonomous driving for the individual and society are not realized, even though it would be technically possible.

To solve this practical problem, ways are sought to get the driver to build trust. The research question addresses which factors are important for the driver to develop trust towards the autonomous system. In research, the previous results on the problem are difficult to apply. Despite many existing studies, there is no overview of applied concepts in business and research. Many terms for "trust" are used indiscriminately.

Objective of the Work

To specifically address the issue, interviews with experts from research and industry will be conducted. This will create a knowledge base on which properties are important for autonomous vehicles to generate trust. The concept of FATE properties is used as a basis. It is of particular interest to find out to what extent and in what form the concept known from research has found application in business. The interview guide aims to allow as much of the interviewee's thought processes as possible. This ensures that it emerges whether the FATE concept has possibly already found application in business, even if the terms are not used in a completely differentiated manner as in research. Subsequently, it will be derived which measures can be taken to realize the FATE properties and thereby create trust towards autonomous driving.

Structure of the Work

First, the reader is introduced to the theoretical foundations of FATE properties in the context of autonomous driving. Then, the procedure of planning and conducting the expert interviews is described. In

the subsequent evaluation, the qualitative content analysis according to Mayring is used. The evaluation of the results is then followed by a critical analysis of the expert interviews. Here, especially the comparison of the application of the concept in business and research is an interesting aspect. Finally, the central findings, as well as conclusions for further research, are summarized as an outlook.

Theoretical Foundations

This chapter introduces the topics of "Autonomous Driving" and "Trust and Adoption of AI Technologies" as well as the concept of so-called FATE (Fairness, Accountability, Transparency, and Explainability) properties in intelligent automation.

Autonomous Driving

The topic of "autonomous driving" is repeatedly the focus of the media and various specialist articles. The term is often used in an undifferentiated manner. This shows that there is no uniform understanding of the term (Maurer, 2015). The degree of automation of a vehicle is often described in terms of the division of tasks between the driver and the system. A basic distinction is made between assisting systems and automating systems. Assisting systems to support the driver but do not take over the driving task of the vehicle. Automated systems depend on their maturity, take over the driving task independently in certain areas of application (Uhr, 2016).

Due to the difficulty of differentiating and the fluid transition between assisting and automated systems, the German Association of the Automotive Industry (VDA, 2015) has agreed on a six-stage scheme. The six degrees of vehicle automation, as shown in Figure 1, are briefly explained below.

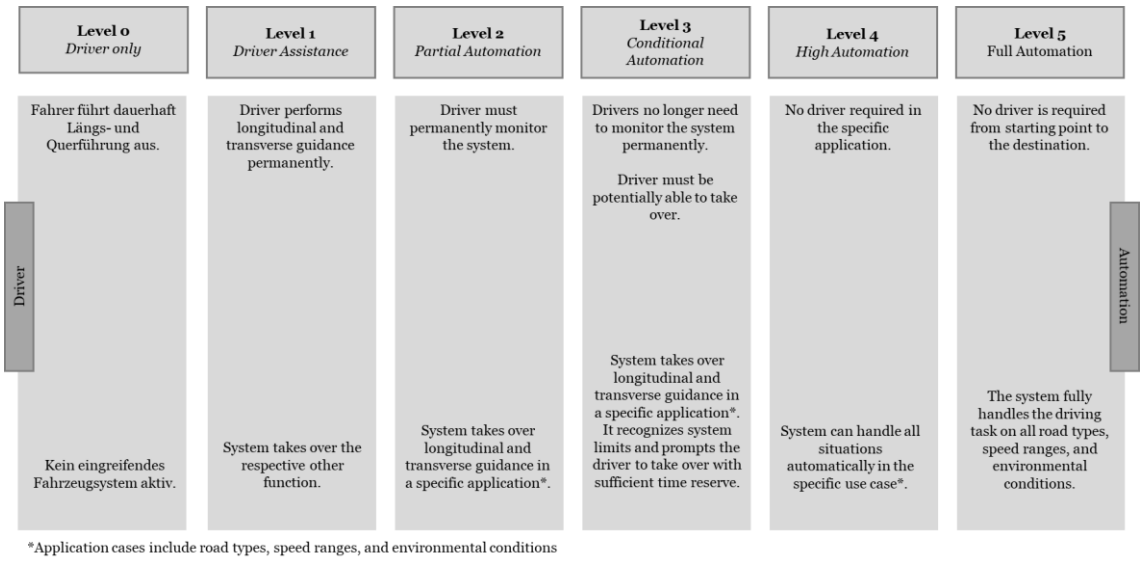

Figure 1: Stages of Autonomous Driving, Based on VDA (2015)

Level 0 describes the state in which *no automation* is used. There are no intervening systems, only warning systems. The driver, therefore, drives the vehicle himself. *Level 1* is *assisted driving*. Here, assistance systems support certain driving tasks of longitudinal or lateral guidance. However, the driver has the responsibility to monitor the driving tasks. Common examples are parking assistance and cruise control (VDA, 2015). The difference between *Level 2 (partially automated driving)* and *Level 3 (highly automated driving)* is that at Level 2, the driver is designated in the system as a permanent supervisor and must be able to take over at any time, whereas at Level 3, the driver is not stopped for permanent supervision. This allows the driver to engage in non-driving activities (Uhr, 2016). However, he must take over vehicle control again in the event of system limits (VDA, 2015). From *Level 4, fully automated driving,* driver interaction is no longer necessary in specific use cases. The passenger car can also cover longer distances without

intervention, and the system recognizes its limits in time to reach a safe state under the rules (Uhr, 2016). *Fully autonomous driving* is described in *Level 5*. Here, the vehicle moves completely on its own without the influence of a human. The driver thus becomes a passenger without driving tasks (VDA, 2015).

To create a certain convergence in the understanding of the term autonomous driving, the term autonomous driving is defined in this paper as the two highest degrees of expression. This definition is based on the definition of the Federal Highway Research Institute of Germany (BAST). This describes autonomous driving as the movement of vehicles, mobile robots, and driverless transport systems that behave largely autonomously. In this context, the vehicle does not make the driver autonomous, but it drives itself (Gasser et al., 2012,).

Trust and Adoption of AI Technologies

The most important prerequisite for the development of new technologies - as is the case with autonomous driving at levels 4 and 5 - is the use of the technologies by consumers. If a product is not adopted by the customers, the development and production are not meaningful.

One model that describes the reason people accept and use technology is Davis' (1985) Technology Acceptance Model (TAM for short). The TAM assumes that two factors - perceived usefulness and perceived ease of use - determine the intention to use technology. Since the introduction of this model, numerous empirical studies have shown that TAM is a robust model of technology acceptance behavior in a variety of information systems (including Davis et al., 1989; Davis & Venkatesh, 1996). Moreover, several studies apply TAM to explain the acceptance and use of driver assistance systems (including Choi & Ji, 2015; Ghazizadeh et al., 2012).

As an extension of the TAM, Ghazizadeh et al. (2012) argue for the consideration of the trust factor to explain the individual acceptance of driver assistance systems. Accordingly, trust in relationships between humans and automation is as important as between humans and humans (Sheridan, 1975). Consequently, trust is an important determinant of acceptance of automation under the intention to use it.

Research has shown that interpersonal trust depends on several factors. Among them are perceived competence, benevolence (or malevolence), intelligibility, and immediacy. Immediacy describes the extent to which the trust giver can quickly assert control or influence when something goes wrong (Bradshaw et al., 2004). Each of these factors or dimensions may be more or less important in a particular situation. Research has also shown that these factors are relevant to trust in automation (Hoffman et al., 2013). However, trust in automation also includes other factors that relate specifically to the limitations and weaknesses of the technology. These factors include reliability, validity, usefulness, and robustness (Hoffman et al., 2013).

Shin (2020) based on the TAM model and the research of Ghazizadeh et al. (2012), among others, established a model to describe the adoption of AI by humans, as shown below in Figure 2.

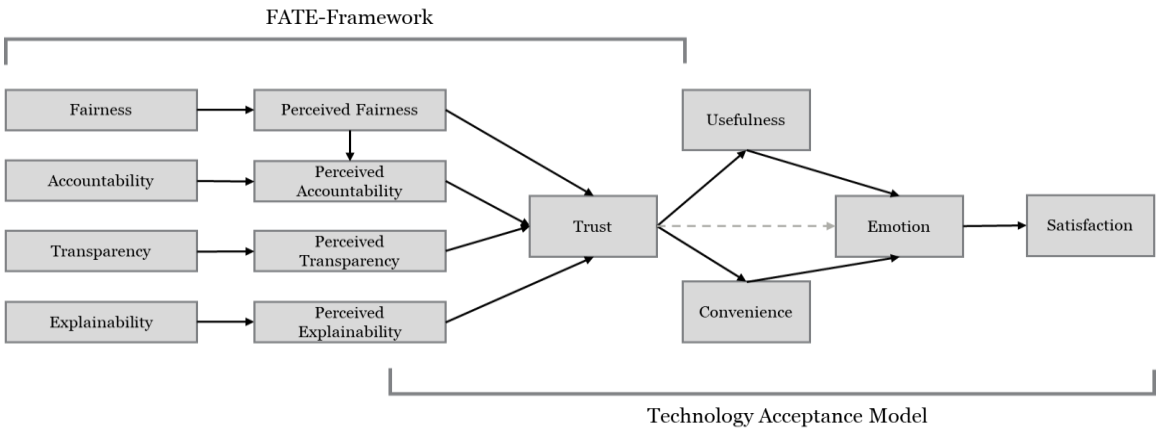

Figure 2. Artificial Intelligence Adoption Model, Adapted from Shin (2020)

The model shows how trust affects the two pillars of the TAM model and how this contributes to the adoption of AI. In particular, the four AI characteristics of Fairness, Accountability, Transparency, and Explainability are new to Shin's model. The user's perception of these four characteristics is seen by Shin (2020) as the cornerstone of whether an AI is trusted or not. The following is a simplified representation of Shin's (2020) model.

FATE Framework

The four properties of Fairness, Accountability, Transparency, and Explainability of Shin's (2020) adoption model are often referred to in an acronym as FATE. FATE is a framework that has been developed in the course of research in the field of "trust in AI" and is applied in particular to the development of algorithms (e.g., Lepri et al., 2017). In recent research, such as that of Shin (2020), the four characteristics are examined in terms of what influence they have on people's trust towards an AI. In the following, the individual characteristics of the FATE framework are examined in more detail.

Fairness

Although Fairness is touted in the context of increasing AI, there is no universally accepted definition of algorithmic fairness (Shin & Park, 2019). Lepri et al. (2017) attempt to define Fairness as the absence of discrimination or bias in decision-making. A distinction is made in research between Group-Level Fairness and Individual Fairness. Group-level fairness is defined as the fair treatment of groups of people with the same attributes (e.g., race, gender, disability) (Feuerriegel et al., 2020). Individual-level fairness is based on the notion that similar individuals are treated in a similar manner (Dwork et al., 2012).

Fundamentally, a critical role is what fairness-related harms the AI's decision has on humans. Contrary to common perception, the reality of running AI is that AI does not always work fairly. Without a careful approach to Fairness in every process of an AI design, a system can produce discriminatory outcomes for certain groups or individuals (Shin & Park, 2019). Given the many complex sources of unfairness, it is often not possible to fully "mitigate" a system or guarantee Fairness (Dignum, 2021). The goal then is to mitigate fairness-related harms as much as possible and avoid creating further unfair biases and their consequences (Diakopoulos, 2016).

Accountability

The second property can be translated literally and substantively from English as accountability. The perceived accountability of algorithms is the expectation that a user's beliefs, actions, or feelings about algorithms must be justified (Aggarwal & Mazumdar, 2008). Accountability in algorithms and their application starts with the designers and developers of the system that relies on them (Diakopoulos, 2016).

Consequently, unintended consequences are a critical condition against algorithm accountability. The result is an accountability gap that goes unaddressed and uncommunicated, exposing the organization to unexpected risks for which it may later be held responsible. How to collect citizen data transparently, what to collect fairly, and who is responsible for data management are also important considerations. Ultimately, these issues depend on how citizens perceive FATE, because the more they believe their data will be handled in a fair, transparent, and accountable manner, the more they will allow companies to collect their data. Thus, societal attitudes towards algorithms are shaped by public discourse. However, public understanding is limited by a technical barrier (Shin & Park, 2019).

Transparency

The third property of FATE analysis is Transparency. Translated in terms of content, this refers to the comprehensibility of an algorithm on the part of the user. Often, further decisions are made based on a single output, which is why the traceability of algorithms, and their result generation are becoming increasingly important. In intelligent automation, traceability is particularly important when it comes to transferring decision-making power from the user to the autonomous system. The user develops an understanding of the algorithm's mode of operation using the comprehensibility of the result and is thus already more likely to leave the decision-making power to the system. The goal is to make the emergence of recommendations in algorithmic processes obvious. Once the user understands the emergence of the output,

he can decide for himself how certain he considers the result to be. Moreover, it is then easier to interpret the result and possibly make further deductions (Shin 2020).

Miller (2018) divides the transparency of autonomous systems into two complementary approaches. First, transparency is enhanced by considering how well humans can understand the output when developing decision-generating systems. Second, transparency is ensured by "explicitly explaining decisions to humans" (Miller, 2018).

Explainability

The fourth property can be translated literally and in terms of content from English as Explainability. Above all, the mode of operation, the effect, and the degree of control of the algorithm can be better understood through this. However, there is often a trade-off here between organizational readiness and the Explainability of the algorithm. The main point to note here is that "Explainability" can provide business benefits. These are the eight benefits: Model Performance, Decision Process, Control, Security, Trust, Ethics, Accountability, and Regulatory (Oxborough et al., 2018).

According to Miller (2018), Transparency and Explainability are very closely related. Thus, the said second approach of Transparency is more a property of Explainability. Moreover, it equates to the concepts of interpretability and explanation. Here, attention must also be paid to the difference between an explanation and a justification. While a justification "argues" in terms of the decision or the generated output of the system, an explanation is rather there to explain the actual decision-making process.

As Figure 3 shows, Explainability here is a human-agent interaction problem, which in turn is an interface between artificial intelligence, social science, and human-computer interaction (Miller, 2018).

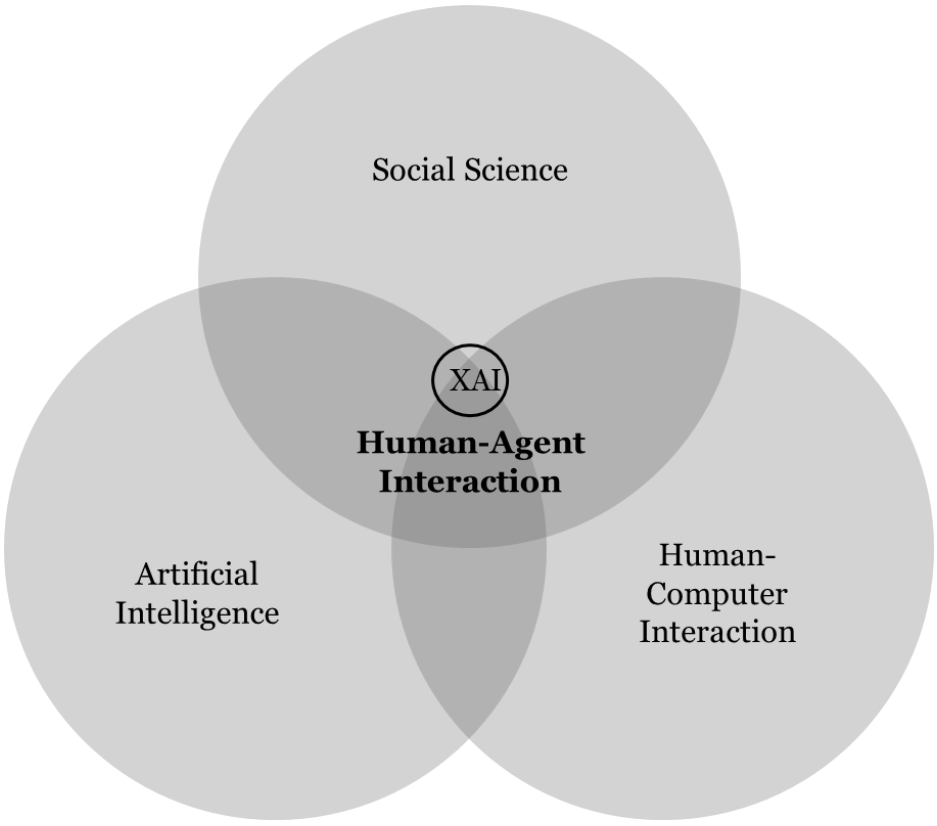

Figure 3: Areas of Explainable Artificial Intelligence, Adapted from Miller (2015)

Methodology for Addressing the Research Question

This chapter details the methodology used to address the research question.

Expert Interviews

In the context of this work, information is generated through expert interviews. These were conducted with experts from industry and research. In this subchapter, the selection and description of the experts, as well as the structural interview design, are explained.

Selection and Description of the Interview Partners

To generate a broad and deep knowledge, six experts from industry and research have been interviewed in the context of this work. The experts are briefly described below.

Experts from the industry. Two different professional fields were deliberately selected as industry experts: Software Developers and CTOs. This was done to take in different points of view and gain a broader understanding of the topic. The two software developers work in departments that specialize in the field of autonomous driving. The focus of their activities is on "Robotics & AI" and "Perception and Functional Safety for Urban Automated Driving". The CTOs work in a leading global IT services company. One of them leads global Autonomous Driving programs, the other the Autonomous Driving and AI division.

Research experts. Shedding light on other perspectives, two experts from the research community were also interviewed. Both experts work in the mobility and innovation systems research group of a large German research institute. One of the two experts' main areas of research is the exchange of information between the system and humans and its influence on trust in the use of autonomous vehicles.

The following table 1 provides an overview of the demographic data of the interviewees.

Characteristics	**Interview partners (n=6)**
Gender	67% male, 33% female
Geographical location	100% Germany
Education	67% Master's degree, 33% Doctorate
Work experience	0%<1 year; 33% 1-5 years; 17% 5-10 years; 50%>10 years
Organization size (#employees)	0%<10; 0% 10-500; 33% 501-5000; 0% 5001-100000; 67%>100000
Branch of the organization	17% OEM; 17% Tier-1; 33% IT services; 33% Research

Table 1. Overview of the Demographics of the Interview Partners

Conception of the Interview Guide

A semi-structured interview method was used to conduct the interviews. In a semi-structured interview, the questions are formulated in advance and the course of the interview is clearly defined. This makes the results more comparable. However, there are no predetermined answers to choose from. This also allows the experts to talk about things that were not considered in the preparation of the interviews (Renner & Jacob, 2020).

The interview guide used is divided into four main categories. In the first section (*topic block 1),* general questions about the topic area of "autonomous driving" are addressed. The aim is to develop a common understanding of the term "autonomous driving". Building on this common basis, the topic area "Trust in Autonomous Driving" is queried in *topic block 2*. Open-ended questions were deliberately chosen here to obtain unbiased information from the experts. The subsequent third section (*topic block 3*) forms the core

of the expert interview. Here, the FATE framework is first discussed in general terms, before the individual FATE properties are then examined in more detail. The last section contains concluding questions. Here, the interview partner is once again deliberately allowed to name further aspects. This section aims to obtain information that has not yet been mentioned.

Conducting the Expert Interviews

The communication platform "Zoom" was used to conduct the expert interviews. The appointment invitations were sent to the respective interview partners via e-mail or LinkedIn. The interviews took place on three different days in blocks of one hour each. On the first interview day (30.05.2021) the interviews were conducted in English, all other interviews in German. These took place on 11.06.2021 and 25.06.2021. A detailed list of the expert interviews conducted is given in Table 2.

Company	Occupation	Date	Time
OEM	Software Developer in the field of Autonomous Driving - Robotics & AI	30.05.2021	10.00-11.00
Tier-1	Software developer in the field of perception and functional safety for urban automated driving	30.05.2021	15.15-16.00
IT service	CTO in the field of Autonomous Driving and AI	11.06.2021	09.00-10.00
Research	Vehicle interior design in the field of Autonomous Driving	11.06.2021	11.00-12.00
IT service	CTO in the area of Global Autonomous Driving Programs	11.06.2021	13.00-14.00
Research	Industrial and Strategic Design in the field of Autonomous Driving	25.06.2021	10.00-10.45

Table 2. List of Interview Partners

Evaluation Method

After conceptualizing and conducting the interviews in the previous paragraphs, this chapter will focus on the qualitative content analysis of the interviews.

Qualitative Content Analysis According to Mayring

Qualitative content analysis is an approach to the empirical, methodologically controlled evaluation of large texts that arise during the data collection in social science research projects. Mayring has decisively shaped this method by further developing and improving the techniques previously used. Today, qualitative content analysis is one of the most widely used methods for the systematic analysis of texts (Mayring & Fenzel, 2019).

The text material embedded in its communication context is evaluated according to content-analytical rules without rashly quantifying (Mayring, 1994). Rather, the text material is to be analyzed systematically and step by step (Ramsenthaler, 2013). It is this systematic approach that primarily distinguishes this method from others. The basic process of qualitative content analysis consists of the rule-governed assignment of textual content to categories. The categories referred to in this context represent aspects of analysis and are based closely on the source material (Mayring & Fenzel, 2019). A repeated revision and adaptation of the created categories is recommended to ensure an adequate representation of the textual content. The categories can either be developed inductively from the material or determined deductively in advance. Whether an inductive or deductive approach is chosen depends, among other things, on the quantity and quality of the material. It is important to know whether the topic area has already been sufficiently investigated or not. Only then can categories be formed in advance (Mayring, 1994).

In the context of this work, some categories were formed in advance for topic blocks 1 and 2 based on the underlying literature and research. By evaluating the text material, it will be tested which of these categories apply and how often they are advertised. Nevertheless, the categories will be dynamically expanded and supplemented based on the existing interview material. If necessary, new categories will also be created during the analysis and the category system will be extended. This procedure leads to a deductive-inductive hybrid approach.

In contrast, an inductive analysis approach is used in topic block 3. Since the topic of FATE in autonomous driving is a research gap and there is no existing literature, the categories are formed only during the analysis. The exact procedure of the inductive Mayring analysis is available by the authors upon request.

Conducting the Content Analysis

The content analysis is carried out separately for the first three topic blocks of the interview guide. The online version of the QCAmap program was used to conduct the content analysis.

Analysis Topic Block 1. In the first thematic block, an inductive-deductive Mayring analysis is used to analyze the advantages and problems of autonomous driving. Of particular interest is whether the issue of trust is seen as a fundamental problem by the experts. The analysis is based on certain categories that have been formed in advance and with the addition of suitable literature. These are listed in the following table 3 for the respective analyses.

Advantages of autonomous driving	**Problems in the field of autonomous driving**
Security	Legal issues
Time	Ethical issues
Mobility	Security
Efficiency	Trust
Environmental protection	
Ride comfort	

Table 3. Benefits and Problems of Autonomous Driving, Based on Kallmeyer (2019)

After the first coding guideline for all categories is developed, the text passages are processed step by step and assigned to the corresponding categories. If it was not possible to assign all text passages to the existing categories, new categories were formed by inductive category formation. The final coding guidelines are available by the authors upon request.

Analysis Topic Block 2. In the second part, also using an inductive-deductive approach, the effects of a lack of trust on the user side are examined and shown in Table 4.

Impact of the "trust" issue	**Qualities to build trust**
Technology Acceptance (Hegner et al., 2019)	Human-human relationships (Hoffman et al., 2013)
Perceived risk (Topolšek et al., 2020)	Human-AI relationships (Hoffman et al., 2013)
Profitability (Richter & Hess, 2021)	FATE (Shin, 2020)
Purchase intention (Panagiotopoulos & Dimitrakopoulos, 2018)	

Table 4. Effects of the Trust Problem and Characteristics to Build Trust

Likewise, various characteristics of autonomous vehicles and the AI behind them that build trust are examined. On the part of the experts, the aim is to find out whether the FATE features are seen as an important tool for building trust. This makes it possible to find out, without having previously discussed FATE, what role FATE or individual FATE characteristics play in the development of autonomous vehicles. As in the analysis of the first thematic block, the text passages are also assigned to the categories and if this is not possible, new categories are formed. The final coding guidelines are available by the authors upon request.

Analysis Topic Block 3. The third thematic block of the interview guide, as already mentioned in chapter 3.1.2, deals with the FATE framework and its characteristics. However, as there is hardly any literature on this, the inductive approach of the Mayring analysis is used in this chapter. This differs from the Deductive-Inductive approach in that no categories are defined at the beginning, but only during the text analysis. This analysis aims to find out whether the experts are aware of the FATE framework, what added value such a framework offers, how the experts define the individual characteristics, whether and why they are important, and what measures can be taken to promote them.

Evaluation of the Results

In this chapter, a summary of the results of the Mayring analysis is given, starting with the results of the analysis of the first thematic block. The final coding guidelines are available by the authors upon request.

Results of Topic Block 1. Within the analysis of the advantages of autonomous vehicles, a large part of the categories formed based on the literature could be confirmed. In particular, the increase in the safety of the user and his environment, the free availability of the passenger's travel time, and the improvement of traffic flow were highlighted in the interviews. Only advantages regarding environmental protection were not mentioned by the experts.

All the problems listed in Table 2 were mentioned by the experts during the interviews and were thus confirmed. Several experts cited the lack of user confidence in autonomous driving vehicles as a problem. This suggests that the industrial companies and research institutions understand a relationship of trust needs to be established to generate social acceptance of the technology. However, according to their statements, the companies and institutes often still lack the empirical values of the users to build this up in a targeted manner. For this reason, the category of *empirical values* was additionally developed inductively within the framework of the Mayring analysis.

All final categories used to analyze Topic Block 1 are available by the authors upon request.

Results of Topic Block 2. In the second block of topics, the effects of the lack of trust on the user side were examined. In response, most experts indicated a lower acceptance of autonomous vehicles. In this context, one expert cited a lower purchase intention on the part of consumers. The lack of trust and the resulting lack of technology acceptance also means that OEMs are comparatively hesitant to push into the development of high levels of autonomy. Entering the market too quickly would lead to a loss of *short-term profits* due to a lack of market penetration.

The question of which AI properties help to build trust with the user was hardly addressed. Only transparent communication was mentioned by two experts. Rather, the experts see the actual use and the build-up of experience during the use of autonomous vehicles as a decisive point for creating trust in the technology. In this respect, AI should be designed in the early development stage in such a way that its use and the resulting experiences are positively shaped. This point of consideration was summarized in the inductively formed category "Other".

Results of Topic Block 3. An unsystematic analysis of the third topic block shows that none of the experts is familiar with the FATE framework, but that the individual characteristics are certainly used. The experts' basic understanding of the individual FATE characteristics is largely consistent with the definitions in the literature (see Chapter 4.2).

Through the inductive Mayring analysis, some points could be identified that show an added value of the FATE framework. On the one hand, the FATE framework increases the consistency of system development. By considering societal aspects and adding experts from a wide range of ethical topics, the system development achieves a higher consistency. In addition, the FATE framework can help to ensure that the

end-user and his or her needs are considered and integrated early in the development process. As a result, products become more customer-centric and adoption probability is higher. Some experts also believe that the FATE framework can contribute to better communication between AI and humans. The exact categories are available by the authors upon request.

In addition to the added value offered by the FATE framework, the further inductive analysis identified several points that underpin why the individual FATE characteristics are important. Consideration of fairness is particularly important in the development of AI systems to ensure adherence to ethical principles and can help to make data selection processes more balanced. For two experts, Accountability is perceived as important to define clear liability boundaries. Transparency is important for more than half of the experts to give users a basic understanding of the functionality and operation of autonomous vehicles. This includes, among other things, information about which models and algorithms run in the background and with which data the system is ultimately enriched. Explainability is seen by most experts as necessary to explain the information understandably disclosed in Transparency. It is important to explain to the user in an intuitive way how the autonomous vehicle works, why it makes which decisions, and what information it processes.

Most of the findings could be drawn from the third inductive Mayring analysis in topic block 3. Here, it was systematically examined which measures the experts recommend promoting the four FATE characteristics. In total, six measures for promoting *Fairness* could be identified. These relate, in addition to a clear definition of the term Fairness, both to fairness in ethical aspects (e.g., fair data selection) and fairness concerning the use of the autonomous vehicle (e.g. disability-friendly design of the car). Measures that can promote *Accountability* according to the experts were named four times. Here, the experts refer primarily to clear communication of responsibility. Be it from the outset on the part of the carmaker or during the use of the autonomous vehicle, for example, to inform the occupant as to when he or she is responsible for the vehicle. In the area of *Transparency,* too, communication is the main key for the experts to be able to build trust between the autonomous vehicle and its user. Among other things, transparent vehicle interiors, clear and transparent communication of the vehicle's activities and, the implementation of verification tests along with communication of the results were mentioned. The measures for generating greater *Explainability* overlap strongly with those of Transparency. This also reflects the experts' opinion that Transparency and Explainability are related concepts and cover similar aspects. However, one example that was explicitly mentioned in the context of Explainability is the analysis of users' prior knowledge. The information about user knowledge should help to better adapt the AI and the communication of the vehicle's activities to the occupants. This in turn should lead to a better understanding of the users regarding the vehicle.

A detailed overview of all identified measures is available by the authors upon request. In the following chapter, the individual measures to promote the FATE characteristics are also discussed in more detail.

Critical Analysis of the Expert Interviews

In this chapter, the findings of the interviews are once again consolidated and critically examined. The results are first compared and then discussed based on the existing literature.

Comparison of Industry and Research

The introductory questions deal with the industry as well as the respective role in the expert's company to clarify how the term "autonomous driving" is used, its advantages, and potential areas of application.

Topic Block 1: Explanation of terms, advantages, and problems of autonomous driving. In describing autonomous driving, the experts refer to the SAE levels, which have already been explained in the second chapter of the theoretical foundations. There is a development perspective as well as a social and economic perspective. A distinction was made between AD (Autonomous Driving) and ADAS (Advanced Driver Assistance Systems). The terms refer to automated driving and the associated driver assistance systems.

The experts emphasized that autonomous driving leads to more safety and efficiency in road traffic and that this is due to the decreasing number of fatalities. During the congestion problem, more efficient road traffic and traffic routes are seen due to intelligent rerouting of individual routes.

A general problem is due to tech companies, such as Tesla, which create a distorted expectation on the part of the user due to the inappropriate naming of their driving systems. Tesla uses the name "Self-Driver", which does not represent the actual level of knowledge of the driving system. In this way, people mistakenly rely on a system that does not do justice to its functional extent and triggers a negative basic attitude in the driver. The verification and validation of driver assistance systems must be thoroughly checked before they are ready for the market.

Topic Block 2: Trust. Trust is correlated with the comfort level. In other words, the more comfort is created in the vehicle, the more willing the user is to relinquish control to the system. However, the problem is based on a simple mechanism. The mechanism of verification. Camera systems that track and observe the user inside the vehicle make it clear that the system would rely on interaction with the user. This would reduce the user's trust because he would realize that the system would not function in its entirety without him, and he would retain an essential function while driving.

The origin of the mistrust towards AI or automated driving can be traced back to the lack of knowledge on the part of the user. The empirical values and the expectations of people have been enriched with positive and negative impressions for decades. Due to the radical change in road traffic, the empirical values of people towards the system are still insufficient to think through the consequential effects of the AI's scope of action, he said. Unfortunately, AI cannot yet build on the experience of the users, which means that intuitive development is not possible. To generate a positive attitude on the part of the users, the AI would have to build on these empirical values. After all, users would attach themselves to what they know. Therefore, it makes sense to evoke a positive experience on the part of the user and to build up the existing knowledge with new knowledge, which in turn can be linked to a positive experience.

On the other hand, people are very adaptive when it comes to using available technologies. From a market theory point of view, the development of driver assistance systems would not continue if there was no demand. Consequently, people are already ready for radical innovation.

The trust model would need to promote characteristics such as Emotional Bonding to facilitate communication between the vehicle and its occupants. The emotional bond could be seen as a digital companion. If the AI becomes a personalized system, i.e., has its name and voice, the feeling of security is reinforced. Communication could be conducive to trust with appropriate colors, round shapes, appropriate lighting mood. In addition, the end user's willingness to learn is an important characteristic for strengthening trust in the AI. The end user's general willingness to learn is necessary to create this purely positive driving experience.

Topic Block 3: Fairness. Fairness is understood as a conservative system that leads to more safety while adhering to the traffic rules. If the system is based on these rules, it is fair. One example is driving across national borders, where different traffic rules apply. In addition, the experts agree that equal gender treatment of passengers is an essential point. This means that systems do not change their behavior towards the occupant and act in a gender-neutral way. In addition, there must be a balanced enrichment of personal data. About the individual and their characteristics, such as cultural backgrounds.

When it comes to data selection, care should be taken to ensure that neural networks only work with data that are relevant for the evaluation.

Finally, the term "fairness" should be seen in the context of the morality machine. That is, how the AI should decide in extreme situations.

Topic Block 3: Accountability. In the event of accidents, it is important to have measures in place to ensure that the parties involved in the accident are held adequately accountable. In the event of damage, it should accordingly be checked whether a system failure or an anthropogenic cause of the accident was the underlying cause. In such exceptional cases, one expert emphasized that there is a worldwide standard, the so-called "Automotive Safety Integrity Levels" (ASIL). ASIL is part of the ISO 26262 series of standards, which include the specific requirements in electrical and/or electronic (E/E) systems of road vehicles. In the context of functional safety, the increasing technological complexity must be considered. Software and hardware contents are important components in the implementation of necessary mechatronics. At the same time, the risks of possible systematic failures and random hardware failures are increasing. To address these risks, the ISO 26262 series of standards formulate guidance that includes appropriate requirements and processes (ISO, 2018).

Two terms were also mentioned during accountability: "reliability" and "liability". In his opinion, the limits of liability are opaque and uncertain. One expert also mentioned that when a human is in an autonomous vehicle, the responsibility tends to lie with the car manufacturer or the person who programmed the AI. Quasi where the technology fails. However, he did not rule out the possibility of an approach in which liability could be adequately shared. Another expert believes that at Level 3, liability still lies with the driver, as accompanying driver assistance systems are used. During legislation, a sensible solution would still have to be found in each fully autonomous vehicle.

There is also an opinion that the AI decides accurately using decision trees in each scenario and that this decision is traceable.

One expert stressed that the actions, feelings, or beliefs that a user develops towards the system must be justified. Based on this definition, the speech function is equipped with a factual and emotional part, he said. The AI could misinterpret the user's speech even though the emotional part of the statement was made clear. The expert described this emotion recognition as a measure to create accountability.

Topic Block 3: Transparency. Transparency had already taken place decades ago through the acoustic imitation of a turn signal. Thus, the real sound of the turn signal was imitated during operation to provide feedback to the driver. This feedback issue had also been observed in the context of electric vehicles. Accordingly, the driver needed acoustic feedback when the engine was on, as expected from an internal combustion engine. Transparent communication would fail in many points.

The experts agree that the important processes that take place in the background of an AI should be described more transparently to the user. An adequate level of detail, as well as the right type of communication, must be considered. Terms of use must be brought closer to the driver in an understandable way. The focus of the car companies is not to show how mathematical equations arrive at a certain result, but rather to show high-level information.

Google is working with model scorecards that can be used more widely in autonomous vehicles in the future. In parallel, an evaluating authority could contribute to increasing transparency under the obligation of secrecy.

Further measures would be smart materials that convey a message to the user in the form of a user interface and with the aid of touch screens or LEDs.

Topic Block 3: Explainability. Models can be trained on certain data to make a prediction. The model should explain why it made the prediction and with what accuracy. Other opinions emphasize that AI does not have to explain all background processes. Explainability is finally seen as a form of error analysis. In the USA, for example, it would have to be published regularly if there had been disengagements with autonomous test vehicles. In addition, the interaction between humans and machines would have to be an animated and collaborative model, which was known from other industries. Thus, augmented reality should become a future standard in the context of Explainability.

Another interviewee stressed how important it is to make the limits and the scope of action of AI clear to the user. In the context of social robotics, the robot "Pepper" is a good example of this. This robot has a very humanoid appearance and creates a certain expectation on the part of the customer that it has certain functions at its disposal. However, this is where the problem lies. Because Pepper has hands, the user would think that he could handle any object, but he could not. Thus, the physical design of the robot or the AI plays an essential role in communication.

In addition, the focus is more on an intuitive product environment and less on explaining technical user manuals. It is more about what the system communicates to the user during use and what information the customer ultimately wants. In the case of new types of flight taxis, safety can only be guaranteed by constant feedback from the system. For example, in the event of severe turbulence and other severe weather scenarios. In any case, it should be regulated and explained what is happening and how the system should act to ensure user confidence.

Discussion of the Results Based on the Literature

This chapter examines the FATE characteristics in terms of their definitions. The aim is to find out whether the experts' understanding of the individual characteristics is consistent with the definitions from the literature.

Fairness

The interviews shed light on how the observance of traffic rules domestically and across national borders, as well as the equal treatment of occupants in terms of gender, must be formulated as a requirement for the AI. In extreme situations, the system should be seen as a "morality machine" that always acts fairly for each scenario. Also, in the context of data enrichment and data selection, the data should be weighed about humans and their characteristics. Thus, expert testimony is consistent with the definition of fairness as the absence of discrimination and bias in decision-making (Lepri et al., 2017).

Accountability

The results make it clear that there is still disagreement among the experts as to who is liable in which scenario. Measures that are already being implemented to realize accountability are based on ASIL. Based on this industry standard, it is accordingly ensured how accountability is described to pronounce liability according to the risk situation. According to the definition, perceived accountability of algorithms is the expectation that a user's beliefs, actions, or feelings about algorithms must be justified (Aggarwal & Mazumdar, 2008). This has been taken up several times by the experts.

Transparency

According to the definition of Transparency, it is particularly important to the experts that AI should be a transparent and comprehensible system (Shin, 2020). The intuitive product environment shows only the high-level information that is necessary for building trust. Furthermore, smart materials are used to create a transparent user interface.

Explainability

In chapter two, Explainability is defined as the user's understanding of the algorithm's operation, effect, and degree of control (Oxborough et al., 2018). Failure analysis that measures the degree and effect of algorithms is standard practice for Autonomous Test Vehicles. An intuitive product environment could be realized through augmented reality and better represent the degree or scope of action of the system. In the context of social robotics, understanding how robots work is important to assess as a user what functions the system can perform. The way systems work is characterized by giving the user feedback about the current situation.

Explainability is a very large topic area and counts as the most important term for most of the interview partners. Explainability as well as Transparency are said to be closely correlated and can explain the system at hand well. Thus, they confirm the findings of Miller (2018) that Transparency and Explainability are related terms. Transparency would also rank high among the experts, reinforcing the building of trust with the inmate. After that, Accountability would be classified. Reliability and liability are mentioned in connection with accountability and should be placed much more centrally. Reliability must grow through the entire system. Liability alone would slow down technological progress.

Since AI represents a conscious system, Fairness is less relevant and predefined, to begin with. The traffic rules for humans were the same for everyone anyway. On the other hand, the term fairness was also mentioned as a basic requirement.

Conclusion

This chapter first summarizes the main findings of the thesis concerning the research question. This is done by explaining the key factors and findings of the interviews about the FATE framework. Potential future research to further investigate the research question is then highlighted.

Central Findings on the Research Question

Considering the interaction between humans and automobiles as a socio-technical system, FATE can be seen as a framework that forms an orientation during the development of the system. However, these guidelines should not only be used for the development of the software but also the entire vehicle development, so that an optimal acceptance of the autonomous system can be created. In the area of human-agent interaction, Explainability is again in "Explainable AI" (see Figure 3).

The concrete concept of FATE is not used in practice and is still foreign to the experts from the industry. During the respective interviews, however, it became apparent that the substance of the concept is an integral part of the development process of autonomous systems. The meaningfulness of the four characteristics is always followed when building trust in algorithms, while no differentiation is made between the terms, and many different personal definitions are understood under them. The individual FATE characteristics are also always attributed to having a trust-promoting effect on the human-AI relationship. It was also mentioned that the experts see the actual use and the building of experience during the use of Autonomous Vehicles as a crucial point for trust in the technology to be established. In this regard, AI should be designed in the early development stage in such a way that its use and the resulting experiences are positively shaped. The early integration of the FATE framework can also provide a solution for this by bringing user orientation into the development process.

To increase the application of the FATE concept, some aspects should be considered. A weighting and clear division of the characteristics would ensure a uniform understanding. Furthermore, the interview results show different assessments regarding which characteristic should be the focus of the framework. Also, the connection between the respective characteristics is addressed in some research papers. This is mostly not done in the context of the FATE concept, but when the individual properties are addressed.

Based on the findings from the interviews and Miller's approach (Miller, 2018) of using the terms Transparency and Explainability synonymously, it makes sense to include the property "Ethics" in this framework instead of Explainability. This has also been done in some research (Kasinidou et al., 2021; Kleanthous et al., 2021; Microsoft, 2021). Ultimately, however, a clearly defined approach to the application of the framework would have a positive contribution to its establishment.

During the seminar work, various measures were identified that make a positive contribution to the realization and better implementation of the FATE properties and thus have a positive influence on users' trust. These identified measures should be considered when applying the framework. One main measure, which was frequently mentioned in different contexts, is communication. Clear communication tailored to the user promotes Transparency and explicability. In addition, simple and clear communication about which actions are the responsibility of the user, which are the responsibility of the system, and at which point the transfer of responsibility takes place is important.

Conclusion for Further Research

The seminar paper specifically dealt with the interview analysis of selected experts from industry and research. To tie in with the present work, a literature review of technology papers that use FATE or other frameworks that raise technology acceptance on the part of users is recommended.

A possible further approach would be to conduct expert interviews again, asking to what extent the FATE framework can help to prevent the problems analyzed in topic block 1. In this way, some of these problems could be tackled in a more targeted manner and, at best, solved more quickly and effectively.

In addition, the measures for promoting the FATE characteristics identified in the coding guidelines C.3.3 can be analyzed in a further step based on existing literature or a further expert survey (similar to the Delphi method). This would bring consistency to the findings of this study and provide a basis for targeted recommendations for action. In addition, this could result in the possibility of designing the FATE framework in a more targeted manner.

The question of how the FATE framework can be integrated into the development of systems or even in existing systems also opens a broad field of research.

Ultimately, all these conclusions and research tasks lead to a common point. It is necessary to find out how the FATE framework must be designed so that it can be easily, unambiguously, and effectively applied within the development process of autonomous systems.

References

Aggarwal, P., & Mazumdar, T. (2008). Decision delegation: A conceptualization and empirical investigation. *Psychology & Marketing*, 25(1), 71-93.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J., & Rahwan, I. (2018). Blaming humans in autonomous vehicle accidents: shared responsibility across levels of automation. *Nature Human Behaviour*, 4(2), 1-40.

Beiker, S. A. (2012). Legal aspects of autonomous driving. *Santa Clara L. Rev.*, 52(4), 1145-1156.

BGBl (2010). Ordinance on the admission of persons to road traffic (Fahrerlaubnis-Verordnung - FeV), Appendix 4 - Suitability and conditional suitability for driving motor vehicles. https://www.gesetze-im-internet.de/fev_2010/anlage_4.html.

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

Bradshaw, J. M., Jung, H., Kulkarni, S., Johnson, M., Feltovich, P., Allen, J., Bunch, L., Chambers, N., Galescu, G. , Jeffers, R., Suri, N., Tayson, W., & Uszok, A. (2004). Toward trustworthy adjustable autonomy in KAoS. In *Trusting Agents for Trusting Electronic Societies*, (p. 18-42). Heidelberg: Springer.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, 31(10), 692-702.

Davis, F. D. (1985). A technology acceptance model for empirically testing new end-user information systems: Theory and results (Doctoral dissertation, Massachusetts Institute of Technology).

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management science*, 35(8), 982-1003.

Davis, F. D., & Venkatesh, V. (1996). A critical assessment of potential measurement biases in the technology acceptance model: three experiments. *International journal of human-computer studies*, 45(1), 19-45.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62.

Dignum, V. (2021). The Myth of Complete AI-Fairness. In *International Conference on Artificial Intelligence in Medicine*, (p. 18-42). Cham: Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214-226.

Fehrenbacher, K. (2015, 16. Oktober). How Tesla is ushering in the age of the learning car. Fortune. https://fortune.com/2015/10/16/how-tesla-autopilot-learns/

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: challenges and opportunities. *Business & information systems engineering*, 62(4), 379-384.

Gasser, T. M., Arzt, C., Ayoubi, M., Bartels, A., Bürkle, L., Eier, J., Flemisch, F., Haecker, D., Hesse, T., Huber, W., Lotz, C., Maurer, M., Ruth-Schuhmacher, S., Schwarz, J., & Vogt, W. (2012). Legal consequences of increasing vehicle automation. Reports of the Federal Highway Research Institute. *Fahrzeugtechnik subseries*, Bremerhaven: Wirtschaftsverlag NW, 83, pp. 1-124.

Ghazizadeh, M., Peng, Y., Lee, J. D., & Boyle, L. N. (2012). Augmenting the technology acceptance model with trust: commercial drivers' attitudes toward monitoring and feedback. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Los Angeles: Sage Publications, 56(1), 2286-2290.

Hegner, S. M., Beldad, A. D., & Brunswick, G. J. (2019). In automatic we trust: Investigating the impact of trust, control, personality characteristics, and extrinsic and intrinsic motivations on the acceptance of autonomous vehicles. *International Journal of Human-Computer Interaction*, 35(19), 1769-1780.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, 28(1), 84-88.

Heitmann, X (2021, 21. April). Milestone in the automated public transport. ioki. https://ioki.com/en/milestone-in-the-automated-public-transport-passenger-operation-for-self-driving-shuttles-on-demand-started-in-karlsruhe/

ISO (2018). ISO 26262-9:2018(en), Road vehicles - Functional safety - Part 9: Automotive safety integrity level (ASIL)-oriented and safety-oriented analyses. ISO. https://www.iso.org/obp/ui/#iso:std:iso:26262:-9:ed-2:v1:en/

Kasinidou, M., Kleanthous, S., Orphanou, K., & Otterbacher, J. (2021). Educating Computer Science Students about Algorithmic Fairness, Accountability, Transparency and Ethics. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education*, 1, 484-490.

Kallmeyer, F. (2019, 11. December). Autonomous driving - opportunities and challenges. Zukunft Mobilität. https://www.zukunft-mobilitaet.net/170765/strassenverkehr/autonomes-fahren-chancen-und-herausforderungen-sae-level5/#fn-170765-68

Kleanthous, S., Otterbacher, J., Bates, J., Giunchiglia, F., Hopfgartner, F., Kuflik, T., Orphanou, K., Paramita, M. L. , Rovatsos, M., & Shulner-Tal, A. (2021). Report on the CyCAT winter school on fairness, accountability, transparency and ethics (FATE) in *AI*. *ACM SIGIR Forum*, 55(1), New York: ACM, 1-9.

Kraftfahrt-Bundesamt (2021). Vehicle stock overview on 01 January 2021. KBA. https://www.kba.de/DE/Statistik/Fahrzeuge/Bestand/Jahresbilanz/bestand_jahresbilanz_node.html

Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, 126-137.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4), 611-627.

Maurer, M. (2015). Autonomous driving: technical, legal and societal aspects. In Maurer, M. & Christian Gerdes, J. & Lenz, B. & Winner, H. (Pub.), (pp. 1-8). Heidelberg: Springer.

Mayring, P. (1994). Qualitative content analysis,. In Boehm, A. & Mengel, A. & Muhr, T (Pub.), Understanding texts, 14th ed. (pp. 159-175). Konstanz: Universitätsverlag Konstanz.

Mayring, P. (2015). Qualitative content analysis: foundations and techniques, 12th edition. Weinheim: Beltz.

Mayring, P., & Fenzl, T. (2019). Qualitative content analysis. In Baur, N. & Blasius, J. (Pub.), Handbuch Methoden der empirischen Sozialforschung, 2nd ed. (p. 633-648) Wiesbaden: Springer VS.

Microsoft (2021). FATE: Fairness, Accountability, Transparency, and Ethics in AI. Microsoft. https://www.microsoft.com/en-us/research/theme/fate/ [19.07.2021].

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artificial intelligence*, 267, 1-38.

Nastjuk, I., Herrenkind, B., Marrone, M., Brendel, A. B., & Kolbe, L. M. (2020). What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. *Technological Forecasting and Social Change* 161, 1-26.

Nobis, C. & Kuhnimhof, T. (2018). Mobility in Germany - Mid findings report. Bonn: *Institute for Applied Social Sciences*, 1-136.

Oxborough, C., Cameron, E., Rao, A., Birchall, A., Townsend, A., & Westermann, C. (2018). Explainable AI: Driving business value through greater understanding. https://www.pwc.co.uk/audit-assurance/assets/pdf/explainable-artificial-intelligence-xai.pdf

Panagiotopoulos, I., & Dimitrakopoulos, G. (2018). An empirical investigation on consumers' intentions towards autonomous driving. Transportation research part C: emerging technologies, 95, 773-784.

Ramsenthaler, C. (2013). What is "Qualitative Content Analysis?“. In Schnell, M. et al. (Pub.), The Patient At the End of Life, 1st edition (pp. 23-42). Wiesbaden: Springer VS.

Renner, K. H., & Jacob, N. C.. (2020). What is an interview?. In The interview (pp. 1-17). Heidelberg: Springer Vieweg.

Richter, M., & Hess, J. (2021). Profitability and competitiveness of a robotaxi fleet-A macroscopic traffic simulation in the city of Zurich. *Journal of Mobility and Transport*, 8, 2-13.

Schreurs, M. A., & Steuwer, S. D. (2015). Autonomous driving-political, legal, social, and sustainability dimensions. In *Autonomous driving* (pp. 151-173). Heidelberg: Springer Vieweg.

Sheridan, T. B. (1975). Considerations in modeling the human supervisory controller. International Federation of Automatic Control, Triennial World Congress, 6, 40.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, 1-25.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98, 277-284.

Tesla Inc. (2021). Autopilot. Tesla inc. https://www.tesla.com/de_DE/autopilot?redrect=no

Tian, Y., Pei, K., Jana, S., & Ray, B. (2018, May). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In Proceedings of the 40th international conference on software engineering, 303-314.

Topolšek, D., Babić, D., Babić, D., & Cvahte Ojsteršek, T. (2020). Factors influencing the purchase intention of autonomous cars. *Sustainability*, 12 (24), 10303.

Clock, A. (2016). Automated driving. Challenges for road safety. bfu - Beratungsstelle für Unfallverhütung, bfu-Grundlagen.
http://www.garagenvision2025.ch/download/bfu_grundlage_automatisiertes_fahren.pdf

German Association of the Automotive Industry (2015). Automation: From driver assistance systems to automated driving. VDA, Berlin.

Certifying Fairness of Artificial Intelligence

Emerging Trends in Internet Technologies, Summer Term 2021

Betül Özdemir
Master Student
Karlsruhe Institute of Technology
betuel.oezdemir@student.kit.edu

Daniel Andreas Kohl
Bachelor Student
Karlsruhe Institute of Technology
urroj@student.kit.edu

Sarah Meerkamp
Bachelor Student
Karlsruhe Institute of Technology
sarah.meerkamp@student.kit.edu

Luca Vetter
Bachelor Student
Karlsruhe Institute of Technology
luca.vetter1@web.de

Susanne Piekarek
Master Student
Karlsruhe Institute of Technology
susannepiekarek@web.de

Abstract

Background: *Artificial Intelligence (AI)-embedded systems offer many benefits, however, also come with fairness issues and risks of discrimination, such as racial bias. A promising approach for ensuring fairness are certifications as a means of (independently) assessing the fairness and non-discrimination of AI-embedded systems.*

Objective: *However, research and practice still struggle with designing and performing fairness certifications. In this study, we clarify how to certify fairness of AI-embedded systems.*

Methods: *We enhance the understanding by performing a literature review on fairness in AI and conducting expert interviews.*

Results: *Our study discusses three key structural building blocks for fairness certifications: fairness criteria to be assessed (certification content), potential issuers and auditors (certification sources), and auditing methods to evaluate fairness (certification process).*

Conclusion: *With this study, we provide a first conceptualization of important building blocks for a fairness certification to highlight its feasibility and usefulness in the domain of AI.*

Keywords: artificial intelligence, fairness, discrimination, certification, assurance seal

Introduction

Artificial intelligence (AI) is one of the most discussed technology trends in research and practice in recent years, referring to the ability of machines to perform cognitive functions associated with human minds (Rai et al., 2019, p. 1073). AI technology is not only used as an artificial human being (i.e., AI in robotics) or to

perform data analytics and inference on data, but is more and more converging into existing information systems (IS) to augment or automate functionalities referring to AI-embedded systems (Glikson & Woolley, 2020). For example, AI models are nowadays embedded in image analysis systems in radiology to provide a higher diagnostics speed (Miller & Brown, 2018), in chat systems to automate customer support (Adam et al., 2021), or autonomous vehicles (Hengstler et al., 2016; Renner, Lins, Söllner, et al., 2021).

Whereas AI-embedded systems may provide glaring opportunities (e.g., novel intelligent functionalities and process automation), using AI-embedded systems is also associated with risks, such as infringing individuals' privacy or the vast presence of racial bias (Floridi, 2019; Floridi et al., 2018). Indeed, two of the most discussed issues of AI-embedded systems are unfairness and discrimination. These systems make inferences by learning existing patterns from data and are prone to biases whereby individuals or whole groups are treated unequally (Feuerriegel et al., 2020). A large number of incidents with AI-embedded systems have shown that such systems may tend to discriminate against individuals and groups while learning from data that captures past injustices (Mehrabi et al., 2021) or lead to discriminatory results due to mishandling or misuse of the AI-embedded systems (Cremers, 2019), among others. For example, researchers used deep learning to identify skin cancer from photographs (Esteva et al., 2017). Yet, their algorithms have learned to make inferences from mostly white persons and thus are more likely to make errors when classifying photographs of people of color (Zou & Schiebinger, 2018).

Given such fairness risks and other trust-related issues with AI, users are still hesitant to adopt AI-embedded systems or may even be relatively hostile toward them (Glikson & Woolley, 2020; Hengstler et al., 2016; Renner, Lins, Söllner, et al., 2021; Renner et al., 2022). Whereas it is in users' best interest to use fair AI-embedded systems, users are mostly unable to determine by themselves whether an AI is fair and free of discrimination or not. Users commonly have no control over extant AI-embedded systems and their inner workings. Besides, users lack experience and profound knowledge of how an AI model's logic provides decisions and functionalities due to the system's novelty, complexity, opaque nature, and the non-determinism of AI behaviors, among others (Eiband et al., 2021; Glikson & Woolley, 2020). Not only users but also providers of AI-embedded systems, policymakers and researchers are looking for novel ways to ensure fairness of such systems to prevent bias and discrimination. At the same time, they also want to communicate the degree of fairness transparently to reduce user uncertainty and increase the adoption of AI-embedded systems.

A promising approach for ensuring fairness are IS certifications as a means of (independently) assessing the fairness and non-discrimination of AI-embedded systems (Cremers, 2019; Hengstler et al., 2016). In general, IS certifications are neutral third-party attestations of specific system characteristics, operations and management principles (Lins & Sunyaev, 2017; Löbbers & Benlian, 2019). Previous research has shown that issuing certifications in highly vulnerable contexts (e.g., using cloud services) can increase users' trust by reducing information asymmetry related to technologies' characteristics and providers' behaviors (Lansing et al., 2018; Löbbers et al., 2020). In the context of AI, researchers and practitioners have started to discuss how to certify AI-embedded systems (e.g., Matus & Veale, 2021; Morik et al., 2021), to derive related AI standards containing best practices (e.g., "AI Cloud Service Compliance Criteria Catalogue" (AIC4); Federal Office for Information Security, 2021), and to develop methods and metrics to assess AI's characteristics (e.g., IBM's Toolbox 360 Fairness comprising fairness metrics).

While these recent (research) endeavors provide valuable contributions, we still lack knowledge on designing and performing fairness certifications. Reasons for this knowledge gap are diverse: knowledge on fairness is scattered across disciplines (e.g., IS, computer science, and ethics); the community has so far not agreed on fairness standards and best practices (Cremers, 2019; Feuerriegel et al., 2020); assessing fairness is challenging because fairness requires a socio-technical perspective (Feuerriegel et al., 2020); and systems are threatened by biases that can emerge along the complete AI pipeline (i.e., bias regarding data, modelling and inadequate applications; Barocas & Selbst, 2016; Cremers, 2019) . Answering calls for fairness certifications to ensure non-discrimination of AI-embedded systems (e.g, Cremers, 2019; Independent High-Level Expert Group on Artificial Intelligence, 2019; Morik et al., 2021), we investigate the following research question (RQ):

RQ: How can fairness of AI-embedded systems be certified?

To answer our research question, we adopt a two-step approach by first reviewing the scattered literature on AI's fairness to identify certification criteria and auditing methods to assess the degree of fairness. We

then conduct semi-structured interviews with experts to deepen our knowledge on fairness certifications and extend our findings.

We structure our findings along the key structural building blocks of certifications proposed by Lansing et al. (2018). First, we turn to a certification's content and discuss certification criteria that can be assessed to evaluate the fairness of an AI-embedded system, including criteria related to the development team, training data, and selection of attributes. Second, we summarize first thoughts about the source of the certification, particularly who can be the issuer and auditor of a fairness certification and the example requirements for these bodies. Finally, we look at the certification process, defining how an auditor evaluates conformance of the AI-embedded system with the certification criteria and discussing the applicability of traditional auditing methods in the context of fairness certifications.

Our study contributes to research and practice by providing a starting point on how to signal the fairness of AI-embedded systems. We provide a first conceptualization of important building blocks for a fairness certification to highlight its feasibility and usefulness in the domain of AI. Besides, we show that a social-technical perspective is needed for the content of AI-related certifications, and fairness-specific criteria must be considered. Finally, we reveal similarities relating to the certification process and source of other existing certifications and the need for continuous auditing and monitoring due to the rapid change of training data and the AI models themselves.

Theoretical Background

Fairness and Discrimination of Artificial Intelligence

Various definitions have been put forward to formalize the concept of "fair AI" (Feuerriegel et al., 2020). This study aligns with a more general definition of fairness, referring to the inverse of discrimination, meaning that a fairness-aware AI model produces non-discriminatory predictions (Bantilan, 2017). Fairness is, therefore, the absence of any discrimination resulting from any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making (Mehrabi et al., 2021). Discrimination is commonly differentiated into direct and indirect (Mehrabi et al., 2021). On the one hand, direct discrimination occurs when protected attributes of individuals are used against their best interest; there are even some characteristics where discrimination is illegal (e.g., gender and race). On the other hand, indirect discrimination occurs when individuals appear to be treated based on seemingly neutral and non-protected attributes. However, protected groups or individuals may still be treated unjustly because of implicit effects from their protected attributes (e.g., salary information can be a proxy of gender, leading to discrimination).

The importance of fairness in AI-embedded systems was repeatedly highlighted in prior research and past cases of discriminating AI-embedded systems. An (in)famous example can be found in Amazon's former job application process, where AI-embedded human resource systems were used to screen job applications and identify promising candidates. To ensure fairness, an AI technology should not have access to attributes that could lead to (direct) discrimination, such as gender, race, or disability (Floridi, 2019; Thiebes et al., 2020). However, an incident at Amazon showed that such information might be not available in an explicit manner. Instead, the system's probabilistic algorithms may use other data as a proxy, including birthplace substituted for race, resulting in (indirect) discrimination (Feuerriegel et al., 2020; Hamilton, 2019). Given such incidents, users inevitably demand the fairness of AI-embedded systems to establish trust and ultimately adopt those systems (Shin, 2020; Shin & Park, 2019). Diverse (socio-technical) requirements regarding fairness must be fulfilled to avoid unfair bias and eliminate discrimination in decision-making processes (Feuerriegel et al., 2020). However, it is challenging to test AI-embedded systems for fairness and transparently signal fairness and non-discrimination in AI's decisions and recommendations to the users. We consider certifications, as these have already shown their added value in establishing trust in other technologies such as cloud computing (Lansing et al., 2018) and may also help establish trust in AI (Hengstler et al., 2016).

Certification of Information Systems

IS certifications are neutral third-party attestations of specific system characteristics, operations, and management principles to prove compliance with regulatory or industry requirements (Lansing et al., 2018).

In the context of IS use, organizations typically signal the possession of such certifications by placing assurance seals on their websites or in their system interfaces, thereby demonstrating their compliance with certification requirements.

The number and variety of IS certifications have increased continuously as the use of IS has diversified and expanded. Nowadays, well-known certifications include "Certified Privacy" for webshops, "CSA STAR" for cloud services, and management standards such as "ISO/IEC 27001—Information security management systems" for security. Recently, the EU GDPR and the EU Cybersecurity Act have foreseen certifications as the primary mechanism for organizations to demonstrate compliance with data protection and cybersecurity requirements across industries and legislative regions. More importantly, the current draft of the EU AI Act necessitates independent attestations of important technical characteristics and related system management practices of high-risk AI-embedded systems, such as AI-embedded systems intended to be used for biometric identification of natural persons or recruitment and selection of natural persons.

Prior consumer-related studies have primarily focused on three effects of IS certifications: increasing consumers' trust perceptions, purchase intentions, and perceived assurance by addressing privacy, security, or business integrity concerns of users (Adam et al., 2020; Löbbers et al., 2020). Novel certifications bear the potential to ensure transparency about non-discrimination and fairness in AI-embedded systems (Cremers, 2019). However, we still lack the understanding of how to design fairness certifications in AI-embedded systems due to diverse causes for discrimination and the complex and opaque nature of AI-embedded systems (Eiband et al., 2021; Feuerriegel et al., 2020). Therefore, with this study, we aim to identify and discuss important building blocks of fairness certifications to guide future research and practice and provide a first proof-of-concept for its feasibility.

Research Method

We adopted a two-step research approach. First, we reviewed existing literature on fairness to derive certification criteria and identify methods proposed by prior research to assess fairness. Afterwards, we performed interviews with AI experts to validate and deepen our understanding of fairness certifications derived from literature and investigate complex circumstances that the analyzed literature has not yet explored.

Literature Review

We aimed to review existing research on fairness to identify fairness issues, causes for discrimination, related fairness requirements and means to assess the fairness of AI proposed by prior research. Our descriptive literature review (Pare et al., 2015) was thereby guided by extant recommendations for literature reviews (Jeyaraj et al., 2006; Kitchenham et al., 2007; Webster & Watson, 2002). To identify publications addressing fairness of AI, we searched scientific databases, which cover the top IS conferences and journals (i.e., ACM Digital Library, EBSCOhost, IEEE Xplore, ProQuest, and ScienceDirect). To cover a broad set of publications, we searched each database with the following string in title and keywords: (AI OR Artificial Intelligence OR Machine Learning) AND (Fairness OR Discrimination OR Bias). We did not search in abstracts since this resulted in an excessive number of articles that were not related to our research objectives (e.g., abstracts stating that AI fairness is important). We limited our search to peer-reviewed articles to ensure a high quality of articles, yielding 411 articles as of May 2020.

To identify and filter articles, we first checked the relevance of each article by analyzing title, abstract, and keywords. If any indication for relevance appeared, the article was marked for further analysis. We excluded 363 articles that were duplicates (25), grey literature (i.e., editorials, work-in-progress) and books (2), not available in English (3), or not applicable to our study (333), such as articles only calling for fair AI but not providing any details. This first assessment resulted in a sample of 48 potentially relevant articles. Afterwards, a fine-grained relevance validation was made by accessing and reading the articles, resulting in a final sample of 17 relevant articles. In this second relevance assessment, we excluded non-research articles (10) and articles that only mentioned fairness, bias, or discrimination of AI-embedded systems but lacked a thorough discussion on them (21).

After the literature search was completed, we carefully read and analyzed the 17 articles to identify criteria and methods to ensure non-discrimination of AI-embedded systems. In particular, we searched the articles

for fairness requirements, recommendations to achieve fair AI, and reasons for unfairness and discrimination to identify potential criteria that can be checked during the certification to assess the systems' degree of fairness. Similar, we looked for metrics, methods and tools proposed by prior research to assess fairness. We recorded for each extracted criterion and method a name, a description, and the source, following the approach by Jeyaraj et al. (2006) and Lacity et al. (2010). A list of so-called master-variables was created to aggregate the identified criteria and methods across the articles. A master-variable is an aggregation of similar criteria or methods consisting of a name and a description (Jeyaraj et al., 2006; Lacity et al., 2010). If an identified criterion/method fit into an existing master-variable, we assigned it accordingly; otherwise, a new master-variable was created. In total, we identified 20 master-variables related to certification criteria (e.g., bias training for employees) and 15 master-variables related to assessing fairness (e.g., explorative data analysis).

Expert Interviews

We performed interviews with AI experts to validate and deepen our understanding derived from literature and investigate complex circumstances that the analyzed literature has not yet explored (Kuechler et al., 2009). We conducted nine semi-structured one-to-one expert interviews. To recruit potential interviewees, we applied a purposeful sampling strategy that focused on selecting individuals who are especially knowledgeable about our phenomenon of interest (i.e., fairness and non-discrimination in AI-embedded systems). Consequently, we included only experts who were engaged in AI-embedded systems with multiple years of experience in the AI domain (i.e., in average 3 to five years), such as operating in the automotive industry with a job position as a CTO [interview identifier I1], or manager [I5]. Additionally, we interviewed a sales manager from the insurance industry [I2], a consultant with a recruiting background [I3] and five people from an AI-related research field, including PhD researchers [I4 & I9], professors [I6 & I7], and an assistant professor [I8].

We applied a semi-structured interview method based on an interview guide (Yin, 2014). While a certain basic structure was necessary because we aimed to gather further information on identified master-variables from prior research, semi-structured interviews also leave interviewed experts with a sufficient degree of freedom to talk about aspects that might not have come to our attention (Myers, 2013). The interviews guide was constantly improved in terms of clarity and comprehensibility of the questions. We mostly asked interviewees what factors and causes are important when considering fairness in AI-embedded systems and how certification might help in guaranteed fairness and non-discrimination. We applied a non-judgmental form of listening, maintained distance, and strived to sustain an open and non-directive style of conversation to ensure impartiality and avoid bias (Myers, 2013; Yin, 2014). We recorded and transcribed each interview. The interviews lasted 29 minutes on average.

We conducted deductive, open, and theoretical coding (Corbin & Strauss, 2015) to analyze the interview data using ATLAS.ti 9. We first started with deductive coding by assigning master-variables identified in the literature review to the interview findings to validate findings from prior research and gather additional information. In this step, we created 35 codes for our master-variables and assigned them to 134 textual segments. Afterwards, we performed open coding and obtained additional 46 codes related to 220 textual segments. We tried to identify new information on fairness certifications that have been neglected in prior research so far during this step. For example, we coded the phrase "we realize again and again that we have completely different or very different approaches to problems." [I8] as "heterogeneity of the development team" as potential criteria to assess during the certification.

Finally, theoretical coding was used to integrate all findings into one' core category'; we formulated a storyline that coherently conceptualized the main phenomena (Corbin & Strauss, 2015; Urquhart et al., 2010). In doing so, we moved beyond description to a more abstract conceptualization level (Urquhart et al., 2010). We turned to the certification literature to identify a suitable core category and identified the typology of certifications' key structural building blocks and structural elements proposed by Lansing et al. (2018) as suitable categories to integrate our findings. Lansing et al. (2018) revealed that certifications typically comprise three building blocks: (1) content (i.e., the implied assurances), (2) source (i.e., the issuing and auditing authorities), as well as (3) process (i.e., the type of attestation process). In comparing our interview findings with this typology of certifications' key structural building blocks, we found a means to organize our results in a conceptually meaningful way. We assigned our codes and master-variables to related structural building blocks to build a first draft of a fairness certification. A summary of the building

blocks and its elements can be found in Figure 1. We will discuss our findings for each building block in detail in the following.

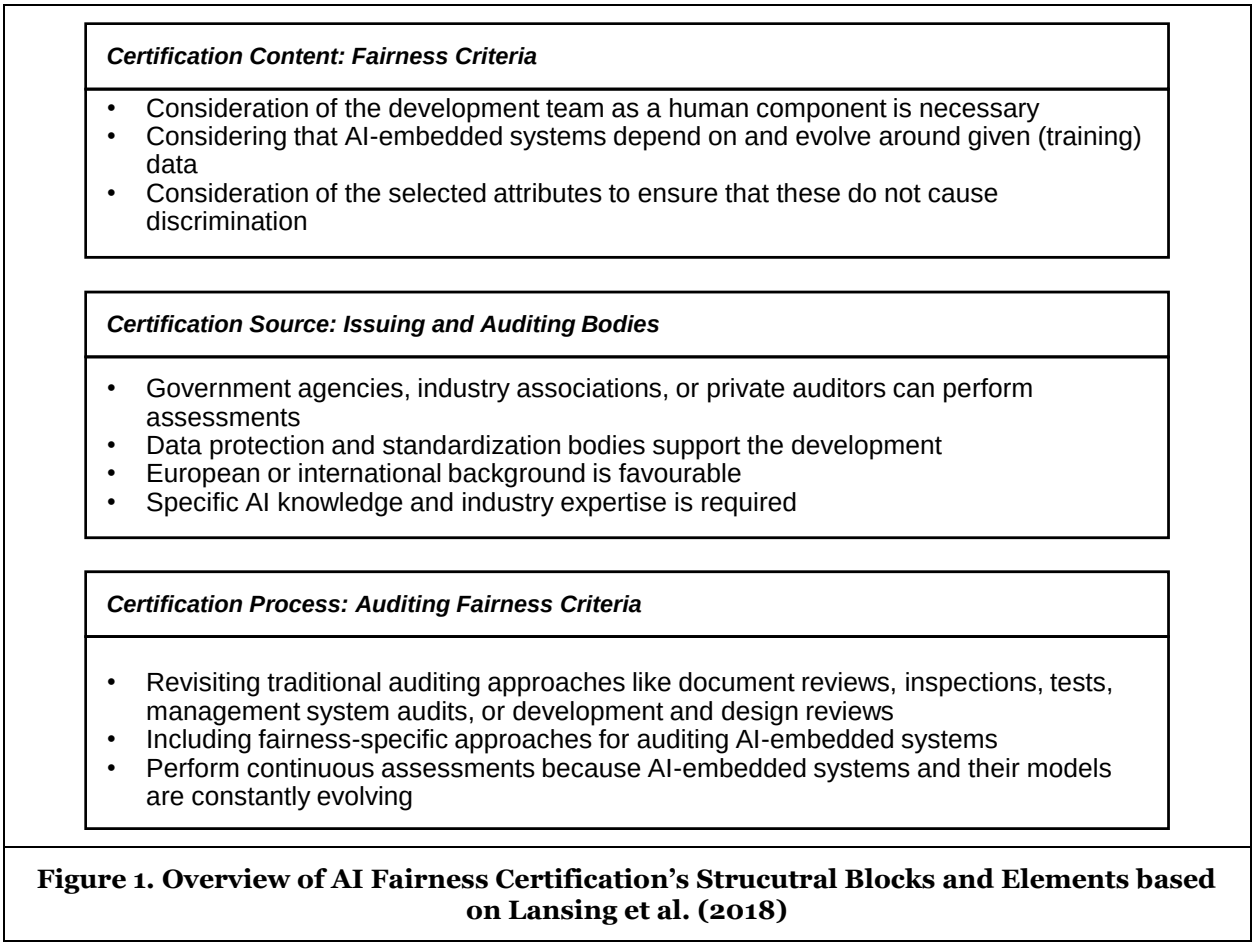

Figure 1. Overview of AI Fairness Certification's Strucutral Blocks and Elements based on Lansing et al. (2018)

Certification Content: Fairness Criteria

The first structural block of an AI fairness certification relates to its content and, most importantly, comprises "the specific certification dimensions and constituting evaluation criteria contained in a certification" (Lansing et al., 2018, p. 1073). We clustered identified criteria in three areas: development team, training data, and selection of attributes.

Development Team

With algorithms being developed and implemented by people, there is always a human component playing a large role in the algorithms' predictions, resulting in a potential risk of discrimination (Hagendorff, 2019, 2020).

Discrimination may occur due to a possible lack of awareness regarding (un-)consciously discriminatory assumptions that are made within the development and testing teams [stated by interviewees I3, I8, I9]. This scenario is also specified as "pre-existing bias" (Friedman & Nissenbaum, 1996) and is the result of a so-called uncomplex value enrolment process (Hagendorff, 2019). Besides, discrimination may occur due to "emerging bias" (Friedman & Nissenbaum, 1996) describing more complex and non-transparent value enrolment processes caused by machine learning techniques in AI-embedded systems (Hagendorff, 2019). Mostly, AI-embedded systems are not fully automatic working processes and are still intervened and supervised by people. People have different preferences, biases, and emotions. No decision concerning AI-

embedded systems may be neutral even if the development depends on stated decisions, inadvertently causing or reinforcing forms of social discrimination.

Nevertheless, prior research already provides recommendations on addressing these fairness issues. Whether these recommendations are followed can also be validated during the certification process. First, it is necessary to assemble a diverse development team and a diverse testing team while especially supporting normally underrepresented groups (Coates & Martin, 2019). Different aspects such as gender, ethnicity, ideologies, and experience should be incorporated, whereas the development team should represent the users of AI-embedded systems [I3, I8, I9]. Thus, it is possible to consider fairness criteria more comprehensively since "by looking at things from different perspectives, you have the option of discovering certain things or inconsistencies in advance" [I8]. Second, bias training (Cardenas & Vallejo-Cardenas, 2019) can sensitize possible prejudices or inequities to address those more effectively. In collaboration with sociologists, a common understanding of social injustice and the principles of fairness may be created [I9]. Third, these practices alone cannot be effective if there is no fundamental understanding and shared goal within the development team for the AI-embedded system (Rajkomar et al., 2018). This goal needs to be discussed and reviewed with underrepresented groups of people to directly identify discrimination so that it can be eliminated at the earliest possible stage. Finally, each person in a development team needs to question what biases exist in individual processes and how they can be eliminated or even reduced (DeBrusk & Wyman, 2020).

Training Data

Considering that AI-embedded systems depend on and evolve around given data, it is important to examine the very data used for training and development [I4, I6]. Historical data may not be neutral and may contain implicit or explicit biases (Cardenas & Vallejo-Cardenas, 2019). As a result, "training data that feed the machine learning components may be shifted in the direction in which society is currently shifted, or from the time in which data originated" [I6]. Thus, fair representativeness of all groups affected by an AI-embedded system in training datasets cannot always be guaranteed [I6].

In the course of a certification, it can be checked whether special attention was paid during data collection concerning the possibility of profiling individuals who will be directly or indirectly affected by the planned AI-embedded systems, as well as structurally disadvantaged groups based on, for example, ethnicity, gender, or disabilities (Hermanin & Atanasova, 2013). Besides, it should be considered "which data are needed for the training and which data are relevant" [I4]. Thus, before selecting the data, the task of the AI-embedded system must be precisely defined [I4]. Nevertheless, there is always the risk that not all relevant users or groups of users have been included in the data set [I4]. Thus, special attention must be paid to the quantity of the data [I1, I3]. An AI-embedded system can be trained extensively to make more targeted decisions given a large data set. Additionally, as our society becomes more diverse and ideally less discriminatory over time, it is important not to end the development of the AI-embedded system with its first use but periodically train and retrain it with new, updated data sets (DeBrusk & Wyman, 2020). Therefore, the certification criteria should also request an AI data management system that ensures ongoing fairness evaluation.

Selection of Attributes

Furthermore, an AI-embedded system relies on so-called attributes for predictions about the behavior of the corresponding users. These attributes describe various characteristics associated with the observed population. Thus, the selection of relevant attributes plays a crucial role during the development phase, as it directly influences the results (Bellamy et al., 2019).

An appropriate approach to select these attributes is essential in developing a fair AI-embedded system [I6, I7, I8]. Lack of knowledge about neglected groups of people can lead to discriminatory results, which must be prevented "at all costs" [I1, I3, I5, I6, I8]. Certification criteria to avoid discrimination can require the context-dependent definition of protected attributes [I3, I7]. The interviewees agreed that different attributes need to be protected depending on the application area. In addition, sensitive data may be used for positive discrimination, for example, by targeting hiring women to achieve gender balance. Nevertheless, AI-embedded systems should not access these attributes if they do not play a role in decision-making [I2].

Besides, the concept of "unawareness" may be used to protect attributes (Rajkomar et al., 2018). Protected attributes are recognized, isolated, and then not explicitly included in machine decision processes (Hagendorff, 2019, 2020). However, the deletion of sensitive data poses various problems. First, it may render AI-embedded systems less reliable because essential data for processing may be missing [I8, I9]. Second, eliminating sensitive data does not directly mean that discrimination can be curbed "as there is a correlation to indirectly acting attributes" and non-sensitive data [I7]. For example, the protected attribute of a person's ethnicity may be excluded, but ethnicity may be inferred from their place of residence in certain circumstances. Thus, discrimination may not be excluded, whereas protected characteristics and the unprotected need to be considered (Valentim et al., 2019). Third, the correlation may not be obvious, whereas unprotected and correlating attributes may not be eliminated.

Another approach may be oversampling to select attributes for ensuring fairness in an AI-embedded system (Parikh et al., 2019). In this method, data sets describing individuals from certain underrepresented subgroups are recognized, isolated, and multiplied based on the protected attributes. This results in (approximately) the same amount of data for all ethnic subgroups, and discrimination may be reduced. In contrast, undersampling mostly uses undersampling-multivariate methods, where a combination of labels and attributes is used as a starting point for sorting (Valentim et al., 2019). Although this method is very efficient and can work effectively against discrimination, there are always cases where unequal treatment cannot be prevented. Similar to eliminating protected attributes, this may also allow unprotected attributes to correlate with protected attributes (Valentim et al., 2019).

Certification Source: Issuing and Auditing Bodies

The second structural building block of an AI fairness certification relates to its source, namely the organizations involved in a certification's issuance and audit (Lansing et al., 2018). Whereas prior research agrees that a certification should be conducted by an independent body having high expertise, our literature review reveals no information on issuing and auditing bodies. Reflecting on the organizations typically involved in IS certifications (as revealed by our review of IS certification practices) and findings from our interviews provides first insights about who such bodies should be and what requirements they should fulfil.

Potential Issuing and Auditing Bodies of Fairness Certifications

A certification involves different bodies: (1) an issuer who may develop the criteria catalogue and certification process and finally awards each certification in case the AI-embedded system fulfils the criteria; and (2) auditors who conduct the actual certification audits to assess certification adherence of AI-embedded systems (Lansing et al., 2018).

In the context of AI certifications, government agencies, industry associations, and private auditors can take the role of issuing and auditing bodies. They can be further supported by standardization bodies like the International Organization of Standardization (ISO) and the European Committee for Electrotechnical Standardization (CENELEC). They are currently developing novel AI standards describing how to develop and test reliable, secure, and fair AI, among others. Yet, standardization bodies (most commonly) develop standards and criteria but do not audit or issue certifications themselves. Instead, private auditors and certification bodies (e.g., PwC, EY, TÜV, DQS) typically audit criteria compliance and issue certifications. We propose that these private organizations will also perform AI certifications soon to meet the ever-growing demand for AI certifications and generate further revenue while building on their favorable market reputation and multi-year expertise. In the context of AI certifications in general and fairness certifications, we also expect the emergence of auditing bodies that are specialized in certain industries or auditing techniques (e.g., performing inspections, penetration tests, or specific training data analyses). Indeed, subcontracting independent auditors is common for most certification bodies to gain access to additional expertise and react to various certification applications.

Interviewees also recommended that data protection bodies should be involved in the certification process because the underlying data protection law also addresses the "prevention of discrimination and stigmatization" [I7]. For example, the GDPR requires that certification bodies, who issue data protection certifications, will inform the supervising data protection authority before a new certification is awarded and if any violations are detected. A fairness certification may similarly consider the involvement of data protection authority in case discrimination results from data processing in AI-embedded systems.

Requirements for Issuing and Auditing Bodies

Interviewees agreed that issuing and auditing bodies should have profound knowledge in AI and rich experience in conducting socio-technical audits [I5, I6, I7, I9]. Since we currently lack best practices and internationally agreed-upon standards for setting up and assessing fair AI-embedded systems, an interviewee requested that the development of novel fairness certifications and its structural blocks should be a result of participatory and democratic processes involving various stakeholders with different backgrounds (e.g., having technical, legal, and organizational expertise) [I9]. Consequently, standardization bodies and research consortia should support certification issuing bodies since they are organized as technical committees that are made up of experts from the relevant industry, consumer associations, academia, NGOs and governments, and because they typically apply a consensus-based approach and take comments from all stakeholders into account. Besides, interviewees agreed that issuing bodies should have a European or international reach because modern AI-embedded systems are typically offered worldwide [I1, I6, I7].

Besides general knowledge in AI, auditing bodies should also have experience and knowledge in the specific application context, industry, or used AI technology (e.g., natural language processing or machine vision) [I5]. Therefore, diverse teams with varying and complementary expertise may conduct a certification audit to assess the system's degree of fairness. For example, it is recommended to include legal experts to facilitate certification audits that consider underlying regulations because definitions of fairness also often comprise "different ideas of justice" [I7].

Certification Process: Auditing Fairness Criteria

The third and last structural building block relates to the certification process, which defines how an auditor evaluates conformance of the AI-embedded system with the criteria (Lansing et al., 2018).

Revisiting Traditional Auditing Approaches

Existing certification practices typically rely on a set of auditing methods to assess adherence to certification criteria. Depending on the certification criterion, a fairness certification may include document reviews, inspections (refer to ISO/IEC 17020), tests (refer to ISO/IEC 17025), management system audits (refer to ISO/IEC 17021), or development and design reviews.

With a document review, an auditor verifies compliance with the certification criteria by reviewing process and system documents, (technical) logs, related certifications, or other documentation. For example, an auditor may verify that the development team has attended bias training through reviewing awarded training certificates. By conducting an inspection, auditors can interact with the system directly to assess, for example, the functioning and the results of the operations. An auditor typically compares the expected results according to the available documentation with the actual results produced by the system. The auditor may have no insight into the internal processing steps of the processing operations ('black box inspection') or may monitor the actual execution of functions to reveal the inner functioning ('white box inspection'). Further, auditors may perform tests to examine assets (e.g., hardware or software code). For example, auditors may perform security tests to determine the correct and strong encryption of stored training data or run penetration tests to provoke a discriminatory system behavior "to see where the borders are" and "at which point the system becomes [for example] racist" [I4]. Management system audits are used to identify that the system owner uses an organizational and technical system to manage the relevant aspects of its activities, products, and services following the organization's policy and the certification criteria. Auditors may conduct interviews, employee observations and tests to attest employees' knowledge and skills and whether processes and the management system are actually 'lived' as promised in the documentation. For example, interviewing or observing AI developers can provide direct evidence of their competence and whether they are aware of fairness objectives set by the organization and potential biases. Finally, a development and design review includes reviewing development methods and procedures and, if required, reviewing the test systems and environments used. The design review may consist of a review of the selected architecture, database diagrams, data flow diagrams, design decisions regarding attributes but also the configuration of the AI-embedded system.

Specific Approaches for Auditing AI-embedded Systems

Our literature review and interviews also revealed several fairness assessments methods that auditors may also use during a certification. First, auditors may calculate two core measures: the normalized or non-normalized difference, which is the mean difference for binary classification normalized by the rate of positive results (Žliobaitė, 2015). However, these core measures alone are not sufficient to measure fairness. They can only be applied to homogeneous populations, considering equal qualifications for a positive decision. In reality, this is rarely the case - for example, different levels of education can explain different salary levels. Thus, the main principle of applying core measures should be to first divide the population into more or less homogeneous segments according to their qualifications and then apply core measures within each segment (Žliobaitė, 2015).

Another approach for identifying unexpected and misleading patterns or variations that could lead to discrimination may be the explorative data analysis (EDA) (Hagendorff, 2019). For this purpose, unsupervised learning methods are used to develop algorithms that search for misleading patterns or deviations in data sets that could potentially lead to discrimination in later use. This approach also allows identifying important variables and missing data points or errors. Unsupervised learning and pedagogically interpretable algorithms may also enable the development of fairness hypotheses for further selective testing and exploration (Veale & Binns, 2017). By applying it one-sidedly to data without sensitive features, certain types of anomalous or potentially problematic patterns may be identified. However, this offers the slightest assurance that the fairness problems have been mitigated.

Besides, sensitivity tests may be used for auditing AI-embedded systems [I1, I4, I9]. Those tests include creating simulated datasets with a high number of excluded variables and performing counterfactual simulations (Parikh et al., 2019). Doing so allows determining how robust the predictions of bias are due to excluded variables.

Visual analysis systems may be used to discover prejudices in AI-embedded systems, such as FAIRVIS (Cabrera et al., 2019). It integrates a novel subgroup detection technology to verify fairness while demonstrating how interactive visualization may help understand non-discrimination. Intersectional bias is detected, and a novel subgrouping technique is used to recommend cross-sectional groups where a model may not be efficient. However, a major difficulty with fairness is that it is mathematically impossible to meet all definitions of fairness simultaneously if the population has different baseline rates, whereas trade-offs between fairness metric performance must be made.

Udeshi et al. (2018) developed Aeqitas, a system that may automatically detect discriminatory inputs that mark fairness violations for a given model of machine learning and a set of sensitive input parameters. Aeqitas is a fully automated and targeted audit generation strategy for the rapid generation of discriminatory inputs while including novel strategies for probabilistic searches over the input domain to detect fairness violations. In doing so, this approach uses the inherent robustness of common machine learning models to design and implement scalable test generation methods. However, Aeqitas has so far only been evaluated with discriminatory input characteristics concerning gender and does not provide a way to localize the cause of discrimination in an AI-embedded system (Udeshi et al., 2018).

A Need for Continuous Assessments

AI-embedded systems and their models are constantly evolving [I1, I4, I9]. Model predictions and subsequent actions should be continuously monitored to ensure that outputs do not amplify existing social distortions (Parikh et al., 2019). Thus, a certain amount of re-auditing will be needed after an initial assessment of fairness. While certifications typically comprise yearly surveillance audits, where the auditor perform at least spot-checks to verify ongoing adherence with the certification criteria, researchers recently proposed continuous certification (CC) approaches (e.g., Anisetti et al., 2020; Lins et al., 2016; Lins et al., 2019; Renner, Lins, & Sunyaev, 2021; Stephanow & Banse, 2017). CC refers to the consistent collection of certification-relevant data about an AI-embedded system, which is then aggregated and processed to validate systems' ongoing compliance with certification criteria (Lins et al., 2019). In general, CC builds on continuous monitoring and auditing and combines these approaches with additional mechanisms for transparent certification-relevant information. Researchers recently started to examine how to perform CC and proposed two complementary CC approaches: test-based CC (i.e., direct, external access of the auditing

body to the system and the underlying infrastructure to check system components and operations; Stephanow & Banse, 2017); and monitoring-based CC (i.e., the system owner monitors its system and infrastructure, collects data, and then makes the certification-relevant data available to the auditing body that will analyse the data automatically; Lins et al., 2019). Whereas research on CC is just in its early stages, we believe that CC approaches are valuable means to deal with the evolving nature of AI-embedded systems and therefore call for future research that reflects CC in the domain of certifying AI.

Discussion

Principal Findings

In this study, we briefly discussed three building blocks of certifications to assess the fairness of AI-embedded systems. Therefore, our study can be regarded as a first conceptualization to highlight the feasibility and usefulness of fairness certifications in the domain of AI.

Reflecting on the building blocks, we want to highlight notable discussion points. First, a socio-technical perspective on fairness is needed for the certification content. On the one hand, criteria should tackle organizational facets, such as confirming that the development team has built up a certain understanding of fairness. In this respect, systems' access to the required information should be justified by the development team through performing a risk assessment regarding potential effects on individuals and groups to be advocated by the operating companies. The risks are "proportional to the importance of the decision being made" [I4]. The use of AI-embedded systems in critical areas, such as in the processing of credit applications [I8] or the diagnosis of diseases, directly impacts the quality of life of individuals and groups and is therefore risky. On the other hand, the AI-embedded system itself should not be discriminatory. Consequently, the criteria should also consider technical facets, such as up-to-dateness of data and model drifts in machine learning. Finally, interviewees also expressed concerns about implementing fairness, which is often not specified in the underlying company-wide guidelines for AI-embedded systems [I3, I6]. This complicates the actual implementation of the fairness criteria, making it necessary to add further recommendations and implementation guidelines in certification criteria catalogues providing the development team with possible procedures for implementing the criteria.

Regarding the certification source, some interviewees raised concerns that providing access to auditing bodies implies that (sensitive) information about the development process must be disclosed [I7]. This demand for transparency about the development process may be problematic for organizations because it could reduce their competitive advantages [I3]. These concerns have been already expressed in other contexts, such as in case of cloud services certifications. However, in practice, confidentiality concerns are not an obstacle because the auditing and certification body sign non-disclosure agreements to reduce potential concerns of system owners. Dishonest and untrustworthy certification bodies may not survive in these competitive environments.

Finally, designing effective certification processes will be the most challenging open issue. The traceability of AI's decision-making processes is an important prerequisite for identifying discriminatory predictions. However, AI models are not always explainable (Barredo Arrieta et al., 2020). Besides, the dynamic nature of an AI-embedded system leads to changes during its runtime. Therefore, it is more difficult or even impossible for an auditing body to understand and assess the decision-making process. In highly complex systems that lack explainability, certifications may disclose discrimination but may not reveal its causes. Accordingly, to facilitate the certification process, documentation about the development should be designed so that an auditing body can gain detailed insights into the processes. The system provider should evaluate whether recently proposed explainability mechanisms may be implemented to ease third-party audits (Barredo Arrieta et al., 2020). Certification bodies should evaluate whether a continuous assessment is suitable to continuously evaluate whether the selected attributes do not discriminate individual groups.

Implications for Research and Practice

With our focus on ensuring fairness and tackling discrimination within AI-embedded systems, we study a novel field of research. Therefore, our key findings are relevant to both researchers and practitioners.

For researchers, our study serves as a first conceptualization of structural blocks for fairness certifications of AI-embedded systems. Thus, we yield a new perspective on how to signal fairness in AI-embedded systems by applying certifications while focusing on fairness as one of the most discussed AI-related issues (Shin, 2020; Shin & Park, 2019). Our work provides a starting point for understanding fairness certifications while providing descriptions on how to define each structural block. We revealed similarities to other IS certifications in related contexts (e.g., concerning the certification process and the issuer) and fairness specifics (e.g., novel criteria are required). Our findings highlight that a socio-technical perspective is needed (i.e., development team, training data, and selection of attributes). This may also be relevant for other AI-related issues such as accountability. Moreover, we have shown the importance of the content of the certifications and revealed that similar to certifications for cloud services, the certification process and the certification body are relevant (Lansing et al., 2018). It is particularly noticeable that continuous auditing and monitoring are necessary due to the rapid change of training data and the AI models themselves.

For practitioners, the implications of our research depend on the different stakeholders in the context of fair AI-embedded systems, such as providers, users, or certification bodies. Providers may implement the aggregated list of criteria and questions to assess the fairness of their AI-embedded systems and derive approaches to improve their systems. Certification bodies may integrate our proposed criteria in their catalogues and reflect whether other auditing methods are suitable. For users, we want to provide first validation that a fairness certification is feasible, providing users in the future with means to determine the fair degree of AI-embedded systems.

Limitations and Future Research

Our study is not without limitations. First, the research on fairness in AI and related certifications is at an early stage. An ever-increasing amount of research is published related to the fairness of AI. Therefore, our literature review and the resulting criteria catalogue only covers criteria and recommendations identified until mid-2020 and should be frequently updated. Second, our study has limitations concerning the number and depth of our interviews to enrich our study. While we conducted nine expert interviews, future research might focus on gathering more information to increase understanding. The statements made by the interviewees only reflect one opinion in their field and are therefore neither representative nor fully impartial. Third, the number of different definitions for fairness and non-discrimination imply that fairness cannot be assessed by a limited group of people. Future research needs to be done on finding a commonly accepted definition that could be a starting point for the certification project.

Our results pave the way for several future research opportunities. First, the opinion of users can also be included when developing IS certifications to elaborate on what is important for consumers regarding fairness certifications (Löbbers et al., 2020). We have shown that certifications may be a powerful means to establish trust in AI-embedded systems. While we have focused on fairness, future research could also address other AI-specific issues, such as signaling the accountability of AI-embedded systems to consumers.

Conclusion

While AI-embedded systems may offer many benefits, they also come with risks, such as individual privacy risks or the presence of racial bias. Given such fairness risks and related issues with AI, users are still hesitant to adopt AI-embedded systems. Independent certifications have shown to be a promising tool in other contexts to signal high system quality and, therefore, may also be used to assess and communicate an AI-embedded system's degree of fairness. However, research and practice still struggle with designing and performing fairness certifications. With our research, we provide a first overview of certification's content and related criteria, the source represented by the issuers and auditors, and the process including evaluating fairness and including traditional auditing methods. To this end, our main contribution lies in the first conceptualization of important building blocks for a fairness certification to highlight the feasibility and usefulness of fairness certifications in the domain of AI.

References

Adam, M., Niehage, L., Lins, S., Benlian, A., & Sunyaev, A. (2020). Stumbling Over the Trust Tipping Point – The Effectiveness of Web Seals at Different Levels of Website Trustworthiness. European Conference on Information Systems, Marakesh, Maroko.

Adam, M., Wessel, M., & Benlian, A. (2021). AI-Based Chatbots in Customer Service and their Effects on User Compliance. *Electronic Markets*, *31*(2), 427-445. https://doi.org/10.1007/s12525-020-00414-7

Anisetti, M., Ardagna, C. A., Damiani, E., & Gaudenzi, F. (2020). A Semi-Automatic and Trustworthy Scheme for Continuous Cloud Service Certification. *IEEE Transactions on Services Computing*, *13*(1), 30-43. https://doi.org/10.1109/TSC.2017.2657505

Bantilan, N. (2017). Themis-ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation. *Journal of Technology in Human Services*, *36*, 15-30. https://doi.org/10.1080/15228835.2017.1416512

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671-732. http://www.jstor.org/stable/24758720

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development*, *63*(4/5), 1-15. https://doi.org/10.1147/JRD.2019.2942287

Cabrera, n. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J. H., & Chau, D. H. (2019). FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 46-56.

Cardenas, S., & Vallejo-Cardenas, S. (2019). Continuing the Conversation on How Structural Racial and Ethnic Inequalities Affect AI Biases. *IEEE International Symposium on Technology and Society (ISTAS)*, Medford, MA, USA.

Coates, D. L., & Martin, A. (2019). An Instrument to Evaluate the Maturity of Bias Governance Capability in Artificial Intelligence Projects. *IBM Journal of Research and Development*, *63*(4/5), 1-15. https://doi.org/10.1147/JRD.2019.2915062

Corbin, J. M., & Strauss, A. L. (2015). *Basics of Qualitative Research*. SAGE.

Cremers, A. B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Rosenzweig, J., Rostalski, F., Sicking, J., Volmer, J., Voosholz, J., Voss, A., Wrobel, S. (2019). Trustworthy Use of Artificial Intelligence: Priorities from a Philosophical, Ethical, Legal, and Technological Viewpoint as a Basis for Certification of Artificial Intelligence. In *https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf*

DeBrusk, C., & Wyman, O. (2020). The Risk of Machine Learning Bias (And How to Prevent It). How AI Is Transforming the Organization MIT Press.

Eiband, M., Buschek, D., & Hussmann, H. (2021). How to support users in understanding intelligent systems? Structuring the discussion. 26th International Conference on Intelligent User Interfaces, College Station, TX, USA.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, *542*, 115-118. https://doi.org/10.1038/nature21056

Federal Office for Information Security. (2021). AI Cloud Service Compliance Criteria Catalogue (AIC4). https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile&v=4

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI. *Business & Information Systems Engineering*, *62*(4), 379-384. https://doi.org/10.1007/s12599-020-00650-3

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261-262. https://doi.org/10.1038/s42256-019-0055-y

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach (Dordr)*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Hagendorff, T. (2019). Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze. *Österreichische Zeitschrift für Soziologie*, *44*(1), 53-66. https://doi.org/10.1007/s11614-019-00347-2

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hamilton, M. (2019). Debating Algorithmic Fairness. *UC Irvine Law Review*, *52*, 261-296. https://doi.org/10.5064/F6JOQXNF

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105-120. https://doi.org/10.1016/j.techfore.2015.12.014

Hermanin, C., & Atanasova, A. (2013). Making "Big Data" Work for Equality. In. https://www.opensocietyfoundations.org/voices/making-big-data-work-equality-0 (last access on 10.11.2021): https://www.opensocietyfoundations.org/voices/making-big-data-work-equality-0

Independent High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf

Jeyaraj, A., Rottman, J., & Lacity, M. (2006). A Review of the Predictors, Linkages, and Biases in IT Innovation Adoption Research. *Journal of Information Technology*, *21*, 1-23. https://doi.org/10.1057/palgrave.jit.2000056

Kitchenham, B. A., Mendes, E., & Travassos, G. H. (2007). Cross versus Within-Company Cost Estimation Studies: A Systematic Review. *IEEE Transactions on Software Engineering*, *33*(5), 316-329. https://doi.org/10.1109/TSE.2007.1001

Kuechler, W. L., Park, E. H., & Vaishnavi, V. K. (2009). Formalizing Theory Development in IS Design Science Research: Learning from Qualitative Research. AMCIS,

Lacity, M. C., Khan, S., Yan, A., & Willcocks, L. P. (2010). A review of the IT outsourcing empirical literature and future research directions. *Journal of Information Technology*, *25*(4), 395-433. https://doi.org/10.1057/jit.2010.21

Lansing, J., Benlian, A., & Sunyaev, A. (2018). "Unblackboxing" Decision Makers' Interpretations of IS Certifications in the Context of Cloud Service Certifications. *Journal of the Association for Information Systems*, *19*(11), 1064-1096. https://doi.org/https://doi.org/10.17705/1jais.00520

Lins, S., Grochol, P., Schneider, S., & Sunyaev, A. (2016). Dynamic Certification of Cloud Services: Trust, but Verify! *IEEE Security & Privacy*, *14*(2), 66-71. https://doi.org/10.1109/MSP.2016.26

Lins, S., Schneider, S., Szefer, J., Ibraheem, S., & Sunyaev, A. (2019). Designing Monitoring Systems for Continuous Certification of Cloud Services: Deriving Meta-requirements and Design Guidelines. *Communications of the Association for Information Systems*, *44*, 406-510. https://doi.org/10.17705/1CAIS.04425

Lins, S., & Sunyaev, A. (2017). Unblackboxing IT Certifications: A Theoretical Model Explaining IT Certification Effectiveness. In *Proceedings of the 38th International Conference on Information Systems*: Transforming Society with Digital Innovation, ICIS 2017; CoexSeoul; South Korea; 10 December 2017 through 13 December 2017,

Löbbers, J., & Benlian, A. (2019). The Effectiveness of IS Certification in E-Commerce: Does Personality Matter? *Journal of Decision Systems*, *28*, 1-27. https://doi.org/https://doi.org/10.1080/12460125.2019.1684867

Löbbers, J., Lins, S., Kromat, T., Benlian, A., & Sunyaev, A. (2020). A Multi-Perspective Lens on Web Assurance Seals: Contrasting Vendors' Intended and Consumers' Perceived Effects. *Electronic Commerce Research*. https://doi.org/https://doi.org/10.1007/s10660-020-09415-2

Matus, K., & Veale, M. (2021). Certification Systems for Machine Learning: Lessons from Sustainability. *Regulation & Governance*. https://doi.org/10.1111/rego.12417

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1-35. https://doi.org/10.1145/3457607

Miller, D. D., & Brown, E. W. (2018). Artificial Intelligence in Medical Practice: The Question to the Answer? *The American Journal of Medicine*, *131*(2), 129-133. https://doi.org/https://doi.org/10.1016/j.amjmed.2017.10.035

Morik, K., Kotthaus, H., Heppe, L., Heinrich, D., Fischer, R., M ̧cke, S., Pauly, A., Jakobs, M., & Piatkowski, N. (2021). Yes We Care! - Certification for Machine Learning Methods through the Care Label Framework. *ArXiv*, *abs/2105.10197*.

Myers, M. D. (2013). *Qualitative Research in Business & Management*. Sage Publ, London.

Pare, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing Information Systems Knowledge: A Typology of Literature Reviews. *Information & Management*, *52*, 183-199. https://doi.org/10.1016/j.im.2014.08.008

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. *Jama*, *322*(24), 2377-2378. https://doi.org/10.1001/jama.2019.18058

Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's Comments: Next-Generation Digital Platforms: Toward Human–AI Hybrids. *MIS Quarterly*, *43*(1), iii-ix.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med*, *169*(12), 866-872. https://doi.org/10.7326/m18-1990

Renner, M., Lins, S., Söllner, M., Thiebes, S., & Sunyaev, A. (2021). *Achieving Trustworthy Artificial Intelligence: Multi-Source Trust Transfer in Artificial Intelligence-capable Technology* 42nd International Conference on Information Systems (ICIS), Austin, TX, USA.

Renner, M., Lins, S., Söllner, M., Thiebes, S., & Sunyaev, A. (2022). *Understanding the Necessary Conditions of Multi-Source Trust Transfer in Artificial Intelligence* 55th Hawaii International Conference on System Sciences (HICSS), Online.

Renner, M., Lins, S., & Sunyaev, A. (2021). A Taxonomy of IS Certification's Characteristics. In *2021 2nd International Conference on Internet and E-Business* (pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3471988.3471989

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system:Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541-565. https://doi.org/10.1080/08838151.2020.1843357

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277-284. https://doi.org/10.1016/j.chb.2019.04.019

Stephanow, P., & Banse, C. (2017, 14-17 May 2017). Evaluating the Performance of Continuous Test-Based Cloud Service Certification. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, Spain.

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Udeshi, S., Arora, P., & Chattopadhyay, S. (2018). Automated Directed Fairness Testing. *33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 98-108.

Urquhart, C., Lehmann, H., & Myers, M. D. (2010). Putting the Theory Back into Grounded Theory: Guidelines for Grounded Theory Studies in Information Systems. *Information Systems Journal*, *20*(4), 357-381. https://doi.org/10.1111/j.1365-2575.2009.00328.x

Valentim, I., LourenÁo, N., & Antunes, N. (2019). The Impact of Data Preparation on the Fairness of Software Systems. *IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, 391-401.

Veale, M., & Binns, R. (2017). Fairer Maschine Learning in The Real World: Mitigating Discrimination Without Collecting Sensitive Data. *Big Data & Society*, *4*, 1-17. https://doi.org/10.1177/2053951717743530

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii-xxiii. http://www.jstor.org/stable/4132319

Yin, R. K. (2014). *Case study research. Design and methods*. SAGE.

Žliobaitė, I. (2015). A Survey on Measuring Indirect Discrimination in Machine Learning. *ArXiv*, *abs/1511.00148*.

Zou, J., & Schiebinger, L. (2018). AI can be Sexist and Racist - It's Time to Make it Fair. *Nature*, *559*, 324-326. https://doi.org/10.1038/d41586-018-05707-8

Fairness of Medical Artificial Intelligence: A Literature Review

Emerging Trends in Digital Health, Summer Term 2021

Jasmin Eipper
Master Student
Karlsruhe Institute of Technology
ufruy@student.kit.edu

Amelie Schwärzel
Master Student
Karlsruhe Institute of Technology
amelie.schwaerzel@student.kit.edu

Justus Thiel
Master Student
Karlsruhe Institute of Technology
Justus.thiel@student.kit.edu

Luisa Weber
Master Student
Karlsruhe Institute of Technology
ugrif@student.kit.edu

Abstract

Background: *With the increasing digitization and growing integration of Artificial Intelligence (AI) in healthcare (e.g. for diagnosis), the risk of incorporating unfair and biased behaviours in medical applications caused by biased AI is rising. Thus, the integration of AI in healthcare requires the need to develop trustworthy and unbiased applications that make fair decisions regardless of a patient's demographic profile.*

Objective: *To contribute to the development of fair and unbiased AI applications in healthcare, the aim of this paper is to provide an overview of the current state of research in the emerging research field on fairness of medical AI. In particular, it aims to elaborate issues and causes related to the occurrence of biased AI and provide possible solutions.*

Methods: *To provide the overview of the extant literature, a systematic literature review was conducted. By searching in eight scientific databases (ACM Digital Library, AIS Electronic Library, arXiv, IEEE Explore, Ebsco Buisness Source, medRxiv, Pubmed, Scopus) and applying inclusion and exclusion criteria, 35 relevant articles have been identified and further analyzed.*

Results: *The emergence of biased AI applications cannot be considered solely as a data or a system problem. Accordingly, the occurrence of unfair AI behaviors has different reasons. For example, under- or over-representation of a particular group in an AI's training data can lead to bias, as it is not a baseline that is representative of the entire population. Unclear instructions to an algorithm lead to incorrect process steps. In addition, the black-box nature of many AIs makes it difficult to identify problem areas due to a lack of transparency. By learning and forming false correlations, models also cause biases. In addition, models must make a trade-off between fairness, model performance, and model explainability, which make the development of a fair, powerful, and transparent AI an NP-hard optimization problem. The interfaces of developers and contributors as well as users with the AI are equally a cause of the emergence of unfair medical AI.*

Conclusion: *Developing AI for healthcare applications is a complex problem that must holistically consider and optimize the interplay of bias as a data and system problem.*

Keywords: artificial intelligence, medicine, healthcare, fairness, bias, taxonomy

Einleitung

Das Gesundheitswesen gilt als einer der vielversprechendsten Bereiche für den Einsatz von Systemen der künstlichen Intelligenz (KI) (Deloitte 2019). Lernende maschinelle Systeme haben das Potenzial, viele Aspekte der Patientenversorgung zu verändern und administrative Prozesse im Gesundheitssektor zu verbessern (Davenport & Kalakota 2019). So werden Technologien zur Diagnose von Krankheiten und Applikationen in der medizinischen Bildgebung eingesetzt, um Krankheiten frühzeitig zu erkennen (Sonntag 2019) oder um Patienten, Ärzte und das Gesundheitssystem in Form von computergestützten Hilfssystemen zu entlasten (Sonntag 2019).

Neben den möglichen Vorteilen weist eine Reihe an Forschungsarbeiten neuerdings jedoch vor allem ethische Bedenken hinsichtlich des Einsatzes der KI-Technologien im medizinischen Sektor auf (Morley et al. 2020). Künstliche Intelligenzen zeigen diskriminierende Verhaltensweisen. So deckte die Analyse eines in US-Krankenhäusern weit verbreiteten Algorithmus zur Zuteilung von Gesundheitsleistungen auf, dass dieser dunkelhäutigen Personen systematisch diskriminiert (Obermeyer et al. 2019). Die Studie stellte fest, dass dieser Algorithmus Dunkelhäutige trotz schwerwiegenderer Krankheit mit geringerer Wahrscheinlichkeit als Weiße an Programme überweist, welche die Versorgung von Patienten mit komplexen medizinischen Bedürfnissen verbessern sollen (Obermeyer et al. 2019).

Im Allgemeinen können das Gesundheitswesen und die Medizin auf eine lange Geschichte der Voreingenommenheit und Diskriminierung zurückblicken. Zahlreiche Studien belegen, dass Patienten, welche einer Minderheit angehören, trotz ähnlicher Schwere der Krankheit, klinischer Vorstellung und Krankenversicherung eine schlechtere Versorgungsqualität zukommt (Ahmad et al. 2020). So werden beispielsweise Personen schwarzer Hautfarbe im Vergleich zu Weißen oder Latinos seltener mit Opioiden behandelt, wodurch sie eine schlechtere Behandlungsqualität bei gleichen Krankheitssymptomen erfahren (Tamayo-Sarver et al. 2003). Bereits Martin Luther King Jr. beschrieb die Ungerechtigkeit im Gesundheitsbereich als "[…] die schockierendste und unmenschlichste [Form der Ungleichheit], weil sie oft zum physischen Tod führt" (Martin Luther King Jr., Chicago, 25. März 1966).

Ausgehend von diesem Hintergrund besteht durch den gestiegenen Einsatz der KI-Anwendungen im medizinischen Sektor die Gefahr, dass mögliche existierende unfaire Verhaltensweisen und Ungleichheiten in der medizinischen Versorgung verschärft werden. Demnach bedarf es der Notwendigkeit der Entwicklung fairer Algorithmen und Modelle, welche für alle Patienten unterschiedlicher Herkunft, unterschiedlichen Alters oder Geschlechts dieselben Ergebnisse liefern. Um die Fairness solcher Applikationen zu gewährleisten und die Anwendung der KI im Gesundheitswesen weiterhin vertreten zu können, muss erkannt werden, wodurch mögliche Verzerrungen sowie unfaires Verhalten einer KI entstehen und wie dieses behoben werden können. Um der Entwicklung fairer und unvoreingenommener KI-Anwendungen beizutragen, gilt es als Ziel dieser Arbeit, einen Überblick über den aktuellen Forschungsstand des jungen und dynamischen Forschungsfeldes zur Fairness medizinischer KI bereitzustellen. Um aufzuzeigen, welche Erkenntnisse die vorhandene Forschungsliteratur liefert, wird im Rahmen des vorliegenden Literaturreviews die relevante Forschungsliteratur identifiziert sowie ausgewertet und kategorisiert. Im Detail sollen in der vorliegenden Arbeit folgende Forschungsfragen beantwortet werden:

1. In welchen Fällen tritt unfaire medizinische KI auf?
2. Wodurch wird dieses unfaire Verhalten verursacht?
3. Wie kann man unfairer medizinischer KI begegnen?

Zur Erschließung der aufgeworfenen Fragestellungen gilt das erste Kapitel der theoretischen Fundierung, weshalb in diesem die relevanten konzeptionellen Grundlagen des Themas aufgeführt werden. Dafür sollen die Themen der Künstlichen Intelligenz sowie der Fairness zu ihren jeweiligen Besonderheiten charakterisiert und zusammengeführt werden, indem das grundlegende Problem der Fairness von Künstlicher Intelligenz dargelegt wird. Neben der Erläuterung der methodischen Vorgehensweise in Kapitel drei folgt im vierten Abschnitt dieser Arbeit ein systematischer Überblick über den aktuellen Stand der Forschung zum Thema der Fairness medizinischer KI. Um dem Forschungsfeld beitragen zu können, werden im fünften Kapitel die zuvor aufgestellten Forschungsfragen anhand der erarbeiteten Literatur beantwortet. Kapitel sechs schließt die vorliegende Arbeit ab, indem die zentralen Ergebnisse präsentiert

und diskutiert werden. Auf Grundlage des Theorieteils und den Ergebnissen der vorliegenden Untersuchung werden schließlich zentrale Implikationen für Forschung und Praxis identifiziert.

Theoretischer Hintergrund zu Künstlicher Intelligenz und Fairness

Zur Erstellung eines gemeinsamen Verständnisses werden in diesem Kapitel zunächst essenzielle Begriffsdefinitionen und Konzepte eingeführt. Die folgenden Abschnitte dienen dazu, einen Überblick über Künstliche Intelligenzen und Fairness zu geben, um die Konzepte im Anschluss in Verbindung bringen zu können, indem das grundlegende Problem der Fairness von Anwendungen der Künstlichen Intelligenz sowohl im Allgemeinen als auch im medizinischen Kontext erläutert wird.

Definition Zentraler Begriffe

Künstliche Intelligenz (KI)

Künstliche Intelligenz (KI) ist ein Überbegriff für Algorithmen, welche „intelligentes" Verhalten zeigen (Baumgartner 2021). Folglich beschreibt KI Computerprogramme, welche sich selbst steuern und infolgedessen Aufgaben unterschiedlicher Domänen lösen können, für welche unter normalen Gegebenheiten menschliche Intelligenz erforderlich wäre (Kellmeyer 2019). Aus diesem Grund wächst der Umfang ihres Anwendungsbereiches stetig und unterstützt zusehends Menschen bei Entscheidungen (Markus et al. 2021). Das Machine Learning (ML) gilt dabei als der sich am schnellsten entwickelnde Teilbereich der KI (Baumgartner 2021). Im Allgemeinen beschreibt ML Lernen, ohne dies zu programmieren (Kellmeyer 2019). Methoden des ML ermöglichen Algorithmen, durch das Lernen charakteristischer Merkmale in den Daten, Muster zu erkennen, welche ursprünglich nicht in den Programmieranweisungen auftauchen. Entsprechend kann sich ein ML-Algorithmus aufgrund seiner Trainingsdaten weiterentwickeln und anhand dieser lernen, wie eine Kategorisierungs- oder Regressionsaufgabe möglichst gut zu lösen ist. Maßgebliche Einflussfaktoren für die Qualität der Ergebnisse sind hierbei neben der Verfügbarkeit des Datensatzes, auch dessen Qualität und Diversität. Jedoch muss beachtet werden, je mehr Freiraum das System bekommt und je selbstständiger es arbeitet, um Aufgaben optimiert zu lösen, desto schwieriger ist für den Anwender die Art und Weise der Lösungsbestimmung des Systems nachzuvollziehen (Baumgartner 2021).

Fairness und Bias

Allgemein versteht man unter dem Begriff Fairness die Abwesenheit von Vorurteilen oder Bevorzugung einer Person oder Gruppe aufgrund ihrer angeborenen oder erworbenen Eigenschaften (Mehrabi et al. 2019). Das erwünschte vorurteilsfreie Verhalten von Personen tritt in unterschiedlichsten Lebensbereichen verschiedener Situationen auf. Dabei gerät in Abhängigkeit der Diskriminierungsart eine andere angeborene oder erworbene Eigenschaft des betroffenen Subjekts in Bedrängnis, wie beispielsweise das Geschlecht oder die ethnische Herkunft (Seyyed-Kalantari et al. 2020). Beispielsweise definiert Fairness im medizinischen Kontext eine gesundheitliche Chancengleichheit, welche auch sozial benachteiligten Gruppen, wie etwa ethnischen Minderheiten, die gleiche medizinische Versorgung ermöglicht wie sozial starken Gruppen (Braveman 2006). Ein weiteres Beispiel ist die algorithmische Fairness in der KI. Diese ist gewährleistet, wenn verschiedene Personengruppen unabhängig von ihren persönlichen Merkmalen ein gleiches Ergebnis erhalten (Baumgartner 2021). An den aufgeführten Beispielen ist klar zu erkennen, dass Fairness eine Erscheinung der Gerechtigkeit ist.

Demgegenüber steht die Voreingenommenheit durch existierende Vorurteile, wodurch ein unfaires Verhalten auftritt. Vorurteile wiederum sind in den Verzerrungen der Sicht-/Verhaltensweise einer Person begründet. Aus diesem Grund ist der Gegenspieler von Fairness die Verzerrung (Bias) (Baumgartner 2021). Ein Bias stellt somit eine „durch falsche Untersuchungsmethoden […] verursachte Verzerrung […]" des Ergebnisses dar (Duden 2021).

Algorithmus und Modell

Ein Algorithmus ist als eine schrittweise Prozedur zur Lösung eines Problems definiert, die auf Daten trainiert wird (Panch et al. 2019). Dabei findet die Problemlösung des Algorithmus durch die Befolgung schrittweiser Anweisungen statt. Die Befehlsausführungen werden dabei so befolgt, dass die Eingabedaten

immer nach dem gleichen Schema in Ausgabedaten umgewandelt werden (Baumgartner 2021). Für den sachgerechten Ablauf lassen sich dabei wichtige Eigenschaften an einen Algorithmus stellen (Kellmeyer 2019):

1. die Menge der Anweisungen muss eindeutig und widerspruchsfrei sein
2. jeder Schritt muss realisierbar sein
3. die Beschreibung muss endlich sein
4. der letzte Schritt soll ein Ergebnis hervorbringen
5. bei einer Wiederholung unter genau denselben Umständen wird dasselbe Ergebnis der Prozedur erzeugt
6. bei jedem gegebenen Schritt in der Prozedur gibt es nur eine Möglichkeit weiterzumachen

Aufgrund der schematischen Abarbeitung der Prozedur wird einem Algorithmus eine objektive Problemlösung zugesprochen (Baumgartner 2021). Um die Objektivität zu gewährleisten, muss ein Algorithmus zudem so konstruiert sein, dass er sensitiv für Bias ist (Abbasi-Sureshjani et al. 2020). Einige Algorithmen zur Erstellung einer KI gelten als Black-Boxen, da sie keinen transparenten Aufschluss in logischer oder mechanistischer Form über ihre internen Funktionsweisen geben. Dies ist vor allem für eine Überprüfung mit ex-ante und post-hoc-Inspektionen im Falle einer Fehlfunktion von Nachteil (Ienca & Ignatiadis 2020).

Modelle werden durch Daten angetrieben und liefern die Ergebnisse eines Algorithmus für maschinelles Lernen (Chen et al. 2019). Folglich erzeugen sie Vorhersagen (Markus et al. 2021), dienen zur Diagnose (Chen et al. 2019), können Informationen klassifizieren und geben den Daten eine Sinnhaftigkeit. Sie modellieren komplexe Input-Output Abhängigkeiten, ohne vorherige explizite und detaillierte Informationen über sie zu benötigen (Ienca & Ignatiadis 2020). Dies fördert einerseits die Reduktion der Subjektivität in der Dateninterpretation, was beispielsweise schnellere medizinische Entscheidungen zulässt (Ienca & Ignatiadis 2020), andererseits kommt es zu Kausalitätsproblemen mit dem Risiko der Ableitung kausaler Beziehungen von kaum korrelierten Daten (Ienca & Ignatiadis 2020). Wenn ein KI-Modell, welches mit einem gegebenen Datensatz trainiert wurde, Ergebnisse produziert, die nicht von den Entwicklern intendiert waren, sinkt die Vertrauenswürdigkeit des Modells (Chen et al. 2019). Allerdings ist ein vertrauenswürdiges Modell besonders im medizinischen Kontext für die Vorhersage von Krankheiten wichtig (Afrose et al. 2021). Modellerklärbarkeit und -interpretierbarkeit spielen hier eine zentrale Rolle (Gade et al. 2020).

Fairness Künstlicher Intelligenz

Die Fairness im Gesundheitswesen gilt als vielseitiges Problem, das eine genaue Abwägung zwischen verschiedenen Fairnesskonzepten in der KI und unterschiedlichen Vorstellungen von Fairness im Gesundheitswesen erfordert (Ahmad et al. 2020). Demnach wird in den folgenden Unterkapiteln näher auf den Zusammenhang von Künstlicher Intelligenz und Fairness eingegangen, um dies anschließend im medizinischen Kontext weiter kontrahieren zu können.

Fairness von KI im Allgemeinen Kontext

Künstliche Intelligenz zielt durch berechenbare, rationale Algorithmen darauf ab, den Entscheidungsprozess von emotionalen Anwendern zu beschleunigen und von Vorurteilen zu befreien (Marcinkowski & Starke 2019). Es wird allerdings festgestellt, dass diese technischen Programme nicht frei von Verzerrungen sind. Neben algorithmischen Verzerrungen und Bias in den Programmmodellen ist die Verwendung von nicht repräsentativen/verzerrten Daten beim Trainieren der Applikation ein häufig vorkommender Grund für das Auftreten von Bias in der KI. Das Resultat sind unfaire Ergebnisse (Brault & Saxena 2020). Allgemein sollten jedoch die Chancen auf Gleichbehandlung für jedes Individuum gegeben sein. Dementsprechend sollte vor der Verwendung einer KI analysiert werden, ob es sich dabei um ein faires Verfahren handelt und somit überhaupt für den breiten Einsatz geeignet ist.

Eben dies stellt die Wissen- und Gesellschaft vor große Herausforderungen. Als Beispiel dient das von Mehrabi et al. (2019) beschriebene Tool – Correctional Offender Management Profiling for Alternative Sanctions – kurz COMPAS. COMPAS ist eine KI-Applikation, die in den USA verwendet wird, um die Wahrscheinlichkeit einer erneuten Festnahme von verhafteten Individuen zu schätzen und dementsprechend ein Strafmaß festzulegen. Dabei werden dunkelhäutigen Bürger mit einer deutlich

höheren Wahrscheinlichkeit als erneut straffällig eingestuft als hellhäutige Mitbürger (Mehrabi et al. 2019). Es ist offensichtlich, dass eine KI nicht in die Zukunft blicken und somit falschliegen kann. Das wirft die generelle Frage auf, ob der Einsatz einer solchen Applikation grundlegend sinnvoll ist. Um eine solche Frage beantworten zu können müssen verschiedene Perspektiven auf Fairness und Bias betrachtet werden.

Mehrabi et al. (2019) definieren allein 23 verschiedene Arten von Bias, die sich in KI-Anwendungen niederschlagen können. Ein sehr zentraler Bias ist laut Mehrabi et al. (2019) der historische Bias, der bereits in soziodemografischen Problemen der Gesellschaft vorhanden ist und selbst bei perfekter Datenerhebung und feature selection auftritt. Ein weiterer Bias ist der Messungs-Bias, der aus der Art und Weise, wie wir bestimmte Funktionen aussuchen, einsetzen und messen entsteht. Übertragen auf COMPAS könnte der Bias dadurch entstehen, dass Minderheiten öfter kontrolliert werden und somit verhältnismäßig auch öfter straffällig werden (Mehrabi et al. 2019). Auf der technischen Seite entsteht unter anderem der bereits genannte algorithmische Bias. Dabei ist zu beachten, dass dieser nicht durch die Inputdaten hervorgerufen wird, sondern einzig durch den Algorithmus hinzugefügt wird (Mehrabi et al. 2019). Bezogen auf COMPAS ist es somit denkbar, dass ebenfalls algorithmischer Bias zu unfairen und diskriminierenden Verurteilungen mit langfristigen Folgen führen kann. Ein unfairer Algorithmus ist demnach ein Algorithmus, dessen Entscheidungen zugunsten einer bestimmten Gruppe von Menschen verzerrt sind.

Feuerriegel et al. (2020) beschreiben weiterhin, wie sich Fairness aus verschiedenen Blickwinkeln verhält. Es wird unterschieden zwischen individueller Fairness und Gruppenfairness. Bei Letzterer bezieht sich die Fairness auf soziodemografische Attribute, die die Zugehörigkeit zu einer Gruppe rechtfertigen. Die Gruppenzugehörigkeit allein sollte keine Diskriminierung hervorrufen. Allerdings existiert keine allgemeine Definition für Gruppenfairness. Vielmehr existieren mehrere heterogene Ansätze, deren gleichzeitige Erfüllung mathematisch unmöglich ist (Feuerriegel et al. 2020). Individuelle Fairness basiert auf dem Verständnis, dass ähnlich situierte Individuen gleichbehandelt werden sollen und zielt somit auf Fairness unabhängig von einer Gruppenzugehörigkeit ab (Feuerriegel et al. 2020).

Zhou et al. (2021) beschreiben ferner Metriken wie Unbewusstheit und kontrafaktische Fairness, die der Messbarkeit von fairer KI dienen sollen. Unbewusstheit der KI gegenüber sensiblen Attributen soll dafür sorgen, dass die Applikation diese Attribute nicht in den Entscheidungsprozess miteinbezieht und somit die Fairness gesteigert wird. Mit kontrafaktischer Fairness ist gemeint, dass die gleiche Entscheidung getroffen wird, unabhängig von gewissen Attributen, wie beispielsweise dem Geschlecht des/der Betroffenen. Die starke Limitation dieser Metriken ist offensichtlich und somit in ihrer Aussagekraft vernachlässigbar (Zhou et al. 2021).

Daraus folgt, dass sowohl eine einheitliche Definition als auch aussagekräftige Metriken zum Bewerten der Fairness von KI-Anwendungen vor dem Hintergrund diverser Betrachtungsmöglichkeiten in Bezug auf Bias und Fairness zum Zeitpunkt der Verfassung der vorliegenden Arbeit in der Literatur nicht vorliegen und somit nicht beschrieben werden (Zhou et al. 2021).

Fairness von KI im Medizinischen Kontext

Jüngste Studien und Praxisbeispiele haben gezeigt, dass künstliche Intelligenzen auch im medizinischen Sektor potenziell unbeabsichtigte Folgen wie Voreingenommenheit, Diskriminierung, fehlerhafte oder unerwartete Ergebnisse hervorrufen können (Trocin et al. 2021). Aufgrund der Kritikalität der verwendeten automatisierten Anwendungen und der Sensibilität der Daten im medizinischen Bereich gilt es, die Notwendigkeit der Implementierung verantwortungsvoller KI-Praktiken im Gesundheitswesen hervorzuheben (Morley et al. 2019). Thiebes et al. (2020) identifizieren fünf TAI-Prinzipien („Beneficence", "Non-maleficence", "Autonomy", "Justice" und "Explicability"), welche befolgt werden müssen, um die Vertrauenswürdigkeit einer KI zu gewährleisten. Die Gerechtigkeit bzw. Fairness gilt hierbei besonders im medizinischen Sektor als zentraler Aspekt. Sie befasst sich mit der Etablierung ethischer Prinzipien und menschlicher Werte, die ein KI-basiertes System erfüllen muss, um Fairness zu fördern, Verzerrungen zu reduzieren und die Zuverlässigkeit der Ergebnisse gewährleisten zu können (Thiebes et al. 2020; Trocin et al. 2021).

Auch im medizinischen Kontext existiert keine einheitliche Definition einer „fairen KI". Sowohl Ahmad et al. (2020) als auch Parikh et al. (2019) weisen jedoch darauf hin, dass zur Untersuchung der Fairness im medizinischen Bereich unterschiedliche Dimensionen der zugrundeliegenden Verzerrungen berücksichtigt werden müssen. Die erste Dimension befasst sich mit den Berechnungen einer zugrundeliegenden KI.

Rechnerisch bzw. statistisch gesehen bezieht sich eine Verzerrung auf einen Analysealgorithmus, welcher ein Ergebnis liefert, das von der wahren zugrundeliegenden Schätzung abweicht (Parikh et al. 2019). Tiefergehend beschreibt algorithmischer Bias im Gesundheitswesen somit jegliche Fälle, in welchen die Anwendung eines Algorithmus bestehende Ungleichheiten in Bezug auf den sozioökonomischen Status, den ethnischen Hintergrund oder auch das Geschlecht verstärkt und sich folglich negativ auf die Ungleichheiten im Gesundheitssystem auswirkt (Panch et al. 2019). Die soziale Dimension bezieht eine strukturelle Verzerrung und eine Ungleichheit bei der Versorgung mit ein, welche aufgrund eingebetteter Praktiken, wie der allgemeinen menschlichen Voreingenommenheit, systematisch zu suboptimalen Ergebnissen für bestimmte Gruppen führen kann (Ahmad et al. 2020; Parikh et al. 2019). Als dritte relevante Dimension nennen die Autoren eine kognitive Voreingenommenheit der Ärzte und Entscheidungsträger. Diese kann als Resultat systematischer Denkfehler, etwa bei der Verarbeitung und Interpretation von Informationen aus der Umgebung, zu Ungleichheiten bei Entscheidungen und Urteilen führen (Ahmad et al. 2020).

Weiterhin muss in Bezug auf die Entwicklung fairer medizinischer KI sowohl die Abhängigkeit der Stakeholder als auch die Abgrenzung der individuellen Fairness von der Gruppenfairness beachtet werden. Während Ärzte sich eher darum sorgen, wie viele der gekennzeichneten Patienten auch wirklich gefährdet sind und die Gesamtheit aller betrachten, erwarten Patienten auf persönlicher Ebene individuell faire Ergebnisse. Für die Gesellschaft wiederum spielt es eine Rolle, ob für alle Patienten die gleichen Voraussetzungen gelten, alle gleich behandelt werden und eine gesellschaftliche Gerechtigkeit über alle Gruppen hinweg zu verzeichnen ist (Ahmad et al. 2020).

Grundlegenderweise haben künstliche Intelligenzen das Potenzial, Ungleichheiten aufzudecken und gesundheitliche Chancengleichheit vorantreiben zu können, sodass allen Menschen eine gleich gute medizinische Versorgung zukommt (Baumgartner 2021). Im Gegensatz zu Ärzten, welche fallspezifisch unterscheiden und entsprechend handeln können, optimiert eine klassische KI Vorhersagen für Mehrheitsgruppen und kann dementsprechend nicht fallspezifisch unterscheiden, wodurch es zu Ungleichheiten kommen kann (Ahmad et al. 2020). Die Entwicklung und Verwendung von KI im Gesundheitswesen stellt demnach eine sehr komplexe Aufgabe dar und bedarf einer genauen Analyse der Gründe des Auftretens der Verzerrungen, um allen Formen der Fairness gerecht werden zu können (Baumgartner 2021), zumal sich Fairness im Gesundheitswesen in vielerlei Hinsicht von anderen Domänen unterscheidet. Hier muss angewandte KI neben algorithmischer Voreingenommenheit auch mit Voreingenommenheit umgehen, welche sich bei der Bereitstellung von Gesundheitsleistungen, beispiels-weise in der Datenerhebung oder Programmierung, einschleichen kann (Ahmad et al. 2020).

Methodische Vorgehensweise

Eine systematische Literaturübersicht, die auch als systematisches Literaturreview bezeichnet wird, gilt nach Kitchenham und Charters (2007) als bewährtes Mittel zur Identifizierung, Bewertung und Interpretation aller verfügbaren Forschungsergebnisse zu einem bestimmten Thema. Insbesondere nützlich ist die Erstellung einer Literaturübersicht zur angemessenen Präsentation neuer Forschungs-aktivitäten, um basierend darauf mögliche Lücken und Bereiche für weitere Forschungen vorschlagen zu können (Kitchenham & Chartes 2007).

Datensammlung

Um einen umfassenden Überblick über den Stand der bisherigen Forschung zur Fairness medizinischer KI zu erhalten, galt es im ersten Schritt des Literaturreviews geeignete vorhandene Literatur zur Thematik zu identifizieren. Dazu gehörte zunächst die Auswahl wissenschaftlicher Datenbanken sowie die Festlegung eines geeigneten Suchterms (Kitchenham & Charters 2007). Zur Abdeckung eines breiten Spektrums an Publikationen über mehrere Quellen hinweg wurden sowohl Informatik-spezifische als auch interdisziplinäre Datenbanken gewählt. Neben *ACM Digital Library*, *AIS Electronic Library*, *IEEE Explore* als Informatik-spezifische, dienten die Datenbanken für peer-reviewte Artikel *Scopus*, *Ebsco Buisness Source* und *Pubmed* der Abdeckung des interdisziplinären als auch medizinischen Kontextes. Zusätzlich wurden aufgrund des jungen und dynamischen Forschungsfeldes der künstlichen Intelligenz im medizinischen Sektor die Preprint-Datenbanken *ArXiv* und *MedRXiv* hinzugezogen. Im Vergleich zu den vorherigen

Quellen beinhalten diese neue, potenziell relevante Forschungsergebnisse, die zwar noch nicht final veröffentlicht wurden, aber für die vorhandene Untersuchung dennoch sehr relevant sein könnten.

Zur Bestimmung eines geeigneten Suchbegriffes wurden die grundlegenden Komponenten der übergeordneten Forschungsfrage identifiziert. Neben einem KI-spezifischen Begriff aus den Kerndomänen ("Artificial Intelligence", "Machine Learning" und "Deep Learning") sollten die Veröffentlichungen einen weiteren Fairness-bezogenen Ausdruck im Titel, im Abstract oder in den Schlüsselwörtern enthalten. Für die Fairness umfasste der Suchterm die Begriffe („fair/fairness", „bias", oder „discrimination/discriminating"). Um den Suchbereich weiter einzuschränken, wurden die Publikationen auf den medizinischen Bereich („diagnose/diagnosis", „digital health" oder „health care") limitiert. Um die Abdeckung aller potenziell relevanter Suchergebnisse zu gewährleisten, wurde mit Hilfe der booleschen Operatoren *AND* und *OR* und unter Berücksichtigung verschiedener Schreibweisen, Abkürzungen und Synonyme folgender Suchterm festgelegt: *("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning") AND (fair* OR bias OR discrimination OR discriminating) AND ("digital health" OR "health care" OR diagnos*).*

Im Anschluss wurden die ausgewählten wissenschaftlichen Datenbanken systematisch durchsucht. Die tatsächliche Suche wurde am 12.05.2021 getätigt und ergab ohne Filterung 12.384 Artikel. Abbildung 1 zeigt die gefundene Anzahl an Veröffentlichungen zwischen dem Jahr 2000 und 2021. Wie der Abbildung zu entnehmen ist, fand ab dem Jahr 2017 ein nahezu exponentieller Anstieg in der Anzahl der Veröffentlichungen statt. Mit insgesamt 1.631 Publikationen wurden im Jahr 2020 bislang die meisten Artikel zum Thema veröffentlicht.

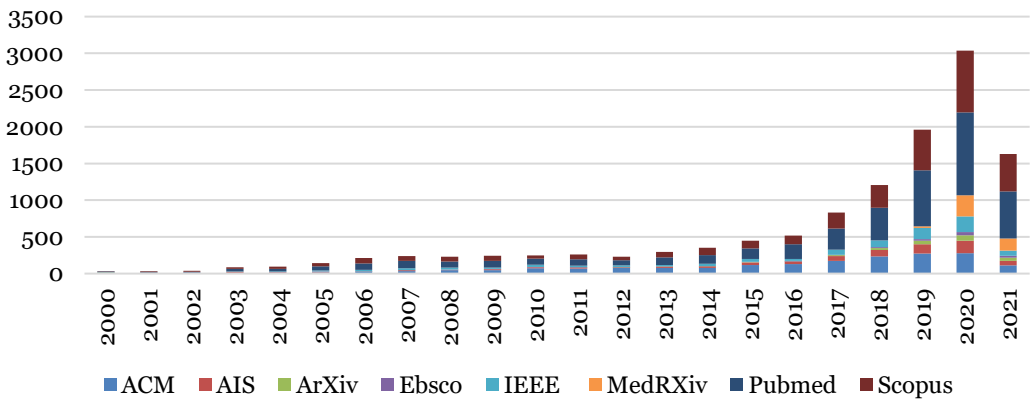

Abbildung 1. Anzahl an Veröffentlichungen pro Jahr und Datenbank

Um relevante Artikel für die weitere Untersuchung zu extrahieren, galt es im nächsten Schritt geeignete Ein- und Ausschlusskriterien zur Literaturauswahl zu definieren (Kitchenham & Charters 2007). Zunächst wurden Artikel ausgeschlossen, die vor 2016 veröffentlicht wurden. Dieser Schritt ist mit der Neuheit des Forschungsfeldes und der wie in Abbildung 1 ersichtlichen seit 2016 beziehungsweise 2017 exponentiell angestiegenen Anzahl an Publikationen der Fairness von KI im medizinischen Sektor zu begründen. Die gefundene Datenbasis reduzierte sich durch den Jahresfilter um 4.131 Artikel auf insgesamt 8.253 Publikationen, die über die verschiedenen Datenbanken hinweg identifiziert werden konnten (siehe Tabelle 1).

Nachdem die Suchergebnisse in das Datenverwaltungsprogramm Citavi exportiert wurden, konnten alle Duplikate ausgeschlossen werden. Die resultierenden 7.741 Artikel wurden im Anschluss in eine Excel-Datei exportiert, um weitere Filterungen vornehmen zu können. Zuallererst wurde geprüft, ob tatsächlich alle Duplikate entfernt wurden und sich in dem Datensatz nur Artikel ab dem Veröffentlichungsjahr 2016 befanden. Hierbei konnten 446 Treffer aufgrund von Dopplungen sowie ein Treffer, der vor 2016 veröffentlicht wurde, identifiziert und entfernt werden. Weiterhin wurden aus dem Datensatz 707 Publikationen entfernt, bei welchen es sich nicht um Preprints oder peer-reviewte Artikel handelte. Bei den meisten dieser Artikel handelte es sich entweder um Tagungsbände, Titel von Konferenzprotokollen oder Bücher. Im Anschluss wurden weitere 38 nicht englisch-sprachige Publikationen entfernt, womit sich die resultierende Anzahl an Artikeln auf 6.550 belief.

Datenbank	**Anzahl der Ergebnisse (ab 2016)**
ACM Digital Library	1.189
AIS Electronic Library	518
ArXiv	205
Ebsco Business Source	99
IEEE Explore	618
MedRxiv	437
PubMed	2.811
Scopus	2.376
Gesamt	8.253

Tabelle 1. Anzahl an Ergebnissen je Datenbank (ab 2016)

Die verbleibenden Artikel wurden im Anschluss auf Basis ihres Titels und Abstracts gescannt, um relevante Publikationen zur Beantwortung der vorliegenden Forschungsfragen zu finden. Generell konnte festgestellt werden, dass es sich bei einer Vielzahl an Beiträgen um Kommentare oder erweiterte Zusammenfassungen handelte. Auch stellte sich heraus, dass sich ein Großteil der Artikel mit generellen und sehr speziellen Anwendungen der KI im medizinischen Sektor befasste und ethische Probleme wie die Fairness der KI weniger in den Vordergrund stellten. Insgesamt 178 Artikel konnten schließlich zur weiteren Analyse identifiziert werden. Auf Basis ihres Volltextes wurde folglich geprüft, ob die Artikel sich tatsächlich mit dem Anwendungsfeld und der Schnittstelle der Fairness der medizinischen KI befassten. Weitere 136 Artikel wurden schließlich als potenziell relevant eingestuft, aber dennoch aus der weiteren Analyse ausgeschlossen, da sie die beiden Konzepte nicht repräsentativ widerspiegelten. Die finale Auswahl belief sich somit auf ein Datenset von 35 Artikeln, die für die weitere inhaltliche Analyse herangezogen wurden.

Datenanalyse

Anschließend an die Datensammlung, welche resultierend 35 relevante Artikel identifizierte, wurden alle Artikel in Gruppen kategorisiert. In Anlehnung an Ahmad et al. (2020) wurde die Einteilung der Gruppen festgelegt und soweit nötig adaptiert. Um der vorhandenen Literatur besser gerecht zu werden, erfolgte eine weitere Unterteilung der Bereiche des Daten- und Systemproblems und eine Neubenennung einiger Kategorien (z.B. Schnittstelle Mensch - KI statt Nutzer und die Unterteilung des Datenproblems in Datenerhebung, Datenqualität und Menge).

Tabelle 2 gibt einen Überblick über die angepasste Gruppierung sowie die Anzahl der Artikel in jeder Kategorie. Dabei ist zu beachten, dass einige Artikel in mehrere Gruppen kategorisiert wurden und somit die Summe der Artikel in Tabelle 2 die Anzahl der 35 relevanten Paper übersteigt. Tabelle 3 im Anhang liefert eine detaillierte Übersicht über die Einteilung der 35 finalen Artikel in die einzelnen Kategorien.

Perspektive	**Kategorie/ Gruppe**	**Anzahl Artikel**
Datenproblem	Erhebung	7
	Qualität	5
	Menge	4
Systemproblem	Algorithmus	8
	Modell	14
	Schnittstellen Mensch - KI	13

Tabelle 2. Systematische Gruppierung der Identifizierten Literatur

Überblick über den Stand der Forschung

Im Folgenden wird ein systematischer Überblick über den Stand der Forschung und die bestehende Literatur zum Thema Fairness medizinischer KI gegeben. Die identifizierte Literatur lässt sich hierbei in die übergeordnete Abgrenzung eines Daten- oder Systemproblems einordnen, auf welche in diesem Kapitel näher eingegangen wird.

Fairness Medizinischer KI als Datenproblem

Eine Vielzahl an Artikeln, die sich mit der Fairness der maschinellen Systeme im medizinischen Sektor befassen, sehen das Problem der verzerrten Ergebnisse in Verbindung mit den zugrundeliegenden Eingabeparametern auf Basis derer die künstlichen Intelligenzen ihre Entscheidungen treffen. Brault und Saxena (2020) weisen darauf hin, dass Verzerrungen in den Datensätzen sowohl bei der Erhebung und Verarbeitung der Daten entstehen können, insbesondere aber auch die Menge der Daten und Qualität berücksichtigt werden muss.

Datenerhebung

Hee (2017) unterscheidet medizinische Daten in zwei Arten: Forschungsdaten, die aus randomisierten kontrollierten Studien erhoben werden und klinische Daten, die aus historischen elektronischen Gesundheits- und Patientenakten abgeleitet werden (Electronic Health Records, Electronic Health Data). Trotz zahlreicher Vorteile der (Wieder-) Verwendung historischer Daten aufgrund der einfachen und kostengünstigen Sammlung (Hee 2017) kritisierten aktuelle Debatten den Einsatz historischer Daten in verschiedenen Anwendungsbereichen. Da die Daten nicht speziell für die klinische Evidenz gesammelt werden, besteht die Gefahr, dass die zugrundeliegende Population aufgrund von fehlenden oder falschen Daten nicht repräsentativ widergespiegelt wird (Hee 2017; Lee et al. 2021).

Neben verwendeten klinischen Daten liegt ein weiterer Untersuchungsfokus auf alternativen Methoden der Datenerfassung wie der Sammlung von Daten aus sensorbasierten Überwachungssystemen (Lee et al. 2021; Ienca & Ignatiadis 2020). Brault und Saxena (2020) widmen sich der Sammlung und Validität von mobilen Gesundheitsdaten und zeigen, dass die aus den Daten gezogenen Schlüsse verzerrt sein können. Angesichts des exponentiellen Anstiegs an vorhandenen elektronischen Gesundheitsakten, (Hirn-) Bildgebungsdaten, und Sätzen aus unkonventionellen Datenquellen geben sowohl Lee et al. (2021) als auch Ienca und Ignatiadis (2020) einen Überblick über die KI-Ansätze in der Neurowissenschaft. Trotz vielseitigem Potenzial äußern sie ihre Bedenken hinsichtlich des Einsatzes der Techniken aufgrund mangelnder Repräsentativität der erhobenen Daten, Voreingenommenheit, der Privatsphäre, Transparenz und anderen ethischen Aspekten (Ienca & Ignatiadis 2020; Lee et al. 2021;).

Larrazabal et al. (2020) und Seyyed-Kalantari et al. (2020) widmen sich Verzerrungen in existierenden öffentlichen medizinischen Bildgebungsdatensätzen zur computergestützten Diagnose sowie dem bildbasierten Screening von Thorax Erkrankungen. Seyyed-Kalantari et al. (2020) analysieren und finden aussagekräftige Muster von Verzerrungen in Form von Disparitäten in True-Positive-Raten zwischen verschiedenen geschützten Attributen wie Geschlecht, Alter, Rasse und Versicherungsart der Patienten in drei öffentlichen Röntgendatensätzen. Ahsen et al. (2019) diskutieren die Auswirkungen von Menschen erzeugter und verzerrter Eingabedaten auf Klassifizierungsalgorithmen, in dem sie die Quelle sowie Mechanik des Rauschens explizit modellieren und Einblicke geben, wie die Verzerrung der Eingabedaten das optimale Design des Algorithmus und dessen Leistung beeinflusst.

Datenmenge

Um zu lernen, wie man bestimmte Aufgaben ausführt, benötigen Machine Learning und Deep Learning Algorithmen eine große Menge und Vielfalt an Trainingsdaten, um Vorhersagen treffen zu können, die von der Vorhersage einer optimalen Kapazitätsauslastung in Krankenhäusern bis zur Prädiktion von bösartigen Läsionen reichen können (Martín Noguerol et al. 2019). Rückgreifend auf die Erhebung der Daten anhand experimenteller Studien haben Forscher es nicht einfach, große, vielfältige Datensätze zu erhalten, da die Studien oftmals zu wenige aussagekräftige Daten liefern (Brault & Saxena 2020). Umgekehrt besteht bei der Analyse großer Datensätze die Gefahr, dass die Daten übermächtig und unstrukturiert (multimedial, grafisch, textuell) sind (Brault & Saxena 2020). Brault und Saxena (2020) führen dies anhand der

Sammlung von mobilen Gesundheitsdaten (Mobile Health Data) auf und argumentieren, dass die Sensoren zur Erhebung der Daten je nach Gerätetyp und Anwendungsbedingungen variieren, was zu zufälligen oder auch systematischen Fehlern in den Daten führen kann.

In der momentanen Debatte ist die Datenmenge und -vielfalt von zentraler Bedeutung, um die Vorurteile der medizinischen KI zu verstehen und abzuschwächen, weshalb sich Forscher zunehmend mit Methoden beschäftigen, große Mengen an Daten aus kleinen Sätzen zu erhalten. Martín Noguerol et al. (2020) weisen in ihrer Analyse auf algorithmische Datenerweiterungsmethoden wie Transfer Learning Algorithmen und Generative Adverse Netzwerke hin. Burlina et al. (2021) demonstrieren einen Ansatz zur Vergrößerung der Trainingsdaten durch die kontrollierte Erzeugung weiterer synthetischer Stichproben, um den Bedarf an mehr Populationen zu decken, die möglicherweise unterrepräsentiert sind oder fehlerhaften Faktoren entsprechen. Anhand ihrer generativen Methodik können sie eine Vorhersagegenauigkeit der diabetischen Retinopathie für die zuvor unterrepräsentierte dunkelhäutige Subpopulation von 71,5 % (zuvor 60,5 %) im Vergleich zu der hellhäutigen Subpopulation mit 72,0 % (zuvor 73,0 %) erreichen (Burlina et al. 2021).

Datenqualität

Die bloße Bereitstellung vielfältiger und großer Mengen an Trainingsdaten ist nach Brault und Saxena (2020) allerdings keine Garantie zur Beseitigung von Vorurteilen. Es bedarf vielmehr einer kritischen Würdigung der Daten und deren Analyse (Brault & Saxena 2020). Die Sicherstellung der Datenqualität gilt als zentraler Faktor und als Muss bei der Wiederverwendung von Forschungsdaten als auch klinischen Daten, um der Gefahr des Treffens falsch negativer Fehler vorzubeugen (Hee 2017; Martín Noguerol et al. 2019). In der vorhandenen Literatur lassen sich im Umgang mit der Datenqualität primär verschiedene vorgeschlagene Lösungsansätze abgrenzen. Martín Noguerol et al. (2019) betrachten den Einsatz von KI und ML im Anwendungsfeld der Radiologie auf strategischer Ebene und weisen anhand einer SWOT-Analyse darauf hin, dass der Erfolg eines ML Algorithmus primär auf der Qualität der Beispiele, die für das Training verwendet werden, beruhen. Die Erzeugung qualitativ hochwertiger und umfangreicher Trainingsdaten erfordert allerdings Experten, die Ground-Truth Beispiele beschriften und Kohorten von spezifischen Patienten, medizinischen Bildern oder Bildgebungsberichten kuratieren (Martín Noguerol et al. 2019). Hee (2017) schlägt einen Prozess der Datenqualitätssicherung vor, der aus mehreren rekursiven Sicherheitsschritten besteht und die Dimensionen der Vollständigkeit, Richtigkeit, Konsistenz und Aktualität der Datenqualitätssicherung einbezieht. Bissoto et al. (2019) diskutieren Messprobleme (Bias) im Zusammenhang mit Labels und Prädiktoren und führen ein kontrafaktisches Experiment durch, indem nach und nach sinnvolle Informationen in den Daten zerstört werden und die Leistung der Modelle zur automatisierten Analyse von Hautläsionen gemessen wird. Die Verzerrungen in den Datensätzen zerstören die Robustheit von Modellen, da Modelle falsche Korrelationen lernen, die nicht in der realen Welt gefunden werden und somit die Leistung der Modelle aufblähen oder herabsetzen (Bissoto et al. 2019). Basierend auf ihren Ergebnissen konkretisieren die Autoren in einer Anschlussstudie, welche falschen Korrelationen von voreingenommenen Netzwerken ausgenutzt werden und schlagen eine Technik vor, anhand derer Modelle entschärft werden können und falsche Korrelationen entfernt werden können (Bissoto et al. 2020).

Fairness Medizinischer KI als Systemproblem

Neben der Reihe an Artikeln, die das Problem der Fairness medizinischer KI als Datenproblem sehen, beschäftigt sich ein weiterer Teil der Auswahl an Artikeln mit der Fairness als Systemproblem. Hierbei wird der Fokus der Artikel auf den Algorithmus, das Modell als auch auf die Interaktion zwischen Mensch und Maschine gelegt.

Algorithmus

Beim Einsatz von KI in der Medizin zeigt sich, dass die eingesetzten Algorithmen für bestimmte Personengruppen einen Bias aufweisen können, welcher zur Diskriminierung dieser führt. Selbst ein Algorithmus, der auf einem relativ ausgewogenen Datensatz in Bezug auf Alter und Geschlecht trainiert wurde, zeigt verzerrte Ergebnisse (Abbasi-Sureshjani et al. 2020). Aus diesem Grund bildet ein Algorithmus eine weitere Ursache für die Entstehung von Bias.

Zu Beginn wird kurz auf die Stärken des Einsatzes von KI-Algorithmen eingegangen, daran anschließend liegt der Fokus auf den Schwächen und Problemen, die durch den Einsatz von KI-Algorithmen entstehen können. Die Stärken und Möglichkeiten eines KI-Algorithmus zeigen Martín Noguerol et al. (2019) in ihrem Artikel auf. Diese ergeben sich durch das automatische Erkennen, Segmentieren und Klassifizieren von Läsionen. Die effizientere und präzisere Diagnose durch den Algorithmus erleichtert die Arbeit des Radiologen und reduziert die menschliche Intervention in zeitaufwändigen, konventionellen Bildgebungsaufgaben (Martín Noguerol et al. 2019). Als Schwäche nennen die Autoren die Übertragbarkeit eines Algorithmus auf andere Aufgaben. Gute Ergebnisse eines ML-Algorithmus in einer Aufgabe lassen nicht schlussfolgern, dass dies auch bei anderen Aufgaben der Fall ist. Um allgemein reliable Ergebnisse zu erhalten, werden verschiedene Algorithmenstrukturen benötigt, die wiederum unterschiedliche Datenattribute brauchen, um richtig zu funktionieren (Ienca & Ignatiadis 2020). Der Mangel an kontextueller Spezifität eines Algorithmus stellt eine Herausforderung im Gesundheitssystem dar, um dem algorithmischen Bias gegenüberzutreten (Panch et al. 2019).

Neben der Spezifität muss außerdem das Auswertungsziel des Algorithmus mit den damit einhergehenden Prozessschritten klar definiert sein und später überprüft werden. Die Autoren Obermeyer et al. (2020) befassen sich in ihrem Artikel näher mit dem Innenleben eines Algorithmus und quantifizieren rassische Unterschiede eines Algorithmus, indem gegebene Inputs, Outputs und mögliche Outputs verglichen werden. Ihre Untersuchungen kommen, wie bereits zu Beginn der Arbeit einleitend erwähnt, zu dem Ergebnis, dass ein Algorithmus die Gesundheitskosten der Krankenhauspatienten vorhersagt anstatt der eigentlichen Krankheit, weswegen dunkelhäutige Patienten seltener als hellhäutige zusätzliche medizinische Hilfe bekamen, obwohl ihre Krankheit schwerwiegender war. Ein ungleicher Zugang zur Gesundheitsversorgung heißt, dass weniger Geld für Dunkelhäutige als für Hellhäutige ausgegeben wird (Obermeyer et al. 2020). An dieser Stelle wird deutlich, dass Algorithmen den menschlichen Bias weitergeben und verstärken. Klassifizierungsalgorithmen, welche von Menschen erzeugte Eingabedaten verwenden, die bereits mit menschlichem Bias versehen sind, können in ihren erzeugten Vorhersagen, die aus solchen Verzerrungen resultierenden Fehler zusätzlich verstärken. Somit verstärkt der algorithmische Bias Ungerechtigkeiten in der Gesellschaft (Panch et al. 2019). Diesem Problem kann durch die Gestaltung der Algorithmen entgegengewirkt werden. Bias-bewusste Algorithmen bieten eine Möglichkeit diesen Bias zu eliminieren (Ahsen et al. 2019). In ihrem Paper untersuchen letztaufgeführte Autoren den optimalen Entwurf eines linearen Klassifizierungsalgorithmus, wenn eine Verzerrung durch den Radiologen mit den einhergehenden Auswirkungen auf die Leistung und das Design des Algorithmus vorhanden ist. Sie kommen ebenso zu dem Ergebnis, dass ein bias-bewusster Algorithmus das Problem beheben kann. Dieser kann die negativen Auswirkungen von Bias eliminieren, wenn der Fehler in der Mammographiebeurteilung aufgrund des Bias des Radiologen keine Varianz aufzeigt (Ahsen et al. 2019). Besonders die akkurate Schätzung des Biasparameters ist essenziell für das optimale Algorithmusdesign (Ahsen et al. 2019).

Außerdem ist ausreichendes Testen und Validieren für die Sicherheit im klinischen Einsatz von hoher Bedeutung. Jedoch kann die Überoptimierung eines Algorithmus während der Trainingsphase mit dem eigentlichen Ziel der Erlangung höherer Genauigkeit in das Gegenteil überlaufen und einen ungewollten Bias hervorrufen (Ienca & Ignatiadis 2020). Den exakten Prozessschritt des Algorithmus zu finden, welcher den Bias hervorruft, ist durch die fehlende Transparenz und Erklärbarkeit nicht gegeben. Dabei sollten Datenwissenschaftler und Ärzte die Art und Weise der Lösungsfindung nachvollziehen (Panch et al. 2019). Das Problem hierbei entsteht durch die Black-Boxen des Algorithmus (Ienca & Ignatiadis 2020; Martín Noguerol et al. 2019; Panch et al. 2019). Interne Arbeitsschritte der KI-Algorithmen entziehen sich kausalen Erklärungen in logischer oder mechanistischer Form (Ienca & Ignatiadis 2020) und infolgedessen verhindern Black-Boxen die optimale Nachverfolgung der Schritte, die den spezifischen Prozess Input-Algorithmus-Output erklären (Martín Noguerol et al. 2019).

Modell

Die Artikel dieses Abschnitts befassen sich mit dem für die Entwicklung leistungsstarker und fairer KI-Anwendungen notwendigen Trade-Off zwischen Fairness, Modellleistung und Modellerklärbarkeit.

Der Status Quo des Einsatzes medizinischer KI weist, primär in Abhängigkeit demografischer und ethnischer Aspekte, starke Verzerrungen in den Ergebnissen auf. Dies zeigt unter anderem eine Studie von Chen et al. (2019), welche die Unterschiede in der Vorhersagegenauigkeit in Bezug auf das Geschlecht und den Versicherungstyp untersucht. Hier weisen weibliche Patienten eine höhere Modellfehlerrate als

männliche, gesetzlich Versicherte eine höhere als privat Versicherte und, im Vergleich der ethnischen Herkunft, schwarze Pateinten die höchste Modellfehlerrate auf (Chen et al. 2019). Wie Obermeyer et al. (2019) beobachten auch Sun et al. (2020) demographisch bedingte Unterschiede einer Modellleistung, welche auch bei einer Verbesserung des Recalls des Modells bestehen bleiben. Obwohl KI aktuelle Analysemodelle verbessern kann, tut sie dies derzeit unter Bedingungen der epistemischen und methodischen Unsicherheit. Im weiteren erkenntnistheoretischen Sinne stellt diese hypothesenfreie Natur der meisten KI-gesteuerten Mustererkennungsmodelle den Begriff der Kausalität in Frage und birgt das Risiko, ungerechtfertigte kausale Beziehungen aus bloß korrelativen Daten abzuleiten. So können kausal schwache KI-Modelle unter bestimmten Umständen die Unsicherheit erhöhen, statt diese zu verringern (Ienca & Ignatiadis 2020).

Trade-Off Leistung vs. Fairness. Standardisierte ML-Modelle optimieren Prognosen für Mehrheitsgruppen, was zu erheblichen Fehlerraten für die Vorhersageergebnisse von Minderheiten führt (Afrose et al. 2021). Im Allgemeinen wirkt sich Fairness negativ auf die Leistung einer KI aus, da sie das Ziel der Genauigkeit sowohl auf Genauigkeit als auch auf Fairness umlenkt. Daher ist ein Trade-Off zwischen der Leistung des Modells und dessen Fairness erforderlich (Ahmad et al. 2020). Chester et al. (2020) analysieren den Verlust der Leistung eines Modells in Abhängigkeit der Trainingsdaten bei dem Versuch, dieses fairer zu gestalten. Durch De-Identifizierung der Inputdaten der Anwendung soll der Einfluss sensibler demografischer Attribute, wie der Rasse oder dem Geschlecht, auf die Entscheidung einer KI begrenzt werden und somit Fairness schaffen (Chester et al. 2020). Jedoch führt dies zu Datenverlust und beeinflusst so Modelle, welche mit de-identifizierten Daten trainiert wurden (Chester et al. 2020). Es kommt also zu einem Verlust der Genauigkeit des Modells, wenn es fairer gestaltet wird. Je nachdem, ob die Gruppen- oder individuelle Fairness, Chancengleichheit oder Ergebnisgleichheit optimiert werden soll, zeigt der Genauigkeitsverlust des Modells unterschiedliche Ausmaße (Chester et al. 2020). Auch Bissoto et al. (2020) beobachten in ihrer Studie eine Verschlechterung der Modellleistung bei einer faireren Gestaltung der KI. Gegensätzlich zu der oft beobachteten negativen Korrelation von Fairness und Modellleistung finden Weng et al. (2020), dass eine Kombination der Lernmethoden FSL und CLS die Modellleistung, besonders für seltene Klassen, erheblich verbessert.

Trade-Off Erklärbarkeit vs. Fairness. KI im Gesundheitswesen benötigt neben einer guten Leistung und Fairness auch Reliabilität und Sicherheit (Gade et al. 2020). Dies kann nach Gade et al. (2020) nur gewährleistet werden, wenn das Bedürfnis nach Erklärbarkeit und Transparenz erfüllt ist. Dieses hängt vom Design der KI und deren algorithmischen Methodik ab (Gade et al. 2020). Erklärbare und transparente KIs erlauben es dem Anwender, der KI Feed-back zu geben und Fehler zu korrigieren, wodurch Modellvorhersagen durch menschliche Erfahrungen optimiert werden können (Gade et al. 2020). Gegensätzlich hierzu finden Liu et al. (2018), dass Fairness-Kriterien in einem einstufigen Feedback-Modell langfristig zu keiner Verbesserung führen, sondern in manchen Fällen Schaden anrichten, in welchen ein uneingeschränktes Modell keinen anrichten würde. Weiterhin ist die Interpretierbarkeit einer Modellentscheidung oder -vorhersage entscheidend für die Akzeptanz medizinischer KI-Anwendungen und ihren Modellvorhersagen, welche Ärzten in schwierigen Fällen als Hilfe heranziehen können (Bissoto et al. 2020). Generell zeigen komplexe Modelle, wie tiefe neuronale Netze, eine höhere Genauigkeit als einfachere, interpretierbare Modelle (Jabbari et al. 2020). In einer Untersuchung der statistischen Parität und Chancengleichheit in Abhängigkeit der Anzahl der Merkmale, welche einem Klassifikator zur Verfügung stehen, beobachten Jabbari et al. (2020) die komplexe Beziehung zwischen Interpretierbarkeit eines Modells und dessen Fairness, welche vier verschiedenen beobachtbaren Trends folgt. Dabei verhält sich der Grad der Fairness-Verletzung mit Zunahme der Modellkomplexität, je nach Korrelation der Merkmale, unterschiedlich (Jabbari et al. 2020).

So wie die Modellerklärbarkeit und dessen Leistung oft in einem reziproken Verhältnis zueinanderstehen, lässt sich ein ähnlicher Zusammenhang für Fairness beobachten. Die Erklärbarkeit von ML-Modellen soll eine stärkere Überprüfung dieser und damit die Möglichkeit fairer und gerechter Modelle bewirken. Allerdings kann eine solche Vereinfachung auch zu einer Leistungsverschlechterung sowie zu weniger fairen Modellen führen. Der Kompromiss ist also dreifach: Fairness vs. Performance vs. Erklärbarkeit (Ahmad et al. 2019). Afrose et al. (2021) sowie Sun et al. (2020) schlagen zur Überwindung dieses Trade-Offs das Trainieren von subpopulationsspezifischen Modellen vor, welche individuell für den jeweiligen Anwendungsort auf die vorherrschenden Anforderungen trainiert und angepasst werden können.

Schnittstellen Mensch-KI

Menschliche Individuen entwickeln für den Prozess der Entscheidungsfindung bewusst und unbewusst heuristische Verfahren. Im Sinne der Effizienz ist dies auch durchaus wünschenswert. Zusammen mit individuellen Vorurteilen, welche unter anderem aus der emotionalen Veranlagung resultieren, entstehen dadurch allerdings kognitive Verzerrung (Kellmeyer 2019; Lee et al. 2021; Starke et al. 2021). Aus diesem Grund wird in folgendem Kapitel detailliert darauf eingegangen, inwiefern die Interaktion von menschlichen Individuen an Schnittstellen mit medizinischer künstlicher Intelligenz Bias und Ungleichheit manifestieren. Dazu werden die Schnittstellen der Nutzer und die der Entwickler mit der KI beleuchtet (Kellmeyer 2019; Lee et al. 2021; Starke et al. 2021).

Schnittstellen zur KI bei der Entwicklung. Es ist unabdingbar, dass Experten aus den verschiedenen Domänen in den Entwicklungsprozess von medizinischer KI involviert sind. Dazu gehören Data Scientists, Programmierer, Mediziner und so weiter. Trotz aggregierter Expertise tritt hier Bias auf, was an unterschiedlichen, aber gleichzeitig auch zu eindeutigen Motiven, sowie den individuellen Verzerrungen der Mitwirkenden und der Projekttreibenden liegt. Damit ist gemeint, dass zwar der Konsens ein faires medizinisches Produkt zu entwickeln besteht, die Experten innerhalb ihrer Domäne allerdings eigene Ansätze verfolgen. Gleichzeitig verfolgt und delegiert ein Projektverantwortlicher ein eindeutiges Ziel, welches sich aus seinen intrinsischen und extrinsischen Anreizen zusammensetzt. Aus unternehmerischer Sicht liegt hier meist ein wirtschaftliches Motiv vor, das heißt es werden gewinnmaximierende Strategien verfolgt während Aspekte wie Ethik und Fairness geringer priorisiert werden (Noor 2020).

Aus medizinischer Sicht entsteht Bias unter anderem durch die individuelle Erfahrung und der ihr einhergehende Verzerrung der mitwirkenden Ärzte. Ist beispielsweise ein approbierter Arzt involviert, der Hautkrankheiten zu einem Großteil bei hellhäutigen Menschen untersucht, wird seine Verzerrung gegenüber anderen Personengruppen in die KI miteinfließen (Kellmeyer 2019; Straw 2020; Lee et al. 2021; Starke et al. 2021; Geis et al. 2019). Durch fehlende Diversität im Entwicklerteam, kann dieser individuelle Bias nicht neutralisiert werden. Auch die Diversität auf der interdisziplinären Makroebene des Entwicklerteams ist meist nicht gegeben, da die damit anfallenden Kosten enorm wären und für die meisten Unternehmen zu hohe Risiken bergen. Dadurch dominieren individuelle Motive, wie beispielsweise ein gewinnmaximierendes Motiv gegenüber der Fairness. Die Unfairness spiegelt sich somit in einem generellen Interessenskonflikt zwischen wirtschaftlich orientierten Unternehmen und der auf Fairness orientierten Gesellschaft wider (Miguel et al. 2020; McCradden et al. 2020). Aus diesem Grund muss der Transparenz an dieser Stelle eine wichtige Rolle zugeschrieben werden. Für jeden Nutzer sollte es ersichtlich sein, welches Produkt auf welche Zielgruppe hin entwickelt wurde (Kellmeyer 2019; Straw 2020).

Schnittstelle der Nutzer mit der KI. Auf der Seite der Nutzer sind Ärzte beziehungsweise generell medizinisches Personal, Krankenkassen und die Patienten anzusiedeln.

Bei der Anwendung der KI gibt es einige Schnittstellen, an denen Bias und Diskriminierung auftreten kann. Dabei ist Wohlstand ein sehr zentraler Faktor. Allerdings ist das kein Problem, dass sich nur auf KI-Produkte zurückführen lässt. Mit freien, kapitalistischen Marktwirtschaften einhergehend spielt die individuelle Verfügung über finanzielle Mittel seit jeher eine ausschlaggebende Rolle bei der medizinischen Versorgung. Als Musterbeispiel dient hier die USA. Beginnend mit der Qualität der Ausbildung von medizinischem Personal über die technische Ausstattung von Arztpraxen und Krankenhäusern bis zu umfangreichen Privatversicherungen, gibt es unzählige Einflussfaktoren auf die Qualität der individuellen medizinischen Behandlung (Hague 2019). Es ist daher naheliegend, dass große Investitionen in innovative, technische Produkte ohnehin nur von wirtschaftlich starken Versorgern, für deren wirtschaftlich starken Patienten getätigt werden können. Inwiefern dieser Sachverhalt un-/fair ist, wird an dieser Stelle nicht weiter ausgeführt. Betrachtet man jedoch die Behandlung innerhalb dieser Personengruppen bleibt das Problem von Diskriminierung, welches in vorangegangenen Kapiteln erläutert wurden, dennoch bestehen (Hague 2019).

Großes Konfliktpotential besteht an der Schnittstelle zwischen Nutzern und KI durch das generelle Vertrauensproblem der breiten Gesellschaft gegenüber künstlicher Intelligenz. Dies ist durch fehlendes technisches Wissen, durch die Black-Box Eigenschaften von künstlicher Intelligenz aber auch dem Stigma an geringem Datenschutz zu begründen. Die Fachliteratur belegt, dass das Vertrauen der Patienten rund um die Erhebung und Verwendung der Daten gegeben sein muss, um den effektiven und ethischen Einsatz von KI sicherzustellen (Murphy et al. 2021; Katznelson & Gerke 2021). Der Datenaustausch zwischen

Google DeepMind und Royal Free London NHS Foundation Trust resultierte in einem solchen gesellschaftlichen Misstrauen. Dabei wurden persönliche, medizinische Daten von 1,6 Millionen Patienten ausgetauscht, um eine App zu entwickeln, die die Behandlung von Nierenschäden verbessern sollte. Es kam die Debatte auf, ob die Daten zu leichtfertig und unnötigerweise in zu großem Umfang weitergegeben wurden. Darüber hinaus wurde hinterfragt, ob die betroffenen Patienten hinreichend über die Verwendung ihrer Daten aufgeklärt wurden, beziehungsweise inwiefern ihre Privatsphäre durch den Austausch verletzt wurde (Murphy et al. 2021; Katznelson & Gerke 2021).

Beantwortung der Forschungsfragen

Auf Basis der beschriebenen Einordnung der Literatur zur Fairness medizinischer KI werden im Folgenden die zu Beginn aufgestellten Forschungsfragen beantwortet. Neben der Analyse möglicher Fälle, in denen unfaire Verhaltensweisen der KI im Gesundheitswesen auftreten können, werden sowohl die Gründe des Auftretens als auch existierende Lösungsansätze identifiziert.

Auftreten Unfairer KI

Um Verzerrungen in den Daten, Algorithmen oder Modellen sowie auch menschliche Verzerrungen beseitigen zu können, muss zunächst identifiziert werden, wo Verzerrungen auftreten können. Hinsichtlich der Vermutung von Ahmad et al. (2020), dass Verzerrungen in allen Bereichen eines klinischen Behandlungspfades auftreten können, werden im Folgenden auf Basis der vorhandenen Literaturauswahl die Prozessschritte einer klinischen Behandlung in Anbetracht des möglichen Auftretens von Verzerrungen im Detail analysiert.

Gemäß dem Prozessmodell des Reimbursement Instituts (siehe Abbildung 2) besteht an erster Stelle des Pfades eine Beschwerde, aufgrund derer sich ein Patient / eine Patientin in eine klinische Behandlung begibt. Sun et al. (2020) weisen darauf hin, dass bereits biologische Unterschiede in der Krankheitsentstehung an der Bildung der verzerrten Ergebnisse beteiligt sind. Frauen weisen beispielsweise bei akutem Myokardinfarkt andere Symptome auf als Männer, was im nächsten Schritt, der Untersuchung, dazu führen kann, dass die Entscheidungen der Gesundheitsdienstleister auf Basis männlicher Kriterien beruhen, wodurch Herzinfarkte bei Frauen somit nicht rechtzeitig erkannt werden und eine Verzerrung auftritt (Sun et al. 2020).

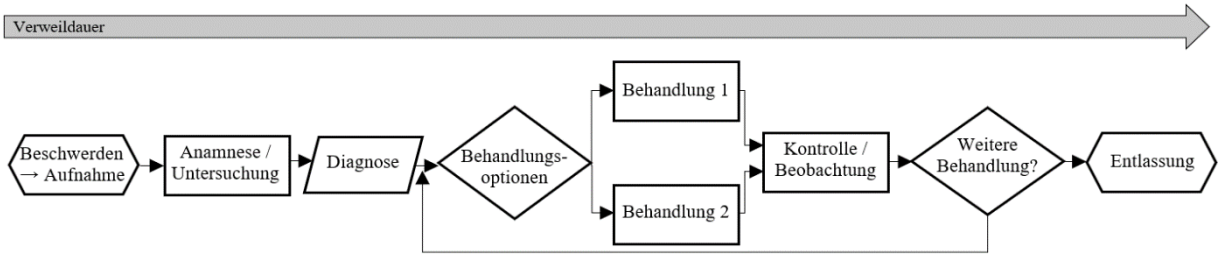

Abbildung 2. Klinischer Behandlungspfad, in Anlehnung an Reimbursement Institute (2021)

Im Rahmen der Anamnese der Beschwerden und der Untersuchung der Patienten steht der behandelnde Arzt vor der Entscheidung des Treffens einer passenden Diagnose und der Festlegung einer geeigneten Behandlungsoption. Wie bereits festgestellt werden konnte, können Ärzte in ihrer diagnostischen oder therapeutischen Entscheidungsfindung allerdings voreingenommen sein, was zu unfairen Ergebnissen führen kann (Kellmeyer 2019). Algorithmen, die vorhandene medizinische Daten objektiv synthetisieren und interpretieren, gelten demnach als vielversprechende Unterstützung zur Entscheidung über eine Diagnose (Ahmad et al. 2020). In der vorliegenden Literatur liegt der Fokus der meisten Studien durch den Einsatz computergestützter Diagnosesysteme bei dem Prozessschritt der Diagnose bzw. der Entscheidung

über mögliche Behandlungsoptionen. Das Auftreten der unfairen KI-Applikationen wird hierbei anhand verschiedener medizinischer Domänen aufgezeigt.

Im Bereich der Radiologie stehen medizinische Bildanalysen zur Erkennung von Krankheiten im Vordergrund der Untersuchungen. Verzerrungen treten hierbei bei der Detektion und Diagnose von lebensbedrohlichen Thorax Erkrankungen (Seyyed-Kalantari et al. 2020), der Erkennung von HIV auf Basis von Gehirn MRTs (Abbasi-Sureshjani et al. 2020), bei Gesichtserkennungstechniken (Larrazabal et al. 2020), der Diagnose von diabetischer Retinopathie (Burlina et al. 2021) der Detektion von Brust- und Lungenkrebs (Afrose et al. 2021) als auch bei der Erkennung und Identifizierung von Melanomen auf Basis der Röntgenbilder auf (Bissoto et al. 2020). Insbesondere die Dermatologie mit der Diagnose von gut und bösartigen Hautläsionen und der Erkennung von seltenen Hautkrankheiten kann hierbei als häufig untersuchtes Forschungsfeld des Auftretens unfairer Ergebnisse identifiziert werden (Bissoto et al. 2020; Weng et al. 2020). Auch in der klinischen Neurowissenschaft und in der psychiatrischen Entscheidungsfindung berichten Forscher hinsichtlich zunehmender Datenverfügbarkeit von ungleichen diagnostischen Ergebnissen und Fehleinschätzungen zu Schmerzbehandlungen aufgrund von unterrepräsentierten Populationen in den zugrundeliegenden Daten (Ienca & Ignatiadis 2020; Lee et al. 2021; Kellmeyer 2019; Starke et al. 2021). Beispielsweise berichten Starke et al. (2021) davon, dass schwarze Bürger häufiger mit Schizophrenie diagnostiziert werden, weiße Bürger eher mit Stimmungsschwankungen. Wie zu Beginn erwähnt, treten Verzerrungen auch im Bereich der Kardiologie auf. Geschlechts- und rassenspezifische Ungleichheiten treten hierbei sowohl bei der Diagnose von Herzkrankheiten auf als auch bei der Entscheidung über die folgende Behandlung: Männer werden bei Brustschmerzen mit häufigerer Wahrscheinlichkeit an einen Kardiologen überwiesen und schwarze Personen erhalten trotz gleicher Symptome mit geringerer Wahrscheinlichkeit Schmerzmedikamente (Tat et al. 2020). Auch neueste Untersuchungen zur COVID-19-Pandemie zeigen unverhältnismäßig große Auswirkungen auf ethische Minderheiten und Personen mit geringem Einkommen, die auf eine algorithmische Verzerrung zurückzuführen sind (Straw 2020).

Indirekt können weitere Verzerrungen im Verlauf des klinischen Prozesses bei der Verlaufsbeobachtung und Kontrolle identifiziert werden. Beispielsweise werden mobile zur kontinuierlichen Überwachung, Beobachtung und Verfolgung von Gesundheitsdaten der Patienten gesammelt und in elektronischen Gesundheitsakten dokumentiert (Brault & Saxena 2020). Die fehlende Zugänglichkeit verschiedener Patientengruppen zu den Möglichkeiten der sensorbasierten Datenerhebung als auch der Eingabe der Daten in Online-Portalen kann zu fehlenden oder fehlerhaften Daten von bestimmten Patientenpopulationen führen und somit schließlich wiederum bei der Verwendung der Daten (z. B. bei der Diagnose) zu unfairen Verhaltensweisen beitragen (Brault & Saxena 2020).

Der letzte wichtige Schritt im Verlauf eines klinischen Aufenthalts gilt der Entscheidung über eine weitere Behandlungsoption oder der Entlassung. Hierbei können Verzerrungen bei der Entscheidung einer Überweisung zu einem weiteren Gesundheitsdienstleister oder der prädiktiven Modellierung zur Vorhersage der Sterblichkeit im Krankenhaus, der Wiederaufnahme oder der Verweildauer anhand von klinischen Daten auftreten (Ahmad et al. 2020; Chen et al. 2019). Letztere prädiktive Vorhersage führen Chen et al. (2019) sowohl am Beispiel der Vorhersage der Patientensterblichkeit auf einer Intensivstation und der Vorhersage einer 30-tägigen psychiatrischen Wiedereinweisung in eine stationäre Einrichtung auf. In ihrer Analyse finden die Autoren basierend auf der Verwendung von klinischen Aufzeichnungen eine Heterogenität in den Attributen nach Rasse, Geschlecht und Versicherungsart (Chen et al. 2019).

Wie sich anhand der vorliegenden Analyse zeigt, können unfaire Ergebnisse durch die Nutzung der Technologien im medizinischen Sektor in jedem Schritt einer klinischen Behandlung auftreten. Insbesondere ist dennoch aufgrund des vielfältigen Einsatzes computergestützter Diagnosesysteme im Gesundheitswesen die Untersuchung eines Patienten in Kombination mit der Festlegung einer Diagnose als kritischer Prozessschritt für das Auftreten unfairer Ergebnisse hervorzuheben.

Gründe für das Auftreten

Die vorhergehenden Kapitel haben gezeigt, dass unfaire medizinische KI als Datenproblem und als Systemproblem auftreten kann. Das folgende Kapitel befasst sich mit den Gründen des Auftretens unfairer medizinischer KI und geht auf die Forschungsfrage ein, wodurch sie verursacht ist. Zuerst werden als

Gründe des Auftretens die Daten und daraufhin der Algorithmus betrachtet. Anschließend wird auf das Modell näher eingegangen, gefolgt von den Schnittstellen Mensch und KI.

Daten

Zu Beginn der vorliegenden Arbeit zeigt die Definition von KI, dass sowohl die Verfügbarkeit des Datensatzes als auch die Qualität und die Diversität dessen als maßgebliche Einflussfaktoren die Qualität der KI-Ergebnisse lenken. Alle Datensätze enthalten bereits im Voraus, aufgrund der Art und Weise der Datenerfassung und –annotierung, Verzerrungen, wenn auch unbeabsichtigt (Bissoto et al. 2019). Dies resultiert in verzerrten Leistungen der ML-Modelle, indem falsche Korrelationen erzeugt werden, welche von den Modellen auf unfaire Weise ausgenutzt werden. Verzerrte Daten treten beispielsweise durch die Unter- bzw. Überrepräsentation bestimmter Gruppen aufgrund von Alter, Geschlecht und ethnischer Zugehörigkeit auf (Brault & Saxena 2020). Diesen Daten ist eine Unausgewogenheit inhärent, wodurch sie nicht repräsentativ und verallgemeinerbar sind (Burlina et al. 2021). Die Unausgewogenheit der Datensätze entsteht wiederum aus verschiedenen Gründen und kann insbesondere auf die Erhebung der Daten, die Menge und die Qualität zurückgeführt werden. Historische Daten (z.B. Electronic Health Data) sind oft zugunsten bestimmter Personengruppen verzerrt. Beispielweise nehmen Frauen, Minderheiten und Menschen mit niedrigem sozioökonomischem Status weniger Gesundheitsdienste in Anspruch und folglich können diese Individuen nicht in dem Datensatz aufgenommen werden, da schlichtweg eine Dokumentation ihrer Daten fehlt (Seyyed-Kalantari et al. 2020). Auch kommen viele (Haut-) Krankheiten in der realen Welt nur sehr selten vor, wodurch öffentliche Datensätze, basierend auf einer natürlichen Patientenpopulation, zum einen sehr unausgewogen sein können und zum anderen nur wenige oder fehlende Beispiele für bestimme Bedingungen enthalten (Weng et al. 2020). Zwar vereinfachen alternative Datenerfassungstechnologien die Sammlung diverser Arten von Daten, dennoch resultieren Verzerrungen, da beispielsweise die Sammlung mobiler Gesundheitsdaten stark von dem Segment der Bevölkerung abhängt, welches überhaupt ein mobiles Gerät zur Kollektion der Daten besitzt (Brault & Saxena, 2020). Die Verantwortlichkeit einer repräsentativen Datensammlung im medizinischen Kontext, welche keinen Bias aufweist, wird aus diesen Gründen an die Wissenschaftler und KI-Entwickler gestellt (Ienca & Ignatiadis 2020).

Algorithmus

Weitere Gründe für das Auftreten unfairer medizinischer KI sind im Algorithmus verankert, da ein Lernalgorithmus eine Voreingenommenheit einführen kann (Srivastava & Rossi 2019). Im Falle der Nichtberücksichtigung demografischer Merkmale (wie z.B. Alter, Geschlecht, ethnische Zugehörigkeit) durch die verwendeten Algorithmen kann die Generalisierung über verschiedene demografische Untergruppen hinweg beeinträchtigt werden (Abbasi-Sureshjani et al. 2020). Infolgedessen kann ein schlecht konzipierter Algorithmus nicht gleich fair über alle Individuen hinweg ablaufen, was jedoch im Kontext der Medizin gegeben sein sollte (Obermeyer et al. 2019). Dem entgegenzuwirken durch die Optimierung eines Algorithmus kann in das Gegenteil umschwenken und selbst Bias hervorrufen. Die Überoptimierung eines Algorithmus, für die Erlangung höherer Akkuratheit während der Trainingsphase, kann einen Bias einführen und zudem die Generalisierbarkeit verringern (Ienca & Ignatiadis 2020). Außerdem führen unklare Anweisungen an einen Algorithmus dazu, dass er beispielsweise anstelle der Krankheit des Patienten die Kosten von ihm verwendet, um den Grad an Gesundheitsbedürfnissen zu approximieren (Katznelson & Gerke 2021). Hierfür bringt der Algorithmus die Patienten zwar in eine Rangfolge, jedoch bildet er diese ausgehend von einem falschen Wert, welcher sich nicht auf die eigentliche Krankheit bezieht. Eine weitere Ursache für Unfairness ist die algorithmische Verarbeitungsverzerrung (Starke et al. 2021). Hierbei stellt die Wahl der Regularisierung- und Glättungsparameter eine Möglichkeit dar, wo in der schrittweisen Prozedur des Algorithmus Bias ansetzen kann. Neben der Verarbeitungsverzerrung ist auch die Undurchsichtigkeit in den Verarbeitungsschritten ein Problem. Black-Boxen hindern die Identifizierung der bias-hervorrufenden Schritte im Algorithmus (Panch et al. 2019). Der Mangel an Transparenz von Black-Boxen erschwert die Nachverfolgung derjenigen Prozessschritte, die eigentlich aufgrund ihrer Hervorrufung von Unfairness eliminiert werden müssten. Demnach erzeugt das Fehlen von Transparenz und zusätzlich der Erklärungsmangel interner Arbeitsschritte Raum für die Entstehung von Bias (Ienca & Ignatiadis 2020).

Modell

Eine Ursache liegt ebenso in dem Training des Modells begründet. Modelle mit hoher Kapazität, die auf großen Datensätzen trainiert wurden, bieten keine natürliche Chancengleichheit. Stattdessen können sie zu potenziellen Ungleichheiten in der Versorgung führen, insofern sie ohne Modifikation eingesetzt werden (Seyyed-Kalantari et al. 2020). Sowohl das Training mit zu kleinen Datensätzen (Bissoto et al. 2019) als auch mit Datensätzen mit seltenen Merkmalen führt dazu, dass ein Modell weniger bereit sein wird, ähnliche Vorkommen zu erkennen (Lee et al. 2021). Dadurch wird es später im medizinischen Einsatz nicht fähig sein, zuverlässige Ergebnisse zu liefern. Das richtige Training eines Modells stellt einen zentralen Punkt dar, an welchem Bias auftreten oder bereits gemindert werden kann. Außerdem verursachen ML-Modelle Bias, indem sie falsche Korrelationen in den Daten bilden oder auch stichhaltige Korrelationen zerstören (Bissoto et al. 2019). Das Modell verlässt sich auf die Informationen, die es durch die Korrelation erhält, um Vorhersagen zu treffen; sollten die falschen Korrelationen gelernt worden sein, führt dies zu Bias (Bissoto et al. 2020). Da die verschiedenen leistungsstarken Modellklassen sehr komplex sind, können sie nicht oder nur kaum erklärt werden (Jabbari et al. 2020). Zudem ist die Vorhersageungenauigkeit eines Modells ein Indikator für unfaire Ergebnisse. Sie kann schlechter sein, wenn anwendungsspezifische Informationen fehlen (Chen et al. 2019). Ebenso wird Bias durch die Verwendung von herkömmlichen Metriken beim Trainieren und Testen von Modellen verursacht, da diese Metriken von Mehrheitsklassen beeinflusst werden (Afrose et al. 2021).

Schnittstellen Mensch-KI

Aufgrund der Beeinflussung von Trainingsdaten durch menschlichen Bias liegt eine Ursache unfairer medizinischer KI im menschlichen Handeln begründet (Geis et al. 2019). Nicht-diverse und/oder nicht-repräsentative (Trainings-) Daten, die die menschliche Voreingenommenheit widerspiegeln, können zu einem Bias bei der Bewertung von Anwendungen im Gesundheitswesen führen (Lee et al. 2021). Die Weitergabe sozial verankerter Vorurteile von Klinikern (Starke et al. 2021), die implizite diagnostische Voreingenommenheit von Ärzten und ihr Handeln auf einer nicht repräsentativen Datenbasis (Straw 2020), wie auch das Training von ML-Modellen auf Basis menschlicher Voreingenommenheit kann zu schädlichen Ausbreitungen des Bias führen (Lee et al. 2021). Dies sind alles Faktoren, die unteranderem für das KI-System eine Herausforderung darstellen, da die von der KI erlernte Grundwahrheit von der menschlichen Interpretation abhängig ist (Tat et al. 2020). Aktuelle Literatur zeigt, dass in der medizinischen Ausbildung eine Lücke bezüglich ethischer KI existiert (Katznelson & Gerke 2021). Die Verantwortlichkeitskluft zwischen Mensch und KI in der Interaktion (Kellmeyer 2019), sprich ob KI nur als zusätzliches Tool in der Diagnose genutzt wird oder Ärzte gar vollständig ersetzt und somit als autonomer Entscheidungsträger fungiert, kann aufgrund der fehlenden Diagnosekontrolle durch den Menschen ein Faktor für die unbemerkte Ausbreitung unfairer KI sein (Ienca & Ignatiadis 2020). Während Unternehmen mit gewinnmaximierenden Strategien vermehrt an dem Aspekt der Wirtschaftlichkeit interessiert sind, orientiert sich die Gesellschaft an dem Aspekt der Fairness. Dieser Interessenskonflikt, zwischen wirtschaftlich fokussierten Unternehmen und an der Fairness interessierten Gesellschaft, spiegelt sich in dem Verhalten einer unfairen KI wider (Miguel et al. 2020; McCradden et al. 2020). Ebenso liegt die Ursache unfairer KI in der Automatisierungsverzerrung des Menschen begründet, welche besagt, dass Menschen generell maschinengenerierte Daten favorisieren und gegenteilige Daten ignorieren (Geis et al. 2019).

Schlussfolgernd lässt sich folgendes über die Ursachen für das Auftreten unfairer KI sagen: Der Mensch nimmt bereits bei der Datensammlung aktiv Einfluss auf möglichen Bias und fortführend hierzu ebenso an anderen Schnittstellen, wie der Entwicklung des Algorithmus etc. Der Einsatz von nicht repräsentativen und/oder verzerrten Daten beim Trainieren eines KI-Modells führt zur Diskriminierung und erzeugt keine fairen Untersuchungsergebnisse gegenüber aller Personen. Zudem sind die einzelnen Prozessschritte des Algorithmus nicht erklärbar und entziehen sich kausalen Erklärungen. Aus diesem Grund kann der Lösungsweg des Algorithmus oftmals nicht nachvollzogen werden und er erscheint nicht vertrauenswürdig.

Lösungsansätze

Zur Überwindung potenzieller Fairnessprobleme bietet die Literatur vielseitige Lösungswege, welche für alle Systemkomponenten Optimierungsmöglichkeiten aufzeigen. Größtenteils stehen die Autoren jedoch

im Einklang darüber, dass primär den zum Training und Testen der KI verwendeten Daten eine besondere Beachtung zukommen muss (Srivastava & Rossi 2019; Obermeyer et al. 2019, Abbasi-Sureshia et al. 2020; Bissoto et al. 2020; Lakeh et al. 2021; Sun et al. 2020; Martín Noguerol et al. 2019; Tat et al. 2020). Diversität in den Trainings-, vor allem aber in den Testdatensätzen, soll zu robusten und zuverlässigen Lösungen führen (Bissoto et al. 2020). Vorrangig sollten diese Daten jedoch normalisiert sowie standardisiert und alle störenden Parameter für das Training der Hauptzielaufgabe verwendet werden, um deren Korrektur effektiver zu gestalten (Abbasi-Sureshja et al. 2020). Das alleinige Hinzufügen von zusätzlich klinisch relevanten Daten hingegen bringt keine zusätzliche Verbesserung der KI-Prognose. Ergo benötigen faire Vorhersagen nicht mehr, sondern bessere Daten sowie bessere Methoden zur Messung bzw. Evaluation der Datenverzerrungen (Bissoto et al. 2019). Auch Brault und Saxena (2020) sehen die ausschließliche Bereitstellung vielfältiger und großer Mengen an Trainingsdaten als unzureichend für die Beseitigung von Vorurteilen. Denn Verzerrungen sind per se unempfindlich gegenüber der Größe der Stichprobe, sodass der systematische Fehler mit der Vergrößerung der Stichprobe dazu neigt, sich aufzusummieren (Brault & Saxena 2020). Dem entgegensetzt kann die Verwendung von Multi-Source Datensätzen, sprich die Verwendung großer Datensätze, auch dazu führen, potenzielle Verzerrungen in der Datenerfassung sowie Fairnesslücken zu reduzieren (Seyyed-Kalantari et al. 2020; Sun et al. 2020). In dieselbe Richtung argumentierend kann die Verwendung von generativen adversen Netzwerken (GANS) als Ansatz gesehen werden, um eine große Menge an gelabelten Daten aus einem kleinen Datensatz zu erhalten (Martín Noguerol et al. 2019). Zusätzlich bilden neuartige generative Methoden eine gute Basis zur Angleichung der Vorhersageergebnisse für Patienten unterschiedlicher Subpopulationsgruppen durch Behebung des Problems fehlender Subpopulations-Trainingsdaten (Burlina et al. 2021). Des Weiteren wird eine Qualitätskontrolle der Datenverarbeitung und eine regelmäßige Annotation durch geschulte Kliniker (Tat et al. 2020), die Qualitätssicherung der Daten mittels den verschiedenen Dimensionen Vollständigkeit, Richtigkeit, Konsistenz und Aktualität (Hee 2017), sowie angemessene Regulationen des daten-wissenschaftlichen Prozesses (Panch et al. 2019) von der Literatur als wichtig zur Überwindung von Fairnessproblemen erachtet. Mögliche Datenverzerrungen sollten bestenfalls im Merkmalsraum behandelt werden, um mögliche Artefakte vom diagnostischen Label zu trennen (Bissoto et al. 2020). Um etwaige menschlich verursachte Verzerrungen der Inputdaten zu beheben, bietet die Nutzung von bias-bewussten Algorithmen einen möglichen Lösungsansatz (Ahsen et al. 2019). Die Qualität der Trainings- und Testdaten stellt also eine besondere Herausforderung dar, denn sie haben maßgeblichen Einfluss auf die Sensitivität und Spezifität eines Algorithmus (Lee et al. 2019).

Ein erster unabdingbarer Schritt zur Erlangung fairer KI-Anwendungen ist jedoch die Identifizierung der Minderheitsvorhersageklasse und der demografischen Minderheitsgruppen im Trainingsdatensatz (Afrose et al. 2021). Da konventionelle ML Prognosen dem Paradigma folgen, in einem Modell Vorhersagen für alle demographischen Klassen zu treffen, sollten unterschiedliche Modelle für unterrepräsentierte Altersgruppen oder Patienten unterschiedlicher ethnischer Herkunft entwickelt werden (Afrose et al. 2021). Dies impliziert, dass nicht das gleiche Fairness-Kriterium für alle Modelle verwendet werden kann, sondern dieses, je nach Anforderung an die KI, angepasst werden muss (Liu et al. 2020).

Da medizinische KI-Anwendungen sowohl ihren Nutzen als auch die Fairness ihrer Ergebnisse optimieren müssen, stellt die Lösung dieser Aufgabe ein NP-schweres Optimierungsproblem dar, denn zeitgleich müssen beide Variablen maximiert werden. Ein möglicher Ansatz zur Lösung dieses Problems sehen Chester et al. (2020) darin, einen minimal akzeptablen Schwellenwert für eine der beiden Variablen zu bestimmen um anschließend den maximal erreichbaren Wert für den anderen Parameter zu bestimmen. Darüber hinaus kann die Fairness einer Modellprognose ebenfalls über eine enge, nicht konkurrierende, kooperative Beziehung zwischen Ärzten und KIs erlangt werden, durch welche der Arzt der KI Feedback zur Anwendung eines Algorithmus geben kann und die KI den behandelnden Arzt bei unsicheren Fällen aktiv um Hilfe bittet (Chen et al. 2019). Dies gelingt allerdings nur, wenn die KI interpretierbare Algorithmen verwendet und auf Bias- und Diskriminierungsprobleme aufmerksam machen kann (Kellmeyer 2019). Darüber hinaus muss auch die Zusammenarbeit von Medizinern, Ingenieuren und Ethikern gewährleistet sein (Ienca & Ingnatiadis 2020). Eine größere Transparenz der KI-Anwendung lässt zudem algorithmische Verzerrung leichter erkennen und hilft fehlerhafte Entscheidungen zu verstehen, um sie zu korrigieren und Eingabevariablen entsprechend anzupassen (Starke et al. 2021). So kann durch transparente, erklärbare KI der klinische Nutzen sowie das Vertrauen in sie gesteigert werden (Lee et al. 2019).

Des Weiteren sollten Algorithmen keine ärztlichen Entscheidungen abnehmen, sondern lediglich als Unterstützung dienen. Entsprechend muss der verwendete Algorithmus designt werden (Obermeyer et al. 2019). Dieser sollte sensitiv für mögliche Bias sein (Abbasi- Sureshjani et al. 2020) und darüber hinaus Kodizes der medizinischen Ethik und ethnische Rahmenwerke integrieren (Straw 2020). Nach Abschluss des Trainings ist es in der Trainingsphase eines KI-Algorithmus ebenfalls unerlässlich, diesen in einem unabhängigen Datensatz zu validieren (Tat et al. 2020). Dementsprechend können etwaige Klassen-ungleichgewichte, beispielsweise durch Resampling Strategien, bereits beim maschinellen Lernen gelöst werden (Weng et al. 2020).

Für eine faire KI-Anwendung ist es unabdingbar, den kompletten Entwicklungsprozess fair zu gestalten (Hague 2019). Dies impliziert die gleiche Repräsentation der Daten, ein geeignetes Algorithmusdesign, die Verwendung angepasster Fairness-Metriken, den Einsatz erklärbarer KI sowie die Anerkennung und Erklärung des Bias eines Modells (Ahmad et al. 2020). Demzufolge sind an die Entwicklung und Nutzung fairer KI in der Medizin drei Anforderungen zu stellen (Brault & Saxena 2020):

1. die Erstellung eines Kataloges der möglichen Verzerrungen und deren Ausmaß, Richtung und Phase, in welcher der Bias auftritt (Eingabe, Analyse, Ausgabe)
2. die Entwicklung methodischer Standards für die Nutzung von KI im medizinischen Kontext
3. die Entwicklung einer kritischen Bewertung von KI in der Wissenschaft im Allgemeinen und insbesondere in der Medizin

Diskussion

Im folgenden Kapitel werden die Ergebnisse der vorliegenden Untersuchung dargelegt sowie kritisch hinterfragt als auch notwendige Implikationen für Forschung und Praxis abgeleitet. Unter Berücksichtigung der Limitationen wird schließlich ein Ausblick auf zukünftige Forschungsarbeiten gegeben.

Zentrale Erkenntnisse

Das Feld der Fairness medizinischer KI hat in letzter Zeit viel Aufmerksamkeit, sowohl in der Forschung als auch in der Praxis erlangt. Aufgrund eines fehlenden Überblicks über den aktuellen Stand der Forschung zum Thema, galt es als Aufgabe dieser Arbeit, ein systematisches Literaturreview durchzuführen. Die Suche in den Datenbanken lieferte eine umfassende Anzahl an Artikeln, die tiefgehende Erkenntnisse zur Beantwortung der aufgestellten Forschungsfragen lieferten. Die Klassifizierung der 35 identifizierten Artikel in Bezug auf ein Daten- bzw. Systemproblem in Anlehnung an Ahmad et al. (2020) erwies sich als besonders hilfreich.

Abbildung 3 zeigt die entwickelte Taxonomie und die Klassifizierung der vorhandenen Literatur in Bezug auf die Fairness der medizinischen KI. Das Datenproblem lässt sich hierbei in die Erhebung, die Qualität und Menge der Daten unterteilen, das Systemproblem weiter in die Bereiche Algorithmus, Modell und die Schnittstelle zwischen Menschen und der künstlichen Intelligenz. Allerdings ist hierbei hervorzuheben, dass einige Artikel Aspekte aus mehreren Gruppen abdecken und demnach nicht einheitlich in die Klassifizierung eingeordnet werden konnten. Dies war insbesondere bei Artikeln der Fall, die die Fairness der KI in Verbindung der Daten mit dem Algorithmus, bzw. dem Modell sahen. Eine detaillierte Einordnung der einzelnen Artikel des Literaturreviews in die Bereiche sowie die vorhandenen Überschneidungen befindet sich im Anhang Teil A. Diese Einordnung kann als Fahrplan gesehen werden, die ihre Betrachter durch die, innerhalb dieser Arbeit klassifizierte, Literatur navigiert. Die zentralen Erkenntnisse, die der ausgewählten Literatur entnommen wurden, werden im Folgenden aufgegriffen.

Betrachtet man die Schnittstellen zwischen Menschen und KI, wird ersichtlich, dass an solchen viele Entstehungsmöglichkeiten für Bias vorliegen. Auf Seite der KI-Entwicklung sind die individuellen Verzerrungen der Entwickler und deren Motive zentrale Ursachen für Bias in den KI-Anwendungen. Dabei stellt ein gewinnmaximierendes Motiv von Unternehmen eine enorme Hürde für die Fairness dar, da es der statistisch signifikanten Diversität von Ressourcen im Weg steht. Der daraus resultierende Bedarf an Kontrollmechanismen deckt nicht in direkter Linie die Interessen der Unternehmen und muss deswegen von außen auferlegt werden. Dies stellt Politik und Wirtschaft vor große Herausforderungen. Auf Seiten der Nutzer, gekoppelt an die Aspekte aus den Bereichen der Modelle und Algorithmen, entsteht Bias gegenüber KI vor allem durch Vertrauensprobleme. Fehlendes technisches Wissen gepaart mit Blackbox-

Eigenschaften der KI und Angst um Verletzung des privaten Datenschutzes verringern die Akzeptanz und somit die Wirksamkeit von KI-Produkten. Aus diesem Grund sind Aufklärungsarbeiten und Schulungen der breiten Bevölkerung von essenzieller Bedeutung. Durch das Verständnis um die Verwendung der eigenen Daten, sowie die Funktionsweise der Technologie kann die Akzeptanz gesteigert werden. Darüber hinaus wurde erkannt, dass KI-Produkte unterstützende und nicht entscheidende Funktionen übernehmen sollte, und somit medizinisches Personal nicht ersetzen sollte.

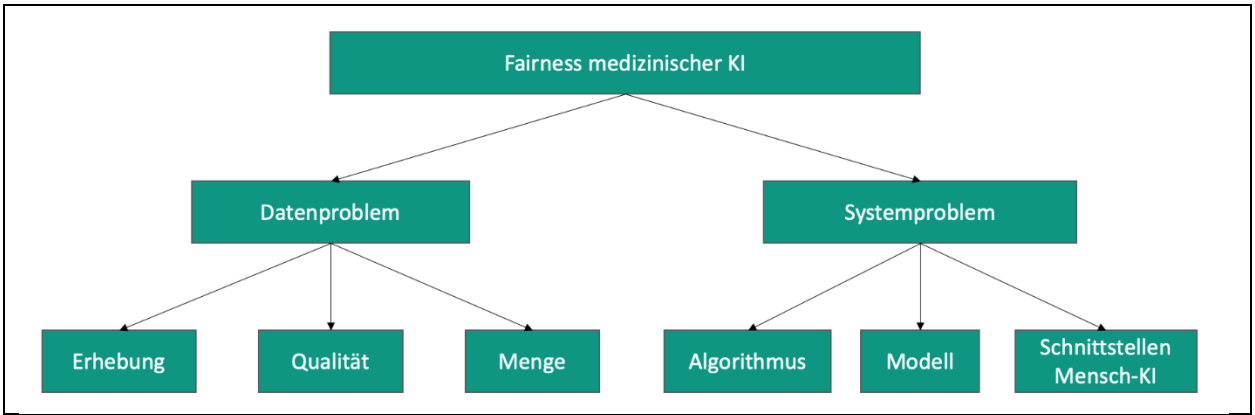

Abbildung 3. Taxonomie zur Fairness Medizinischer Künstlicher Intelligenz

Im Falle der Algorithmen, sowie der verwendeten Modelle wurde dargelegt in welchen Zusammenhängen die Performance, die Fairness und die Erklärbarkeit der KI stehen. Es wurde detailliert erläutert, dass aufgrund der Blackbox-Eigenschaften komplexer Algorithmen und Modelle eine paarweise negative Korrelation besteht. Überträgt man diesen Sachverhalt auf den zuvor genannten Bedarf an Transparenz und Erklärbarkeit der KI, ergibt sich gleichzeitig eine schlechtere Leistung und möglicherweise sogar geringere Fairness der KI. Eine Möglichkeit diesem Problem zu begegnen, wird darin gesehen einer Variablen einen Schwellenwerte zuzuordnen und unter den Nebenbedingungen die anderen Parameter zu optimieren. Darüber hinaus wurde festgestellt, dass anders als in konventionellen Konzepten, Anwendungen spezifisch für verschiedene soziodemografische Gruppen und ethische Minderheiten entwickelt werden sollten.

Das Datenproblem lässt sich weiter unterteilen in die Erhebung, die Qualität und die Menge der Daten. Dabei wurde festgestellt, dass sich in den historischen, klinischen Daten Verzerrungen manifestiert haben, die dann bei Erhebung auch in die KI miteinfließen. Bei der Erhebung großer Mengen an Daten aus großen Datenbanken, muss darauf geachtet werden, dass zwar die Vielfalt an Daten besser abgedeckt werden kann, sie aber aufgrund ihrer Unstrukturiertheit oft zu wenig Aussagekraft besitzen. Der Rückschluss, dass durch den Zugriff auf kleinerer Datenbestände Bias verringert werden kann ist jedoch umstritten. Vielmehr scheint sich die aktuelle Literatur einig zu sein, dass mehr auf die Qualität der Daten als auf die Menge geachtet werden müsse, da die Verzerrungen recht unempfindlich gegenüber der Stichprobengröße sein können. In erster Linie sollen die Daten in Bezug auf das Hauptziel standardisiert und normalisiert werden. Mit diesem Aspekt einhergehend kristallisiert sich der Bedarf an besseren Methoden zu Qualitätssicherung der Daten heraus.

Anhand der zentralen Erkenntnisse lässt sich erkennen, dass innerhalb dieses jungen Forschungsfelds viele Aspekte zum Auftreten von Bias untersucht werden. Dabei ist ein Gradient an Einigkeit in den verschiedenen Debatten zu erkennen. Besonders in dem Umgang mit den Daten, sowohl bei der Erhebung als auch der Menge der Daten beherrscht keine klare Meinung bezüglich der Problemlösungsansätze das Forschungsfeld. Zwar geht die allgemeine Tendenz auf die Qualitäts-sicherung als verheißungsvolle Reduzierung von Bias zu, jedoch werden in der untersuchten Literatur diesbezüglich keine konkreten Konzepte erläutert. Ein Kernproblem dieser Thematik wird dabei darin gesehen, dass keine klare Definition von fairer KI vorliegt, da sich diese je nach Betrachtungswinkel verändert. In Korrelation dazu, existieren keine signifikanten Metriken, anhand derer eine KI-Anwendung bewertet werden kann. Höhere Einigkeit herrscht dagegen bei dem Ansatz Anwendungen für spezifische Personengruppen zu entwickeln und die Gesellschaft hinsichtlich intelligenter Technologien zu unterrichten, um das Vertrauen zu steigern.

Im Übertrag auf das Ziel der Arbeit lässt sich sagen, dass ein Überblick über die Literatur erfolgreich erstellt wurde. Gleichzeitig muss vor Augen geführt werden, dass keine Garantie für die vorliegende Vollständigkeit aktueller Forschung in der ausgewählten Stichprobe gegeben werden kann. Der verwendete, recht allgemeine Suchterm ermöglichte dabei Einblicke in viele Teilbereiche der Medizin, Technik und Ethik, ergab jedoch keinen detaillierten Überblick über einzelne medizinische Bereiche. Dennoch repräsentiert und unterteilt die erstellte Taxonomie die systematischen und systemischen Quellen für Bias medizinischer KI, die in der aktuellen Forschung diskutiert werden. Bezüglich der Eingliederung der Fairness medizinischer KI in die Literatur nichtmedizinischer KI fällt eine klare Abgrenzung schwer. Grundlegend ist die Gesellschaft gegenüber Fairness in KI nicht sonderlich sensitiv solange keine schwerwiegenden Folgen daraus resultieren können, was in den meisten Anwendungsfällen herkömmlicher KI der Fall ist. Im Gesundheitswesen jedoch sind diese Folgen gegeben, im schlimmsten Fall sogar fatal. Aufgrund dessen kommt der Fairness medizinischer KI zurecht eine große Aufmerksamkeit zu und erfordert spezielle Maßnahmen.

Zusammenfassend decken die erarbeiteten Ergebnisse in weiten Teilen unsere Erwartungen. Etwas unerwartet war die Uneinigkeit im Umgang mit Erhebung und Menge der Daten. An dieser Stelle wurde davon ausgegangen, dass es eindeutig zielführend sei große, diverse Mengen an Daten zu verwenden. Des Weiteren wurden konkretere Ansätze zu regulatorischen und datenschutzrechtlichen Richtlinien erwartet.

Implikationen für Forschung und Praxis

Anhand der vorliegenden Untersuchung konnte zunächst einhergehend mit der zugrundeliegenden Literatur keine einheitliche Definition der Fairness im medizinischen Kontext gefunden werden. Somit steht im Wesentlichen nicht fest, welches Ziel eine faire medizinische KI erreichen soll und aus wessen Sicht dies geschieht. Einerseits besteht die Möglichkeit der Optimierung der Fairness für Individuen und Subpopulationsgruppen, andererseits ist die Optimierung stakeholderabhängig. In Anlehnung an die zugrundeliegende rechnerische, soziale als auch kognitive Dimension der Fairness und die Unterscheidung in die individuelle und Gruppenfairness muss sowohl in der Forschung als auch in der Praxis abgewogen werden, welches Fairnesskonstrukt sich am besten für die Anwendung in einem spezifischen medizinischen Bereich eignet. Auch gilt es, einheitliche Fairness Metriken zu definieren, um messen zu können, wann es sich tatsächlich um unfaire Ergebnisse handelt oder die Verzerrung möglicherweise gerechtfertigt ist.

Die zentralen Ergebnisse der vorliegenden Arbeit zeigen, dass die Fairness medizinischer KI nicht als einseitiges Problem gesehen werden kann, sondern es sich um ein komplexes Systemproblem handelt, welches diverse Komponenten wie die verwendeten Daten, den angewandten Algorithmus, das resultierende Modell als auch den menschlichen Faktor beinhaltet. Somit ist eine gesamtheitliche Betrachtung des Zusammenspiels der Systemkomponenten notwendig, um zukünftig praktikable Ansätze zu entwickeln. Aus praktischer Sicht können die Erkenntnisse aus einigen beispielhaften Bereichen zur Optimierung bestehender KI-Anwendungen genutzt werden. Jedoch müssen weiterhin für jeden medizinischen Bereich, wie etwa die Dermatologie oder Neurologie, individuelle und spezifische Anforderungen an eine KI beachtet und definiert werden. Demzufolge muss für jeden einzelnen Bereich entsprechend erprobt werden, welche Inputdaten repräsentativ und von Relevanz sind, welcher Algorithmus und welches Modell am geeignetsten sind und wie der Trade-Off zwischen Leistung, Fairness und Erklärbarkeit optimal umgesetzt werden kann. Informatiker und Praktiker sollten somit gemeinsam mit Medizinern, Epidemiologen und Ethikern kontinuierlich zusammenarbeiten, um die Absichten hinter dem Design der Algorithmen und Modelle zu diskutieren. Auch sollte jederzeit gewährleistet sein, dass die KI-Anwendungen bewusst umprogrammiert werden können, wenn potenzielle Verzerrungen auftreten. Es ist also von zentraler Wichtigkeit, die Technik nicht sich selbst zu überlassen, diese stets kritisch zu hinterfragen und eine kontinuierliche menschliche Beteiligung sicherzustellen.

Um die Verzerrungen durch die angewandten KIs präventiv vorzubeugen, sollte bei der Entwicklung neuer Technologien zukünftig bereits von Beginn an der Aspekt der Fairness einbezogen werden. Beispielsweise sollten angesichts der aktuellen COVID-19-Pandemie Gerätschaften und computergestützte Systeme zur Vorhersage und Diagnose der Krankheit auf Basis repräsentativer Datensätze, die unterschiedliche auftretende Symptome und Merkmale in Abhängigkeit von Geschlecht und Alter einbeziehen, entwickelt und programmiert werden.

Limitationen und Zukünftige Forschung

Das Forschungsfeld der Fairness medizinischer KI stellt einen noch recht jungen und zugleich dynamischen Forschungsbereich dar, in welchem regelmäßig neue Erkenntnisse gewonnen und Durchbrüche erzielt werden. Somit ist nicht auszuschließen, dass in naher Zukunft innovative Entwicklungen einer fairen KI Anwendungsfälle ermöglichen, welche in dieser Arbeit nicht identifiziert und dargestellt wurden. Um dieser Lücke entgegenzuwirken, wurden ebenfalls Preprints aus den Datenbanken *ArXiv* und *MedRxiv* in die Analyse des Forschungsstandes aufgenommen und Erkenntnisse aus der Grundlagenforschung zu dem Thema Fairness in medizinischer KI in die Forschungsimplikationen übernommen.

Dem Problem eines möglichen Publikations-Bias konnte jedoch nicht begegnet werden. So beinhaltet dieses Literaturreview möglicherweise einseitige Forschungsergebnisse, welche auf (signifikant) positiven Ergebnissen beruhen, und beleuchtet keine Erkenntnisse zu erfolglosen Methoden, welche jedoch für die Praxis ebenfalls von Wichtigkeit sein könnten. Dementsprechend kann eine weitere Spezifizierung des Suchterms dazu beitragen, möglicherweise seltener genutzte Ansätze zu identifizieren und diese so ins Blickfeld von Forschung und Praxis rücken. Weiterhin kann durch die fehlende klare Abgrenzung der Fairness in medizinischer KI gegenüber Trustworthy AI bzw. der Fairness allgemeiner KI kein exakt definierter Rahmen für die Literaturanalyse dieser Arbeit gegeben werden. Folglich bietet die vorliegende Arbeit lediglich einen allgemeinen Überblick über den aktuellen Stand der Forschung zu Fairness in medizinischer KI und ist deshalb nicht generalisierbar. Die Übertragbarkeit der gelieferten Erkenntnisse auf einzelne, spezifische medizinische Bereiche muss demzufolge im Einzelnen überprüft und validiert werden.

Infolgedessen sollte sich zukünftige Forschung mit einer allgemein gültigen Definition und Abgrenzung von Fairness im medizinischen Kontext beschäftigen. Weiterhin gilt es, die Übertragbarkeit zentraler Erkenntnisse (z.B. der benötigten Datenmenge und deren enthaltene Informationen) auf einzelne medizinische Anwendungsfelder genauer zu untersuchen. Darüber hinaus muss analysiert werden, inwiefern medizinische KI-Anwendungen länderübergreifend eingesetzt werden können oder ob unterschiedliche ethnische Gesellschaftsanteile einen Nachteil für bestimmte Populationsgruppen mit sich bringen. Einer der wichtigsten Punkte zur Erlangung fairer und leistungsstarker KIs stellt jedoch die Erforschung eines geeigneten Trade-Offs zwischen ihrer Leistung, Fairness und Erklärbarkeit dar, weshalb hierfür zwingend ein praktikabler Ansatz entwickelt werden muss.

Zusammenfassung

Mit dem vorliegenden systematischen Literaturreview ist ein Überblick über den aktuellen Forschungsstand der Fairness medizinischer KI zu der Beantwortung der anfangs aufgeführten Forschungsfragen gegeben. Mit Hilfe eines Suchterms wurden in verschiedenen Datenbanken eine Vielzahl von Artikel gefunden, von welchen 35 als relevante Artikel identifiziert wurden. Die finalen Artikel wurden anschließend systematisch klassifiziert und analysiert. Die erstellte Taxonomie zeigt, dass die Artikel in zwei Hauptgruppen mit zugehörigen Untergruppen zugeordnet werden können, insofern es sich bei unfairer medizinischer KI um ein Daten- oder ein Systemproblem handelt.

Die vorliegende Arbeit kommt zu dem Schluss, dass unfaire medizinische KI in vielen medizinischen Bereichen als auch an vielen Punkten während der klinischen Behandlung auftritt. Die Radiologie, Dermatologie, Kardiologie und Neurowissenschaften sind stellvertretend als Beispiele für viele medizinische Felder zu nennen, in denen das Problem unfairer KI existiert. Bereits Unterschiede beim Auftreten von Krankheitssymptomen können bei der Diagnose und Untersuchung zu ungleichen Handlungsempfehlungen führen. Auch bei andauernden Kontrollen und Beobachtungen des Krankheitsverlaufs ist eine Verzerrung nicht auszuschließen. Unfaire medizinische KI hat dabei unterschiedliche Gründe des Auftretens, weshalb kein eindeutiger Grund für unfaire KI identifiziert werden kann. Die vorliegende Arbeit hat als Ursache der Problematik die Daten und deren Erhebung, Qualität und Menge, den Algorithmus, das Modell und die Schnittstellen von Menschen zur KI identifiziert. Daten mit einer Unter- bzw. Überrepräsentation einer bestimmten Gruppe führen zu Verzerrungen, da diese keine repräsentative Grundlage darstellen, welche stellvertretend für die gesamte Bevölkerung steht. Unklare Anweisungen an einen Algorithmus führen zu falschen Prozessschritten. Zudem erschwert die Black-Box-Eigenschaft vieler KIs die Identifikation der Problemstellen aufgrund eines Mangels an Transparenz. Durch das Erlernen und Bilden falscher Korrelationen verursachen Modelle ebenfalls Verzerrungen. Daneben

müssen Modelle einen Trade-Off zwischen Fairness, Modellleistung und Modellerklärbarkeit eingehen, was die Entwicklung einer fairen, leistungsstarken und transparenten KI zu einem NP-schweren Optimierungsproblem macht. Die Schnittstellen von Entwicklern und Mitwirkenden, wie auch von Nutzern mit der KI stellen ebenso eine Ursache für die Entstehung unfairer medizinischer KI dar.

Für das Problem ergeben sich verschiedene Lösungsansätze, wie beispielsweise bessere Daten für das Trainieren und Testen einer KI sowie bessere Methoden zur Messung und Evaluation der Datenverzerrungen. Für unterrepräsentierte Bevölkerungsgruppen sind unterschiedliche Modelle mit einer größeren Transparenz zu gestalten. Zudem sollten Algorithmen Ärzte bei ihrer Arbeit unterstützen, sodass kein konkurrierendes Verhältnis, sondern vielmehr eine Kooperation vorherrscht. Schlussfolgernd ergeben sich nach Brault und Saxena (2020) drei Anforderungen, welche an die Entwicklung und Nutzung fairer medizinischer KI zu stellen sind: die Erstellung eines Katalogs möglicher Verzerrungen und deren Ausmaß, Richtung und Phase, in welcher Bias auftritt; die Entwicklung methodischer Standards für die KI-Nutzung in der Medizin und die Entwicklung einer kritischen Bewertung von KI im allgemeinen und medizinischen Kontext.

Die Analyse der vorliegenden Arbeit hat gezeigt, dass Fairness medizinischer KI ein komplexes, systemkomponentenübergreifendes Problem darstellt. Aufgrund dessen ist die ganzheitliche Betrachtung des Zusammenspiels der Systemkomponenten nötig, um bei der zukünftigen Entwicklung medizinischer KI Fairness zu gewährleistet.

Literaturverzeichnis

Abbasi-Sureshjani, S., Raumanns, R., Michels, B. E. J., Schouten, G., & Cheplygina, V. (2020). Risk of Training Diagnostic Algorithms on Data with Demographic Bias. *ArXiv Preprint arXiv:2005.10050*, http://arxiv.org/pdf/2005.10050v2

Afrose, S., Song, W., Nemeroff, C. B., Lu, C. & Yao, D. (2021). Subpopulation-specific Machine Learning Prognosis for Underrepresented Patients with Double Prioritized Bias Correction. *MedRXiv Preprint,* https://doi.org/10.1101/2021.03.26.21254401

Ahmad, M. A., Eckert, C., Allen, C., Hu, J., Kumar, V. & Teredesai, A. (2020). Fairness in Healthcare AI. *KDD2020,* https://www.youtube.com/watch?v=MzuoWAk9_AQ

Ahmad, M. A., Patel, A., Eckert, C., Kumar, V. & Teredesai, A. (2020). Fairness in Machine Learning for Healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3529-3530.

Ahsen, M. E., Ayvaci, M. U. S. & Raghunathan, S. (2019). When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis. *Information Systems Research,* 30(1), 97-116.

Bauer, G., Kechaja, M., Engelmann, S. & Haug, L. (2021). Diskriminierung und Antidiskriminierung: Beiträge aus Wissenschaft und Praxis. Bielefeld: transcript.

Baumgartner, R. (2021). Künstliche Intelligenz in der Medizin: Diskriminierung oder Fairness? In G. Bauer, M. Kechaja, S. Engelmann & L. Haug (Hrsg.), *D*iskriminierung und Antidiskriminierung: Beiträge aus Wissenschaft und Praxis. (pp. 149-164) Bielefeld: transcript.

Bissoto, A., Fornaciali, M., Valle, E. & Avila, S. (2019). Constructing Bias on Skin Lesion Datasets. *ArXiv Preprint arXiv:1904.08818*. https://arxiv.org/pdf/1904.08818

Bissoto, A., Valle, E. & Avila, S. (2020). Debiasing Skin Lesion Datasets and Models? Not So Fast. *ArXiv Preprint arXiv:2004.11457*. http://arxiv.org/pdf/2004.11457v1

Brault, N. & Saxena, M. (2021). For a critical appraisal of artificial intelligence in healthcare: the problem of bias in mHealth. *Journal of Evaluation in Clinical Practice,* 27(3), 513-519.

Braveman, P. (2006). Health disparities and health equity: concepts and measurement. *Annual review of public health* (27), 167-194.

Burlina, P., Joshi, N., Paul, W., Pacheco, K. D. & Bressler, N. M. (2021). Addressing Artificial Intelligence Bias in Retinal Disease Diagnostics. *ArXiv Preprint arXiv:2004.11457*. https://arxiv.org/abs/2004.13515

Chen, I. Y., Szolovits, P. & Ghassemi, M. (2019). Can AI Help Reduce Disparities in General Medical and Mental Health Care. *AMA Journal of Ethics,* 21(2), 167-179.

Chester, A., Koh, Y. S. & Wicker, J., Sun, Q. & Lee, J. (2020). Balancing Utility and Fairness against Privacy in Medical Data. *IEEE Symposium Series on Computational Intelligence (SSCI)*, 1226-1233.

Davenport, T. & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal,* 6:(2), 94-98.

Deloitte. (2019). Forces of change: The future of health. https://www2.deloitte.com/za/en/pages/life-sciences-and-healthcare/articles/the-future-of-health.html

Duden. (2021). Bias. https://www.duden.de/rechtschreibung/Bias

Evans, M. & Wilde Mathews, A. (2019.) Researchers Find Racial Bias in Hospital Algorithm. https://www.wsj.com/articles/researchers-find-racial-bias-in-hospital-algorithm-11571941096

Gade, K., Geyik, S., Kenthapadi, K., Mithal, V. & Taly, A. (2020). Explainable AI in Industry. *FAT 2020 Tutorial.* https://www.slideshare.net/KrishnaramKenthapadi/explainable-ai-in-industry-fat-2020-tutorial

Gade, K., Geyik, S., Kenthapadi, K., Mithal, V. & Taly, A. (2020). Explainable AI in Industry: Practical Challenges and Lessons Learned. In *Companion Proceedings of the Web Conference 2020*, 303–304.

Geis, J. R., Brady, A., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Kitts, A. B., Birch, J., Shields, W. F., van Hoven Genderen, R., Kotter, E., Gichoya, J. W., Cook, T. S., Morgan, M. B., an Tang, Safdar, N. M. & Kohli, M. (2019). Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Insights into imaging,* 10(1), 101.

Hague, D. C. (2019). Benefits, Pitfalls, and Potential Bias in Health Care AI. *North Carolina Medical Journal,* 80(4), 219-223.

Hee, K. (2017). Is data quality enough for a clinical decision?: Apply machine learning and avoid bias. In *2017 IEEE International Conference on Big Data (Big Data)*, 2612-2619.

Hofmann, J., Kersting, N., Ritzi, C. & Schünemann, W. J. (2019). *Politik in der digitalen Gesellschaft*: *Zentrale Problemfelder und Forschungsperspektiven.* Bielefeld: transcript.

Ienca, M. & Ignatiadis, K. (2020). Artificial Intelligence in Clinical Neuroscience: Methodological and Ethical Challenges. *AJOB neuroscience* 11(2), 77-87.

Jabbari, S., Ou, H.-C., Lakkaraju, H. & Tambe, M. (2020). An Empirical Study of the Trade-Offs Between Interpretability and Fairness. In *2020 ICML Workshop on Human Interpretability in Machine Learning*, 1-6.

Katznelson, G. & Gerke, S. (2021). The need for health AI ethics in medical school education. *Advances in health sciences education: theory and practice,* 26(4), 1447-1458.

Kellmeyer, P. (2019). Artificial Intelligence in Basic and Clinical Neuroscience: Opportunities and Ethical Challenges *Neuroforum,* 25(4), 241-250.

Kitchenham, B. & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Keele University and Durham University Joint Report.* https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

Larrazabal, A. J, Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America,* 117(23). 12592-12594.

Lee, E. E., Torous, J., Choudhury, M. de, Depp, C. A., Graham, S. A., Kim, H.-C., Paulus, M. P., Krystal, J. H. & Jeste, D. V. (2021). Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological Psychiatry, Cognitive Neuroscience and Neuroimaging,* 6(9). *856-864.*

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M. & Hardt, M. (2018). Delayed Impact of Fair Machine Learning. *ArXiv Preprint* arXiv:1803.04383. https://arxiv.org/abs/1803.04383

Marcinkowski, F. & Starke, C. (2019). Wann ist Künstliche Intelligenz (un-)fair? In J. Hofmann, N. Kersting, C. Ritzi & W. J. Schünemann (Hrsg.). *Politik in der digitalen Gesellschaft: Zentrale Problemfelder und Forschungsperspektiven.* (pp. 269-288) Bielefeld: transcript.

Markus, A. F., Kors, J. A. & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics,* (113), 103655.

Martin Luther King Jr. (1966). National Convention for Medical Committee for Human Rights. Chicago. https://www.goodreads.com/quotes/106932-of-all-the-forms-of-inequality-injustice-in-health-care

Martín Noguerol, T., Paulano-Godino, F., Martín-Valdivia, M. T., Menias, C. O. & Luna, A. (2019). Strengths, Weaknesses, Opportunities, and Threats Analysis of Artificial Intelligence and Machine Learning Applications in Radiology. *Journal of the American College of Radiology: JACR,* 16(9) 1239–1247.

McCradden, M. D., Joshi, S., Anderson, J. A., Mazwi, M., Goldenberg, A. & Zlotnik Shaul, R. (2020). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association: JAMIA,* 27(12), 2024-2027.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ArXiv Preprint* arXiv:1908.09635. http://arxiv.org/pdf/1908.09635v2

Miguel, I. de, Sanz, B. & Lazcoz, G. (2020). Machine learning in the EU health care context: exploring the ethical, legal and social issues. *Information, Communication & Society,* 23(8), 1139-1153.

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M. & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social science & medicine (1982),* (260), 113172.

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J. Malhotra, N., Cai, J. C., Malhotra, N., Lui, V. & Gibson, J. (2021). Artificial intelligence for good health: a scoping review of the ethics literature. *BMC medical ethics,* 22(1), 14.

Noor, P. (2020). Can we trust AI not to further embed racial bias and prejudice? *BMJ 2020* (363).

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science,* 366(6464), 447-453.

Panch, T. & Mattie, H. & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of global health,* 9(2), 10318.

Pandl, K. D., Feiland, F., Thiebes, S. & Sunyaev, A. (2021). Trustworthy machine learning for health care. In *CHIL '21: Proceedings of the Conference on Health, Inference, and Learning,* New York, 47-57.

Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M. & Sunyaev, A. (2020). On the Convergence of Artificial Intelligence and Distributed Ledger Technology: A Scoping Review and Future Research Agenda. *IEEE Access* (8), 57075-57095.

Parikh, R. B., Teeple, S. & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. *JAMA,* 322(24), 2377-2378.

Reimbursement Institute (2021). Behandlungspfad. https://reimbursement.institute/glossar/behandlungspfad/

Röösli, E., Rice, B. & Hernandez-Boussard, T. (2021). Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. *Journal of the American Medical Informatics Association : JAMIA,* 28(1), 190-192.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. (2020). CheXclusion: Fairness gaps in deep chest X-ray classifiers. *ArXiv Preprint arXiv:2003.00827.* https://arxiv.org/pdf/2003.00827

Sonntag, D. (2019). Künstliche Intelligenz in der Medizin – Holzweg oder Heilversprechen? *HNO,* 67(5), 343-349.

Srivastava, B. & Rossi, F. (2019). Rating AI systems for bias to promote trustable applications. *IBM Journal of Research and Development,* 63(4/5), 5:1-5:9.

Starke, G., Clercq, E. de & Elger, B. S. (2021). Towards a pragmatist dealing with algorithmic bias in medical machine learning. In *Medicine, Health Care, and Philosophy,* 1-9.

Straw, I. (2020). The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future. *Artificial intelligence in medicine* (110), 101965.

Sun, T. Y., Walk, O. J. B. D. IV, Chen, J. L., Nieva, H. R. & Elhadad, N. (2020). Exploring Gender Disparities in Time to Diagnosis. *ArXiv Preprint arXiv:2011.06100.* https://arxiv.org/abs/2011.06100

Tamayo-Sarver, J. H., Hinze, S. W., Cydulka, R. K. & Baker, D. W. (2003). Racial and ethnic disparities in emergency department analgesic prescription. *American journal of public health,* 93(12), 2067-2073.

Tat, E., Bhatt, D. L. & Rabbat, M. G. (2020). Addressing bias: artificial intelligence in cardiovascular medicine. *The Lancet Digital Health,* 2(12), e635-e636.

Thiebes, S., Lins, S. & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets (31)*, 447-464.

Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2021). Responsible AI for Digital Health: a Synthesis and a Research Agenda. *Information Systems Frontiers*, 1-19.

Weng, W.-H., Deaton, J., Natarajan, V., Elsayed, G. F. & Liu, Y. (2020). Addressing the Real-world Class Imbalance Problem in Dermatology. *ArXiv Preprint* arXiv:2010.04308. https://arxiv.org/abs/2010.04308

Zhou, N., Zhang, Z., Chen, J. & Singhal, H. (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. *ArXiv Preprint arXiv:2105.06558*. https://arxiv.org/abs/2105.06558

Anhang

Autor(en)	**Titel**	**Daten Problem**			**System Problem**		
		Erhebung	**Menge**	**Qualität**	**Algorithmus**	**Modell**	**Schnittstellen Mensch-KI**
Abassi-Sureshjani et al. (2020)	Risk of Training Diagnostic Algorithms on Data with Demographic Bias				x		
Afrose et al. (2021)	Subpopulation-specific Machine Learning Prognosis for Underrepresented Patients with Double Prioritized Bias Correction					x	
Ahmad et al. (2020)	Fairness in Machine Learning for Healthcare					x	
Ahsen et al. (2019)	When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis	x			x		
Bissoto et al. (2019)	(De) Constructing Bias on Skin Lesion Datasets	x		x			
Bissoto et al. (2020)	Debiasing Skin Lesion Datasets and Models? Not So Fast			x		x	
Brault & Saxena (2020)	For a critical appraisal of artificial intelligence in healthcare: the problem of bias in mHealth		x	x			
Burlina et al. (2021)	Addressing Artificial Intelligence Bias in Retinal Disease Diagnostics		x		x		
Chen et al. (2019)	Can AI Help Reduce Disparities in General Medical and Mental Health Care?					x	
Chester et al. (2020)	Balancing Utility and Fairness against Privacy in Medical Data					x	
Gade et al. (2020)	Explainable AI in Industry: Practical Challenges and Lessons Learned					x	x
Hague (2019)	Benefits, Pitfalls, and Potential Bias in Health Care AI						x
Hee (2017)	Is data quality enough for a clinical decision?: Apply machine learning and avoid bias	x		x			
Ienca & Ignatiadis (2020)	Artificial Intelligence in Clinical Neuroscience: Methodological and Ethical Challenges	x			x	x	
Jabbari et al. (2020) aus: Ahmad et al. (2020)	An Empirical Study of the Trade-Offs Between Interpretability and Fairness					x	
Katznelson & Gerke (2021)	The need for health AI ethics in medical school education						x
Kellmeyer (2019)	Artificial Intelligence in Basic and Clinical Neuroscience: Opportunities and Ethical Challenges						x

Larrazabal et al. (2020)	Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis	x					
Lee et al. (2021)	Artificial Intelligence for Mental Healthcare: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom	x					
Liu et al. (2020) aus: Ahmad et al. (2020)	Delayed Impact of Fair Machine Learning					x	
Martín Noguerol et al. (2019)	Strengths, Weaknesses, Opportunities and Threats Analysis of Artificial Intelligence and Machine Learning Applications in Radiology		x	x	x		
McCradden et al. (2020)	Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning						x
Miguel et al. (2020)	Machine learning in the EU health care context: exploring the ethical, legal and social issues						x
Murphy et al. (2021)	Artificial intelligence for good health: a scoping review of the ethics literature						x
Noor (2020)	Can we trust AI not to further embed racial bias and prejudice?						x
Obermeyer et al. (2019) aus: Evans & Mathews (2019)	Dissecting racial bias in an algorithm used to manage the health of populations				x	x	x
Panch et al. (2019)	Artificial intelligence and algorithmic bias: implications for health systems				x		
Röösli et al. (2021)	Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19						x
Seyyed-Kalantari et al. (2020)	CheXclusion: Fairness gaps in deep chest X-ray classifiers	x	x				
Srivastava & Rossi (2019)	Rating AI Systems for Bias to Promote Trustable Applications					x	
Starke et al. (2021)	Towards a pragmatist dealing with algorithmic bias in medical machine learning					x	x
Straw (2020)	The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future						x
Sun et al. (2020)	Exploring Gender Disparities in Time to Diagnosis					x	
Tat et al. (2020)	Addressing bias: artificial intelligence in cardiovascular medicine						x
Weng et al. (2020)	Addressing the Real-world Class Imbalance Problem in Dermatology				x	x	

Tabelle 3. Übersicht der Paper und deren Klassifizierung, in Anlehnung an Pandl et al. (2020)

Drivers and Inhibitors for the Adoption of Privacy Preserving Machine Learning in Organizations

Critical Information Infrastructures, Winter Term 21/22

Tobias Böck
Master Student
Karlsruhe Institute of Technology
tobias.boeck@student.kit.edu

Daniel Fischer
Master Student
Karlsruhe Institute of Technology
daniel.fischer@student.kit.edu

Amal Labbouz
Master Student
Karlsruhe Institute of Technology
amal.labbouz@student.kit.edu

Leon Sander
Master Student
Karlsruhe Institute of Technology
Leon.sander@protonmail.com

Abstract

Background: *Ever-improving machine learning techniques are leading to its widespread use, bringing benefits to organizations deploying them. The technology has now reached all sectors, leading to mainstream use. Simultaneously, a wide discussion about information security and privacy in machine learning developed, expressing concerns about data access, data storage and data usage. To address these concerns, the field of privacy preserving machine learning (PPML) emerged in recent years.*

Objective: *Since this research field is still in its early stages, scientific PPML contributions to date have mainly focused on technical aspects and have hardly looked at its implementation and diffusion within organizations. This paper will provide insights to explore the opportunities and risks of using PPML from an organizational perspective.*

Methods: *This work is based on an explorative and qualitative research approach in the form of semi-structured interviews with experts.*

Results: *We conducted interviews with 5 experts from the industry.We identified 16 factors, 8 of which drive adoption and 8 of which inhibit it. These factors were divided into technical, organizational, and legal. It can be shown that PPML solutions are not yet widely diffused in the market and that policy makers are lacking awareness about the technology. Increased computational costs, performance losses, and financial uncertainty are major inhibitors of PPML. In contrast, security and privacy guarantees, as well as privacy-compliant collaboration, are important drivers of technology.*

Conclusion: *We provided an insight into the current state of developments in research, politics and industry. The technical factors are highly complex, but are already receiving attention in research. Organizational aspects are currently neglected, while a heterogeneous legal framework irritates organizations.*

Keywords: Privacy Preserving Machine Learning, Organizational Adoption, Federated Learning, Differential Privacy, Secure Multiparty Computation, Homomorphic Encryption

Introduction

Have you been concerned about your information privacy when using digital solutions?

Have you then taken precautions to protect your data and information?

If you are honest, you will answer "yes" to the first question and "no" to the second. Don't worry, we didn't use your data to make this assumption. And besides, the authors of this seminar paper and most people in society would answer similarly. The reason given for this cognitive dissonance is the so-called "privacy paradox" (Norberg et al.,2007), which describes a situation where individuals claim to care about their information privacy while acting contrary to this attitude (Norberg et al., 2007; Spiekermann et al., 2001).

However, due to omnipresent data generation, storage and analytics, privacy is no longer just a personal matter individuals have to take care of, but a concern for security in general(Heurix et al., 2015; Altman et al., 2018). Ongoing incidents at large Internet-centered organizations like Meta, where the case of Cambridge Analytica or the current allegations of former employee Frances Haugen (European Parliamentl, 2021) raise questions about data use and data availability to third parties, disrupt the public's trust in corporations and whole sectors. Simultaneously, the field of machine learning (ML) made rapid progress and enables the use of data-driven algorithms in an increasing number of fields. These include several critical (information) infrastructures (CII) like finance, telecommunications, or healthcare (Mercier et al., 2021). As critical infrastructures continue to digitize, they are thus also affected by loss of information security, as witnessed by the 2014 hacking attack on JP Morgan Chase (Silver-Greenberg et al., 2014), while in the healthcare sector, there are still reservations about the use of machine learning for privacy concerns (Pandl et al., 2021; Panch et al., 2019). With the further diffusion of machine learning solutions into CII, privacy protection is a growing issue, since data used to train and deploy the algorithms might include sensitive information, such as health, gender or ethnicity (Mercier et al., 2021; Altman et al., 2018). For users to trust these ML systems sufficiently, they need evidence about data access, data usage, and data protection (Heurix et al., 2015; Brundage et al., 2020). Therefore, the responsibility to ensure information security is increasingly shifting from a personal sphere to organizations, as can be seen from ever more far-reaching legal frameworks (Rubinstein, 2011). Organizations need to address this development in future, to comply with regulatory requirements and unlock untapped potential in sectors such as healthcare.

Consequently, the questions above could be paraphrased and asked to the providers and operators of ML models:

Do you inform your users sufficiently about the further usage of their data?

Have you taken precautions to protect your users' data and privacy?

This time assumptions about the answers cannot be made because, unlike in the individual sphere, the deployment of privacy enhancing technologies (PET) from the organizational side has not yet been sufficiently researched (Harborth et al., 2018). Nevertheless, with the emerging field of privacy preserving machine learning (PPML), a promising technique is evolving that could solve the tension between information security and machine learning. PPML addresses problems arising from the transmission and use of sensitive information by using various approaches such as federated learning (FL), secure multi-party computing (SMPC), differential privacy (DP) and homomorphic encryption (HE) (Mercier et al., 2021). However, being a subfield of PETs, it is still not known what drivers and inhibitors of using PPML look like from an organizational perspective. Despite the advancement and growing body of research in the field of PPML, publications to date have mainly focused on technical aspects, including categorization, design and performance (Al-Rubaie and Chang, 2019; Morsbach et al., 2021; Ali et al., 2020). And while incentives and barriers for organizations have been exhaustively surveyed for Artificial Intelligence (Chui and Malhotra, 2018; Pandl et al., 2021), research on the diffusion of PPML is missing. Some work identified drivers and inhibitors for implementing PETs in an organizational context (Harborth et al., 2018) but does not explicitly address the area of ML. As a result, we still lack an understanding of what drives or inhibits organizations from adopting PPML. To address this gap, we ask the following research question: *What are drivers and inhibitors of practical implementation of privacy preserving machine learning approaches in organizations?*

We anser this research question by conducting semi-structured interviews with experts from the industry and embed their assessment in existing research on PPML and privacyaspirations. Our study yields 8

factors that drive the use of privacy preserving machine learning and 8 factors that inhibit its use for organizations. In addition, we were able to gain insights into the current status of the field in terms of research, policy and industry development. The identified factors were categorized as technical, organisational and legal in order to distinguish the individual drivers and inhibitors in the fields. With our work we contribute to existing research by bridging the gap between the previous mostly theoretical and technical research in this field with the practical deployment by providing an informed overview and understanding of the current drivers and inhibitors of implementing PPML in organizations. Our results can contribute to future research and approaches to accelerate the diffusion of PPML in the industry

Background

In this section we provide a brief overview of PPML techniques, research regarding the application of PPML in CII and research regarding the adoption of novel technologies in organizations. Table 1 provides a selection of PPML application areas and its characteristics in different contexts.

PPML Overview

When an ML model is trained on sensitive (personal) information, privacy risks include unacceptable access to raw training data, unacceptable inference from a trained model and unacceptable access to the model itself (Brundage et al., 2020). On the one hand, privacy risks thus arise from database related risks such as access control issues or breaches in centralized databases holding unencrypted sensitive data. Historically, anonymization was deemed a sufficient approach to protect sensitive data. However, methods to invert this process have led to the consensus that anonymization is often an insufficient data protection approach (Saranya et al., 2015). A considerable body of literature also examines sophisticated security and privacy issues inherent in the ML domain (Rigaki and Garcia, 2020). This research focuses on (malicious) exploitation of unintended information leakage from centralized and decentralized ML systems. Vulnerabilities include:

- Model reconstruction attacks (Milli et al., 2019) allow the reconstruction of rare private data through attacks on a model's latent space
- Model inversion attacks (Fredrikson et al., 2015) allow the reconstruction of training data utilizing public feature vectors
- Membership inference attacks (Shokri et al., 2017) allow to infer whether specific samples were part of a models training dataset

These diverse risks have motivated the development of various techniques subsumed under the term "privacy-preserving machine learning" (PPML), which aim to provide guarantees regarding data access, data usage and/or data protection to individuals and organizations. The individual techniques in the PPML toolbox each address a certain aspect of privacy, and each comes with specific limitations and costs. In practice, therefore, PPML systems often rely on a combination of techniques. Depending on the perspective, researchers (Brundage et al., 2020; Mercier et al., 2021) usually ascribe the following techniques to the PPML toolbox.

Federated Learning (FL) addresses privacy issues that arise from centralized data storage. In FL-systems, a central server orchestrates the learning on decentralized databases. The coordinator distributes an initial model, which is further trained on local data. Collaborating learners then only report their trained local model updates and subsequently receive a globally updated model in the next iteration. There exist open-source frameworks such as Tensorflow Federated for this purpose (TFFederated). However, as the transmitted local updates could leak potentially sensitive information, additional measure are necessary to ensure privacy (Kairouz et al., 2019).

Differential Privacy (DP) ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis (Dwork, 2008) and thus provides measurable guarantees

of privacy that help to mitigate the risk of exposing sensitive training data in machine learning (TFPrivacy). In the context of CII, for example, if a database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of that individual's data in the database will not significantly affect her chance of receiving coverage (Dwork, 2008). In practice, differential privacy can be achieved by clipping gradients in the stochastic gradient descent optimization method and/or by adding a controlled amount of statistical noise to either data or model updates to obscure data contributions of individual data points. The open-source framework Tensorflow Privacy implements differentially private stochastic gradient descent (DP-SGD) (TFPrivacy). However, this process will inevitably impact the performance of models negatively and therefore constitutes a tradeoff between privacy level and model performance (Morsbach et al., 2021).

Secure Multiparty Computation (SMPC) relies on a set of cryptographic protocols that enable multiple parties to jointly compute the output of some function while keeping their inputs private (Cramer and Damgård, 2005). An exemplary function is Private Set Intersection (PSI), where two parties can jointly compute the intersection of their datasets. Here too, open-source frameworks exist (Multiparty.org).

Homomorphic Encryption (HE) is a form of encryption which allows specific types of computations to be carried out on ciphertexts and generate an encrypted result which, when decrypted, matches the result of operations performed on cyphertext (Yi et al., 2014). For HE, open source implementations exist (Chillotti et al., August 2016) as well. On the downside, both SMPC and HE are subject to a high computation and communication overhead, which poses a problem for the computationally expensive training of ML models.

Furthermore Brundage et al. (2020) include in the PPML toolbox Secure Enclaves (trusted execution environments), which they define as "a set of software and hardware features that together provide an isolated execution environment that enables a set of strong guarantees regarding security for applications running inside the enclave."

Our research indicates that the ambiguity of the term "privacy" can lead to misconceptions when talking about PPML. Without claim for completeness, in Table 1 we therefore synthesize the following overview of how PPML can be applied, which data and data sources it relies on as well as the privacy notion and purpose it serves.

	General B2B PPML	**B2B PPML on Personal Data**	**B2C PPML for End-User Privacy**
Underlying Data	Non-personal data from various domains: IoT, CII, …	Data on individuals (financial records, telecommunication, …)	Data produced by individuals in daily life (searches, typing, …)
Data Sources	Organizational databases, IoT devices	Organizational databases	End-user device
Privacy Notion	Security and residency of input data into a compute operation remain private, and the only thing visible is the result of that operation.	Security and residency of personal and sensitive input data remain private, and the only thing visible is the result of that operation.	Retaining sensitive data on your own device "[…] the right to have some control over how your personal information is used."
Purpose	Building (better) models while protecting corporate secrets and/or complying with regulations	Building models while protecting corporate secrets and complying with regulations, gaining a 360° customer view	Protecting personal sensitive data by keeping it on end-user device
Example	Joint model training in IoT / predictive maintenance scenario	Joint credit score calculation	Gboard,Apple FaceID

Table 1. Relation between PPML Application Fields and Characteristics

Related Research

Research on PPML is advancing rapidly, however it usually takes a technical perspective. Queries in the Google scholar, Web of science and Scopus databases with a combination of the keyword "privacy-preserving machine learning" with "critical information infrastructures" or "critical infrastructure" and ranking by relevance reveal one result applicable to our scope: Mercier et al. (2021) benchmark methods and open-source frameworks to provide a first overview of the applicability of PPML techniques to the time series domain, due to its important role in various critical infrastructure settings. They successfully apply DP and FL to various architectures and datasets, highlight the importance of hyperparameter selection in DP, and confirm drawbacks of HE due to computational effort.

Similarly, the identification of factors influencing the adoption of novel technologies in organizations is an active field of research. Pandl et al. (2021), for example, investigate the adoption of Artificial Intelligence as a Service (AIaaS) in practice and identify 12 drivers and 12 inhibitors. Harborth et al. (2018) research incentives and barriers regarding the implementation of Privacy Enhancing Technologies in a business context. From insights gained in semi-structured interviews, they derive a taxonomy with the three principal categories technical optimization, business model & public perception. They conclude that their findings are too heterogeneous to paint a clear picture

Method

This work is based on an explorative and qualitative research approach in the form of semi-structured interviews with experts. Semi-structured interviews are characterized by several advantages. These include increased reliability, due to the standardization of the questions, as well as increased validity, as it is ensured that the criteria are covered systematically and completely compared to an unstructured narrative interview (Segal et al., 2006, p. 125). Semi-structured interviews, compared to surveys or structured interviews, do not limit the exploratory investigation for obtaining unknown and relevant information and is therefore utilized.

Data Gathering

Interview partners were identified and contacted through web research in the critical infrastructure sector. To gain extensive insight, an extra focus was set to get an interview with experts in every major subject of PPML. The subjects covered are decentralized learning, differential privacy, encryption and secure enclaves. More information on the interviewees in Table 2. In developing the interview guide, a framework according to (Kallio et al., 2016) was applied. The framework is divided into 5 phases.

1. Identify the Prerequisites for the Use of a Semi-structured Interview

In essence the first step is about evaluating if a semi structure interview is the appropriate method for data collection in relation to our research question. Based on our goals to gain real organizational insights, an approach which does not restrict the ability for explorative investigation while still maintaining some form of structure and systematically covering specific questions is most suitable. Both essential points are give by the nature of semistructured interviews.

2. Retrieving and Utilizing Previous Knowledge

We did an extensive research in the literature to understand the topic at hand aswell as to understand the elaborations of the interviewees and to be able to dig deeper when necessary to expand the flow of conversation and gain more insight in relevant points.

3. Formulate a Preliminary Semi-structured Interview Guide

Based on the knowledge from step two and our goals, we created a first formulation of the interview guide.

4. Pilot Testing

The Pilot Testing stage was a reiterative process to confirm the coverage and relevance in relation to our research question and goals. Internal Testing and expert assessment by consultation of our supervisor led us to the final version after three iterations.

5. Presenting the Complete Semi-structured Interview Guide

For further development, testing or comparison, the final semi-structured interview guide can be found in the appendix A.6.

ID	Job Position	Experience in Privacy Field	Sector
A.1	COO	> 5 years	Software
A.2	Researcher	> 5 years	Software
A.3	CEO	> 5 years	Software
A.4	Project Manager	> 10 years	SMPC & SE development
A.5	Head of business development	> 5 years	SMPC development

Table 2. Information on Interviewees

Data Analysis

To analyze our data, we used established coding paradigms. Based on the amount of interviews we conducted, we decided that two iterations of coding are sufficient to extract the information necessary. Staring with open coding in the first iteration to conclude the second iteration with selective coding for a higher data-analysis level. Since our goals are of exploratory nature we used the inductive coding approach and derived cods from the data, starting without preconceived notions of what the codes should be (Linneberg and Korsgaard, 2019, p. 12).

Open Coding as the first level of coding, we summarized statements which could span over multiple sentences to its essence with the aim of identifying distinct concepts and themes for categorization (Williams and Moser, 2019, p. 48).

By analyzing the codes and orienting at our goals we defined nine different themes for categorization: current state in industry, current state in the policy, current state in research, technical inhibitors, economic inhibitors, legal inhibitors, technical drivers, economic drivers, legal drivers.

Selective Coding enables the selection and integration of the data from previous steps into "cohesive and meaning-filled expressions" (Williams and Moser, 2019, p. 52). The meaning-filled expressions represent the final codes as a more abstract conceptualization level.

Results

This chapter presents the results of the semi-structured interviews. 55 companies and experts were requested for an interview and 5 interviews could be conducted. 4 experts work in companies that mainly base their business models on the development of privacyfriendly machine learning models. A large part of the refusals to our requests were justified with the statement that the company does not have any competences in this complex of topics.

In the following, the codes and topics from the deductive procedure are presented. The themes were developed from the codes with a focus on achieving our research goals. For improved comprehensibility, the codes are presented by theme. This chapter references the transcripts of the interviews. These can be received on request.

Current Status

These themes allow to shed light on the status of the technology, its use in practice and in law and enables the achievement of another goal of the research. Table 3 shows the identified codes from interviews related to the current status.

Theme	Code
Current status in industry	Practical implementation is still in an early phase.
Current status in policy	Technological progress is outpacing legislation.
Current status in R&D	Substantial funds are available.

Table 3. Codes in the Context of Current Status

Industry

Practical implementation of PPML is still in an initial phase. PPML solutions have not yet diffused widely in the market. In the experience of one expert, consultative conversations take place with a focus on digitalisation and not with a target on privacy (see Appendix A.1, available upon request). Another expert reports that the push for innovation in privacy preserving technologies comes from US companies, while German companies specialise in legal compliance (see Appendix A.3, available upon request).

Politics

Technological progress is outpacing legislation. Along with the lack of awareness about the developments in this field, policy makers are too slow to act with appropriate legislation. In the view of one expert, policy-makers should not intervene too much, as they usually do not understand what is happening technically (see Appendix A.3, available upon request). Another expert criticises in this context that regulation at EU level takes a long time due to the influence of lobbyists and the factor that large companies are constantly suing against new regulations (see Appendix A.1, available upon request).

R&D

Substantial funding is available. A lot is being done in research on this topic, in on experts' opinion. According to him (see Appendix A.5, available upon request) the Defense Advanced Research Projects Agency (DARPA) has invested around two hundred million USD into PPML research, and venture capital has allocated the same amount to the sector as well.

Inhibitors of PPML Adoption

These themes enables the identification of technical, economic/organizational and legal inhibitors, which corresponds to the goals that have been set. A inhibitor is defined as a factor that is disadvantageous to the adoption and implementation of PPML techniques in organizations. Table 4 shows the identified codes from interviews related to the context of PPML inhibitors.

Technical Inhibitors

Increased computational costs and performance losses. Multiple interview partners (see Appendix A.2, A.4, available upon request) view the increased computation and communication overhead (SMPC, HE) and performance losses (DP) as a barrier to implement PPML techniques in use-cases that currently work without. However, the magnitude of the overhead depends strongly on the specific technique (see Appendix A.4, available upon request).

Technical complexity and model parametrization decisions. The choice of appropriate parameters for modelling needs to be considered from different perspectives. An example of the discussion on the selection of a suitable parameter is the setting of epsilon in differential privacy models, which determines how much

specificity a data collector is willing to sacrifice in order to protect the privacy of its users. According to the experts, if this parameter is chosen too weakly, anonymisation will not take place even if such an approach is followed (see Appendix A.2, available upon request). At the same time, it must be ensured that the models are capable of predicting reliably. This trade off between the choice of an appropriate parameter and the quality of the models is a challenge for companies.

Theme	**Code**
Technical inhibitors	● Increased computational costs and performance losses. ● Technical complexity and parametrization decisions.
Economic and organizational inhibitors	● Technical competences and managerial responsibilities. ● Financial uncertainties. ● Firms subjection to market forces and competitive dynamics.
Legal inhibitors	● Europe's heterogeneous legal landscape. ● Ambiguous privacy notion. ● Privacy level guarantee.

Table 4. Codes in the Context of PPML Inhibitors

Economic and Organizational Inhibitors

Technical competences and managerial responsibilities. The lack of experts with knowledge and skills in the intersection of ML, cryptography and privacy is one of the most serious factors inhibiting the widespread diffusion of PPML (see Appendix A.2, A.3, available upon request). Furthermore, one expert (see Appendix A.5, available upon request) states that in many organizations, it is unclear who's management responsibility (CTO, CIO, CDO,etc.) the implementation of PPML solutions falls into.

Financial conflicts and uncertainties. If the implementation of PPML solutions is considered an investment, two problems arise. For business models that rely on the sale of data, and are thus the most privacy violating, there exists little financial incentive to implement PPML techniques (see Appendix A.3, available upon request). Furthermore, one interviewee (see Appendix A.5, available upon request) states that there exists great uncertainty on how to measure the return on investments into PET and PPML. This circumstance contributes to the lack of incentives for firms to be an early adopter of PPML techniques.

Firms subjection to market forces and competitive dynamics. Businesses act according to market signals, but end-users often act and consume contrary to their own privacy interests (privacy paradox). Therefore, according to one interviewee (see Appendix A.3, available upon request) the lack of market pressure does not create incentives for firms to adjust their products. Furthermore, products marketed as privacy preserving while providing weak or no guarantees, may distort consumer perception (see Appendix A.3, available upon request). In B2B settings, unsolicited implementation of PPML techniques is seen as a competitive disadvantage, due to worse model performance and only used when there is no other way (see Appendix A.2, available upon request)

Legal Inhibitors

Europe's heterogeneous legal landscape According to one Interviewee (see Appendix A.4, available upon request), in Europe there exists an overlap of legal frameworks: the GDPR and ePrivacy laws interact with national laws, (e.g. national statistic laws). Therefore, the feasibility of privacy preserving cross-border data analysis projects is often restricted by legal factors, not technical ones.

Politics must generate awareness of the issue. According to one expert, the definition of privacy differs between the legal and the technical side (see Appendix A.3, available upon request). Another interviewee identifies this aspect as a content-related problem that cannot be solved without a deeper awareness of the topic in politics or further explanatory texts in law (see Appendix A.1, available upon request). Unless this happens, there will be a further discrepancy between the legal and technical understanding of privacy.

Proof that anonymisation parameters provide protection. Along with the problem of the right choice of a parameter that guarantees the anonymisation of data is the proof of anonymisation towards third parties.

Since there exist no uniform understanding of the definition of privacy from a legal and technical perspective, this means that legal proof is not even possible. The heterogeneous legal landscape in Europe hinders the implementation of PPML solutions according to an expert (see Appendix A.4, available upon request), which would simultaneously have an aggravating effect on the control of PPML approaches according to certain criteria, if any existed.

Drivers of PPML Adoption

This theme enables the identification of technical, legal and economic drivers, which corresponds to the goals that have been set. Drivers are factors that provide incentives to organizations to adopt PPML techniques. Table 5 fAm codes from interviews related to the context of PPML drivers.

Theme	**Code**
Technical	● Technological development. ● Safety and privacy guarantees.
Economic and Organizational	● Inter-organizational data leveraging. ● Competitive advantage. ● Versatility of PPML techniques. ● Data protection compliant cooperation.
Legal	● Data residency legislation. ● Enforcement of existing law.

Table 5. Codes in the Context of PPML Drivers

Technical Drivers

Technological development. One Interviewee (see Appendix A.4, available upon request) mentions that both the general technological progress (e.g. decreasing costs of computation) as well as the development of specific technologies (e.g. Intel SGX Secure Enclave) is driving the adoption of PPML solutions.

Safety and privacy guarantees. The use of PPML offers security guarantees that enable use cases that rely on sensitive data. One expert mentions the example of border control, where data could be shared between different nations to create models for assessing whether a passenger should be checked (see Appendix A.1, available upon request). Furthermore, with the correct application of a PPML solution, attacks such as membership interference or model inversion are no longer possible.

Economic and Organizational Drivers

Inter-organizational data leveraging. Multiple interviewees (see Appendix A.4, A.5, available upon request) state that financial incentives of novel business use-cases, enabled by PPML, are a strong drivers for the application of specific techniques (FL, SMPC, SE). One interviewee (see Appendix A.5, available upon request) mentions that he sees collaborative data development to create a "360° customer view" as the strongest driver in the PPML field. Exemplary is a SMPC system between a bank and a telecommunication firm to build even better customer profiles. The interviewee emphasizes that customer consent is fundamental to these business models and that PPML enables a much more diverse perspective on consent ("Sharing your data without sharing it").

Competitive advantage. In some use-cases, firms use PET/PPML techniques to differentiate their products from competitors (see Appendix A.5, available upon request). A famous examples is the differentiation though privacy by Apple vs. Andoid, where customers are consequentially willing to accept a higher price because they percieve Apple as more privacy preserving. One interviewee (see Appendix A.3, available upon request), however, mentions the risk of greenwashing in this context, due to the fact, that each techniques addresses such a limited aspect of privacy.

Versatility of PPML techniques. The use of PPML approaches is not limited to the application in critical infrastructures. It is useful in all areas where sensitive data is used to train models, corporate secrets need to be protected or data is provided for collaborative working. In the application, data protection is a

multifaceted context that relates not only to the domain but also to the application. In the PPML landscape, it is essentially about protecting and preserving the privacy of input data while facilitating the output of an operation with those datasets.

PPML enables data protection-compliant cooperation between companies. PPML techniques can facilitate collaboration between organisations in different sectors. One expert cites as an example the use of PPML techniques to produce improved census and movement statistics by combining data from mobile network operators, while preserving privacy (see Appendix A.4). A further example is the already mentioned use case of cooperation at border controls. Another possibility is the unsiloing of data. Unsiloing enables collaborative work on data, and by using PPML, privacy is respected when companies share data. It enables faster and more efficient discovery of exciting insights.

Legal Drivers

Data residency legislation. According to one interviewee (see Appendix A.5, available upon request) legal decisions mandating or encouraging data residency, such as Schrems II and Indonesian offshore cloud computing taxes, contribute to the adaption of PPML techniques.

Enforcement of existing law. In the opinion of one interviewee (see Appendix A.4, available upon request) cyber crime is hard to fight, because is spans multiple sectors and multiple legislative systems. Therefore PPML techniques offers to national law enforcement agencies the possibility to cooperate more effectively in crime prevention. Furthermore PPML can also be used by multiple organizations for fraud detection models, e.g. in a financial context (see Appendix A.4, A.5, available upon request).

Discussion

This chapter discusses the principal findings and identifies implications for research and practice. Subsequently, imitations are identified and guidance for future investigations is given.

Principal Findings

Based on the qualitative analysis of five expert interviews, we identified a total of 8 factors that drive the adaption of PPML and 8 factors that inhibit its adoption by organizations. Furthermore, our study gave us an insight into the assessment of the current status in industry, policy and research.
Our results suggest that there is some overlap in research on incentives and barriers to PET adoption (Harborth et al., 2018), but that many factors in PPML cover specific circumstances that do not arise while adopting other PETs. These include, above all, technical factors (e.g. accuracy and performance) that are naturally unique to this area, while organizational factors show many parallels (e.g. competitive advantage). On the whole, separate research is needed to understand the drivers and inhibitors of PPML adoption, since this field is more nuanced.
It is claimed that the diffusion of PPML is currently still in its early stages. This assertion is confirmed by the Gartner Hype Cycle, according to which known PPML approaches such as FL and DP are still in the innovation trigger phase and need an estimated 5-10 years to reach the plateau of productivity (Willemsen, 2021). This entails some drivers and inhibitors. Our results indicate that the biggest inhibitors currently lie in technical and organizational aspects. While the challenges in technology, above all performance loss, complexity and parametrization, have already been addressed and thematized in various literature (Morsbach et al., 2021; Ali et al., 2020), the identification of organizational and economical inhibitors in the context of PPML have not yet taken place to our knowledge. Nevertheless, some parallels to research on other fields can be pointed out here. The lack of competence in the area of PPML is an already exiting issue in deploying artificial intelligence as well as artificial intelligence as a service (AIaaS) (Pandl et al., 2021; Chui and Malhotra, 2018). In addition, organizations see both, the lack of measurability of the return of investment and the difficult market conditions as inhibitors, which appear both in the literature on PETs, where privacy enhancing solutions have to compete with conventional approaches (Harborth et al., 2018), and in Artificial Intelligence solutions, where high investment sums in innovative research are needed in prior to market entry (Chui and Malhotra, 2018). As an intersection between these two fields, this situation is exacerbated for PPML, creating a high risk for implementers. Combined with a low willingness to pay on the part of end customers, for example due to the privacy paradox (A.Norberg et al., 2007) introduced in

the beginning or the privacy calcalus (Dienlin and Metzger, 2016), current market incentives are not sufficient for the widespread deployment of PPML.
Nevertheless, looking at the drivers for adopting PPML, the most concise are equally those of organizational and economic factors. In general, we see that organizations understand the benefits of implementing PPML to their businesses. Especially the possibility of interorganizational data-leveraging is an opportunity that is specific to PPML and can open doors to new applications and business models. The importance of overcoming data silos has already been touched upon in various use cases in the literature (Panch et al., 2019; Pandl et al., 2021). This is particularly true in areas such as the healthcare sector. PPML could offer new solutions according to findings from literature and interviews, enabling new business models. This goes hand in hand with the finding that PPML is increasingly seen as a competitive advantage and thus offers first mover potential. While a similar conclusion has been reached in the literature on PETs, there is a warning, however, that with increasing diffusion these competitive advantages will become obsolete (Harborth et al., 2018). Looking at our specific scope of critical information infrastructures, particularly relevant drivers in the technical and legal areas become apparent here. Hence, increased security and protection of data is a common argument among the respondents, making safety and privacy guarantees as well as opportunities in the legal sphere a driver for deploying PPML.
Interestingly, the legal sphere, driven by policy, has a partially incoherent position. On the one hand, organizations feel that policy is currently too slow, contradictory, and lacking in knowledge in the field to provide a sufficient framework, and therefore see it as an obstacle. On the other hand, policy-makers are called upon to engage more with the issue in order to bridge the gap between technical and legal definitions and to establish legal certainty.
Overall, the willingness and motivation to implement PPML also strongly depends on the corporate context. Sector, use case, and the regulations therein are essential factors. Accordingly, the healthcare sector, for example, is seen as having a great deal of potential because of the data silos that have existed here to date. On the other hand, business models that specialize in sharing and selling data have not such an incentive to implement PPML.

Implications for Research and Practice

We have focused our study on the drivers and inhibitors of implementing PPML in organizations and, to the best of our knowledge, this opens up an area that has received little attention so far. Our results are supported by existing literature and partly by findings from other fields (e.g., drivers and inhibitors in the implementation of PETs or AI). Nevertheless, we also find results that require further consideration from the research community, both in their understanding and in the further step of mitigating inhibitors. Overall, the implementation of PPML in practice is complex and situational and therefore requires more in-depth understanding. We have made a start with the identified 16 factors that can be elaborated. For example, the identified technical challenges are very characteristic of PPML, but so far are not found in inhibitors to deploying traditional AI nor in the deployment of PETs. Here we still see a need for research. While the progress on the technical side is advancing and fundings are secured as one of our interview partners claimed, research in organizational factors seems to be insufficient. Future research can deal with this. Thus, new guides for research directions can be built on the basis of our findings.
For practitioners, our findings have also generated added value from which implications can be derived. The implementation of PPML involves many technical resources that need to be considered. Research is still growing rapidly and is just starting to take off. Providers of PPML must therefore quickly implement and incorporate findings from research. This applies, for example, to new best practices on accuracy and performance, as unimplemented new findings and advances could otherwise undermine the first-mover advantage. This is even more challenging as human capital in this field is already very limited. Organizations would therefore need to set up a well thought-out strategy to implement PPML approaches, even if they have previously deployed traditional machine learning successfully or consider using novel MLaaS approaches. This applies at least as much to market strategy. For users, privacy technologies are intransparent and hardly distinguishable in terms of quality due to lack of knowledge. Organizations need to communicate their approaches clearly and well in order to convince users and differentiate themselves from other market players who may use weaker privacy techniques. The added value must be comprehensible to users in order to make the establishment of PPML profitable. Especially practitioners in areas with data silos, where currently hardly any or insufficient ML methods are used, have this competitive advantage. Finally, we see another group of stakeholders that needs to be more involved in the topic and

it's diffusion. Legislators need to gain an understanding of the field and create a clear legal framework to remove hurdles or at least not create unnecessary challenges.

Limitations and Future Research

Alongside new findings, our work also presents some limitations. An entry into the plateau of productivity is only predicted for the next 5-10 years. Thus, organizations are still in the early phase of adoption of PPML. Statements about today's drivers and inhibitors could become rapidly obsolete with the further diffusion of the technology, as the wider dissemination of technologies could change the status quo (e.g. no need for in-house solutions due to offer of PPML models via AIaaS solutions).
Furthermore, this situation currently limits the sources of experience from the industry. We noticed this limitation in our work, as we were able to conduct fewer interviews than we had planned. Several companies we approached did not have any expertise in the field. Others, however, were very specialized in the topic and focused their whole business model around PPML. Most of the participating organizations consisted of start-ups that had focused on PPML since their inception and brought their knowledge from prior work in research. Therefore, the results primarily reflect a niche of very savvy implementers, but not yet a wide part of the industry, which is again due to the early phase of the technology's diffusion. Moreover, we couldn't further develop the existing interview guide to derive improved, nuanced insights. For future research, this may mean collecting more data in the form of interviews and possibly choosing an additional methodology (e.g., a survey or talks at trade fairs) in which a larger circle of organizations - including ones without a PPML-research background - can contribute their experience and knowledge. With this study, we identify and present new insights into the drivers and inhibitors of deploying PPML from an organizational perspective, opening the doors for further research in this area. As machine learning (and thus PPML) and its application areas are a broad field, future research can identify whether drivers and inhibitors behave differently in specific sectors (e.g. healthcare). In addition, the factors we have identified should be further assessed and their mode of action understood so that they can be minimised in the case of inhibitors and harnessed in the case of drivers. Lastly, future research can include the perspective of policymakers alongside companies to understand their ambivalent role

References

Ali, S., Irfan, M., Bomai, A., & Zhao, C. (2020). Towards Privacy-Preserving Deep Learning: Opportunities and Challenges.

Al-Rubaie, M., & Chang, J. M. (2019). Privacy-Preserving Machine Learning: Threats and Solutions. IEEE Security & Privacy, 17(2), 49–58. https://doi.org/10.1109/MSEC.2018.2888775

Altman, M., Wood, A., O'Brien, D. R., & Gasser, U. (2018). Practical approaches to big data privacy over time. International Data Privacy Law, 8(1), 29–51. https://doi.org/10.1093/idpl/ipx027

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims (arXiv:2004.07213). arXiv. http://arxiv.org/abs/2004.07213

Chillotti, I., Gama, N., Georgieva, M., & Izabachène, M. (2020). TFHE: Fast Fully Homomorphic Encryption Over the Torus. Journal of Cryptology, 33(1), 34–91. https://doi.org/10.1007/s00145-019-09319-x

Chui, M., & Malhotra, S. (2018). AI adoption advances, but foundational barriers remain, McKinsey. See: Https://Www. Mckinsey. Com/Featured-Insights/Artificial-Intelligence/Ai-Adoption-Advances-but-Foundational-Barriers-Remain.

Cramer, R., & Damgård, I. (2005). Multiparty computation, an introduction. In Contemporary cryptology (pp. 41–87). Springer.

De Cristofaro, E. (2021). A Critical Overview of Privacy in Machine Learning. IEEE Security & Privacy, 19(4), 19–27. https://doi.org/10.1109/MSEC.2021.3076443

Dienlin, T., & Metzger, M. J. (2016). An Extended Privacy Calculus Model for SNSs: Analyzing Self-Disclosure and Self-Withdrawal in a Representative U.S. Sample: THE EXTENDED PRIVACY CALCULUS MODEL FOR SNSs. Journal of Computer-Mediated Communication, 21(5), 368–383. https://doi.org/10.1111/jcc4.12163

Dwork, C. (2008). *Differential privacy: A survey of results. In international conference on theory and applications of models of computation 2008 (pp. 1-19)*. Springer, Berlin, Heidelberg.

European Parliamentl. (2021, November 5). Facebook whistleblower testifies in European Parliament | News European Parliament. https://www.europarl.europa.eu/news/en/headlines/society/20211028STO16120/facebook-whistleblower-testifies-in-european-parliament

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1322–1333. https://doi.org/10.1145/2810103.2813677

Harborth, D., Braun, M., Grosz, A., Pape, S., & Rannenberg, K. (2018). Anreize und Hemmnisse für die Implementierung von Privacy-Enhancing Technologies im Unternehmenskontext. https://doi.org/10.18420/SICHERHEIT2018_02

Hemphill, T. A., & Banerjee, S. (2021). Facebook and self-regulation: Efficacious proposals – Or 'smoke-and-mirrors'? Technology in Society, 67, 101797. https://doi.org/10.1016/j.techsoc.2021.101797

Hersen, M. (Ed.). (2006). Clinician's handbook of adult behavioral assessment. Elsevier Academic Press.

Heurix, J., Zimmermann, P., Neubauer, T., & Fenz, S. (2015). A taxonomy for privacy enhancing technologies. Computers & Security, 53, 1–17. https://doi.org/10.1016/j.cose.2015.05.002

Implement Differential Privacy with TensorFlow Privacy | Responsible AI Toolkit. (n.d.). TensorFlow. Retrieved June 2, 2022, from https://www.tensorflow.org/responsible_ai/privacy/tutorials/classification_privacy

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and Open Problems in Federated Learning (arXiv:1912.04977). arXiv. http://arxiv.org/abs/1912.04977

Kallio, H., Pietilä, A.-M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. Journal of Advanced Nursing, 72(12), 2954–2965. https://doi.org/10.1111/jan.13031

Khalil, S. (2014). Not everything that counts can be counted and not everything that can be counted counts. The Psychiatric Bulletin, 38(2), 86–86. https://doi.org/10.1192/pb.38.2.86b

Linneberg, M., & Korsgaard, S. (2019). Coding qualitative data: A synthesis guiding the novice. Qualitative Research Journal, 19(3), 259–270. https://doi.org/10.1108/QRJ-12-2018-0012

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2022). When Machine Learning Meets Privacy: A Survey and Outlook. ACM Computing Surveys, 54(2), 1–36. https://doi.org/10.1145/3436755

Mercier, D., Lucieri, A., Munir, M., Dengel, A., & Ahmed, S. (2021). Evaluating Privacy-Preserving Machine Learning in Critical Infrastructures: A Case Study on Time-Series Classification. IEEE Transactions on Industrial Informatics, 1–1. https://doi.org/10.1109/TII.2021.3124476

Milli, S., Schmidt, L., Dragan, A. D., & Hardt, M. (2019). Model Reconstruction from Model Explanations. Proceedings of the Conference on Fairness, Accountability, and Transparency, 1–9. https://doi.org/10.1145/3287560.3287562

Morsbach, F., Dehling, T., & Sunyaev, A. (2021). Architecture Matters: Investigating the Influence of Differential Privacy on Neural Network Design (arXiv:2111.14924). arXiv. http://arxiv.org/abs/2111.14924

Multiparty. (n.d.). GitHub. Retrieved June 2, 2022, from https://github.com/multiparty

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. Journal of Consumer Affairs, 41(1), 100–126. https://doi.org/10.1111/j.1745-6606.2006.00070.x

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. Npj Digital Medicine, 2(1), 77. https://doi.org/10.1038/s41746-019-0155-4

Pandl, K. D., Teigeler, H., Lins, S., Thiebes, S., & Sunyaev, A. (2021). Drivers and Inhibitors for Organizations' Intention to Adopt Artificial Intelligence as a Service. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.215

Rigaki, M., & Garcia, S. (2021). A Survey of Privacy Attacks in Machine Learning (arXiv:2007.07646). arXiv. http://arxiv.org/abs/2007.07646

Rubinstein, I. S. (n.d.). REGULATING PRIVACY BY DESIGN. 50.

Saranya, K., Premalatha, K., & Rajasekar, S. (2015). A survey on privacy preserving data mining. 2015 2nd International Conference on Electronics and Communication Systems (ICECS), 1740–1744.

Segal, D. L., Coolidge, F. L., O'Riley, A., & Heinz, B. A. (2006). Structured and Semistructured Interviews. In Clinician's Handbook of Adult Behavioral Assessment (pp. 121–144). Elsevier. https://doi.org/10.1016/B978-012343013-7/50007-0

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks against Machine Learning Models (arXiv:1610.05820). arXiv. http://arxiv.org/abs/1610.05820

Silver-Greenberg, J., Goldstein, M., & Perlroth, N. (n.d.). PMorgan Chase Hack Affects 76 Million Households. 3.

Spiekermann, S., Grossklags, J., & Berendt, B. (2001). E-privacy in 2nd generation E-commerce: Privacy preferences versus actual behavior. Proceedings of the 3rd ACM Conference on Electronic Commerce - EC '01, 38–47. https://doi.org/10.1145/501158.501163

TFFederated. TensorFlow Federated. (n.d.). TensorFlow. Retrieved June 2, 2022, from https://www.tensorflow.org/federated?hl=de

Willemsen. Hype cycle for privacy. (2021). Webseite, July 2021. URL https://www. gartner.com/en/documents/4003504-hype-cycle-for-privacy-2021.

Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. International Management Review, 15(1), 45–55.

Xie, M., Wang, J., & Chen, J. (2008). A Practical Parameterized Algorithm for the Individual Haplotyping Problem MLF (Vol. 4978, p. 444). https://doi.org/10.1007/978-3-540-79228-4_38

Yi, X., Paulet, R., & Bertino, E. (2014). Homomorphic Encryption. In X. Yi, R. Paulet, & E. Bertino, Homomorphic Encryption and Applications (pp. 27–46). Springer International Publishing. https://doi.org/10.1007/978-3-319-12229-8_2

Societal Perception of Security Threats in Vehicular Fog Computing

Critical Information Infrastructures, Winter Term 21/22

Yannick Erb
Master Student
Karlsruhe Institute of Technology
uqjwb@student.kit.edu

Özge Nur Subas
Master Student
Karlsruhe Institute of Technology
uwrtm@student.kit.edu

Anastasiia Zhyliak
Master Student
Karlsruhe Institute of Technology
upubw@student.kit.edu

Olga Zimmermann
Master Student
Karlsruhe Institute of Technology
utdnv@student.kit.edu

Abstract

Background: *With the increase of data generated by vehicles, currently employed ecosystems such as centralized cloud computing (CC) come to an end of their capabilities regarding, for example, real-time data processing and latency. Vehicular fog computing (VFC) is proposed as a potential solution to overcome this issue. However, the usage of VFC systems comes with security threats that may have lethal consequences. Literature has started to investigate these security threats from an expert point of view employing threat modeling techniques. Besides theoretical considerations, the adoption of novel technology is also influenced by potential users' perceptions of risks associated with it.*

Objective: *This seminar paper sets out to answer the question of "How does society perceive security threats of VFC?" by providing a ranked list of security threats that VFC may come with.*

Methods: *To shed light on the perception of security risks in VFC, we build on the threat model proposed by Klein et al. (2022) and conduct a ranking-type Delphi study among non-domain experts (n=24). This three-phased approach enables the identification of novel security threats from a non-domain expert point of view, the selection of the security threats perceived as most important, and their ranking.*

Results: *We present and discuss an ordered list of nine security threats that are perceived as most dangerous by the participants of the study. We also find that nearly all security threats and categories presented in the model of Klein et al. (2022) are also reported by our participants, except for the Repudiation category.*

Conclusion*: Our results indicate that privacy and safety-related security threats are considered the most dangerous. Although only reaching a low level of consensus (W=0.2322) in the ranking, participants are concerned about the privacy of their data that is collected, stored, and used for decision making in the VFC system and about threats that may result in system failure and potentially in harm to traffic participants. This work adds to the ongoing research on security threats in VFC by providing a non-expert point of view, which enables finding solutions to the security threats perceived as most dangerous to influence adoption positively.*

Keywords: vehicular fog computing, VFC, security threats, ranking-type Delphi study

Introduction

It is predicted that data traffic will increase because vehicles will send 1 to 10 exabytes of data per month to the cloud (Marshall & Cases, 2021). In total, this is an increase of at least 1000 times the current volume. It follows that vehicle designs are becoming increasingly software-centric and dependent on connectivity and cloud and edge computing capabilities. It is still unclear how this volume, and especially the sensitive data, can be handled, as the ecosystems (e.g., centralized cloud computing (CC) (Marshall & Cases, 2021)) currently in use would be overwhelmed by such a large amount of data. Such an ecosystem would, for example, have problems regarding maintaining the latency or with real-time processing of the data.

Fog computing (FC) represents one potential solution to these challenges (Iorga et al., 2018). It is a rather novel computing paradigm that has become increasingly popular in research. It is an enhancement of CC and brings computing closer to the end-user by incorporating physical fog nodes (e.g., gateways, switches, etc.) or virtual fog nodes (e.g., virtualized switches, virtualized machines, etc.) between the edge and cloud layer to offload the cloud (Iorga et al., 2018). VFC is an extension of the FC paradigm to conventional vehicular networks, enabling more ubiquitous vehicles to be supported (Huang et al., 2017). In the VFC architecture introduced by Huang et al. (2017), vehicles are considered intelligent devices with additional computational and communication capabilities to collect useful traffic information due to their mobility and multiple sensors. It leads to better communication efficiency and improvement in terms of latency, location awareness, and real-time responses compared to traditional CC.

Besides its advantages, VFC comes with security risks and threats like *Privacy Leakage* or *Architectural/Software Weakness* that may even have lethal consequences (Klein et al., 2022). The importance of such security threats is compounded by the technology's use in road traffic, where failure could put human lives at risk. In addition, a better understanding of security issues and measures will help build societal confidence in VFC systems, which is necessary for VFC to become more widespread and achieve its promised benefits. Threat modeling methods can be used to gain such an understanding of security risks (Hoque & Hasan, 2019). “Threat modeling is a step-by-step process to analyze, identify, and prioritize all the potential threats and vulnerabilities of a system and solve them with known security solutions” (Hoque & Hasan, 2019). To identify security threats and find methods for protection, various studies in the VFC context (e.g., Klein et al. (2022); Hoque and Hasan (2019); Alrawais et al. (2018)) have been conducted. According to Klein et al. (2022), the examination of security threats and protection measures may lead to increased trust in VFC by society.

This is important because VFC becomes increasingly popular and further adoption by the whole society is necessary for VFC to succeed. The adoption of innovation depends, among other factors, on product-specific factors (Bähr-Seppelfricke, 2000). These include the perceived risk as one of the most important factors negatively influencing the adoption process (Bähr-Seppelfricke, 2000). So, to determine whether VFC may be adopted by society, it is thus necessary to understand how society perceives identified security threats VFC comes with.

Objectives

The focus of this seminar thesis is to identify and prioritize the VFC security threats by taking a society's point of view. We will focus on answering:

How does society perceive security threats of VFC?

To answer this research question, we will gather security threats from individuals in society and then let them rank the identified threats. This ranking is intended to list the security threats from most dangerous to least dangerous to gain a better understanding of society's perception of security threats. To achieve this goal, we will identify the security threats in VFC from the perspective of non-domain experts and map them with the security threats of VFC identified in related research, particularly those introduced by Klein et al. (2022), who surveyed VFC experts to identify security threats. Additional security threats mentioned by the non-experts that were not listed by Klein et al. (2022) will be added up to the new list from this study. This mapping will also help understand and see whether and to what extent society's perceptions of existing security threats are reflected in those of the VFC experts. Further, a subset of the identified threats will be

identified to minimize the set of threats to be considered and to get a better overview of the most dangerous threats and how they are perceived. Lastly, a ranking of the security threats in the identified subset will be performed.

Summing it up, we aim to provide an understanding of the perception of security threats so the most important threats can be looked at closer to find solutions to eliminate the threats and positively influence the adoption process of VFC.

Theoretical Background

Vehicular Fog Computing

The National Institute of Standards and Technology defines FC as "a layered model for enabling ubiquitous access to a shared continuum of scalable computing resources" (Iorga et al., 2018). It "facilitates the deployment of distributed, latency-aware applications and services, and consists of *fog nodes* (physical or virtual), residing between *smart* end-devices and centralized (cloud) services" (Iorga et al., 2018). FC can be extended and applied in the field of vehicles and is called "vehicular fog computing". Figure 1 shows an example of a VFC architecture, namely the three-layer model presented by Huang et al. (2017) and Ning et al. (2019), as adopted by Klein et al. (2022). This model is general enough to cover more specialized architectures (Klein et al., 2022).

The edge layer contains the edge devices, for example, intelligent vehicles that perceive the environment through sensors or traffic lights (Klein et al., 2022). This layer is connected through various interfaces to the next layer, the fog layer, represented by road side units (RSU). The RSU make local decisions (e.g., ending warnings to vehicles about bad road conditions and accidents or scheduling traffic lights). However, the computing power of RSU is limited and thus not sufficient to optimize traffic on a large scale. To resolve the issue posed by insufficient computing capabilities, the fog layer is connected to the next layer, composed of cloud servers accountable for large-scale decisions and data analysis (Klein et al., 2022). A key aspect of VFC is security, while threat modeling, which defines a systematic approach for describing and classifying system security threats, can be used to understand the threats (Hamad et al., 2016).

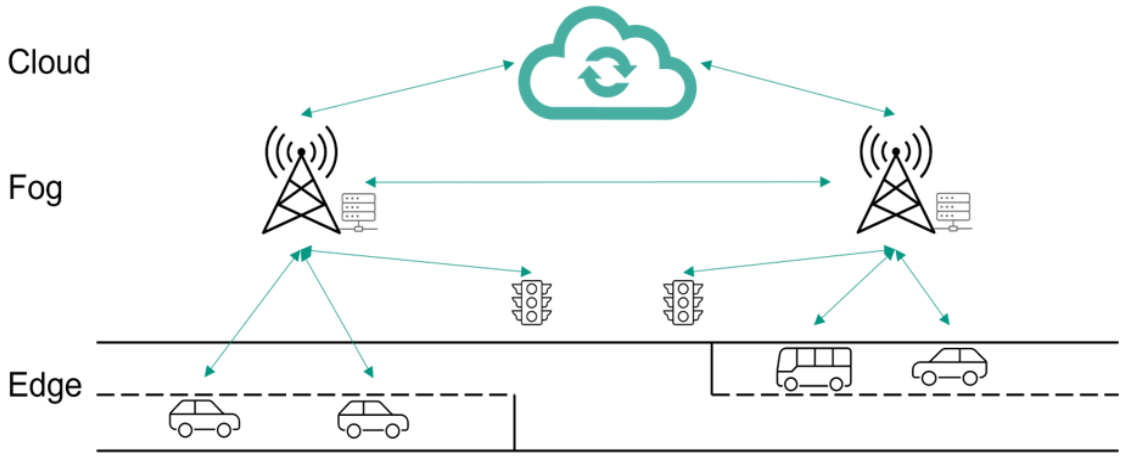

Figure 1. Schematic Representation of a Three-Layer VFC Architecture Adapted from Klein et al. (2022)

Threat Modeling for Vehicular Fog Computing

Among the many existing definitions of security threats, the most general one is given by the Internet Engineering Task Force – it is the potential negative, intentional action or event which is facilitated by a vulnerability and which results in an undesirable impact on a computer system or application (Shirley, 2000).

Gaining a better understanding of the security of VFC, which is a key aspect for developing vehicular information technology, can be achieved by threat modeling (Hamad et al., 2016).

Retrieving a threat model for a system is "a systematic process of identifying and prioritizing the potential threats and vulnerabilities of a system" (Hoque & Hasan, 2019). A threat model aids the process of enhancing a system's security and assessing the risk of attacks. Threat modeling typically involves five steps and components (Hasan et al., 2005; Klein et al., 2022), which namely are (1) *assets*, (2) *entry points*, (3) *attacker model*, (4) *threats and vulnerabilities*, and (5) *mitigation strategies*. However, for this work, we will only look at step four (4) and thus on the identification of threats and vulnerabilities and their perception from a non-expert point of view.

Extant literature has already identified several threats to FC and VFC and proposes different threat models. For example, Klein et al. (2022) provide an overview of research endeavors in the context of security, threat modeling, and (V)FC, stating that research on security threats for VFC is still scarce and scattered across disciplines. Building on this finding, they provide a threat model that synthesizes prior research and empirical insights from expert interviews, which other studies lack (Klein et al., 2022). Thus, while there are other threat models in the context of VFC (e.g., Hoque and Hasan (2019)), the threat model of Klein et al. (2022) seems the most extensive and in-depth.

For their threat model, Klein et al. (2022) rely on a STRIDE-based threat modeling approach by Khan et al. (2017) (Klein et al., 2022). This approach extends the renowned and mature threat modeling method to cyber-physical systems (Shevchenko et al., 2018). Adopted by Microsoft in 2002, STRIDE is an acronym for Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege, which form the categories in which threats may be structured for threat analysis (Microsoft Corporation, 2009). An overview of the identified threats to VFC and their assignment to the STRIDE categories by Klein et al. (2022) is provided in Table 1. Common for all threats in the model, except for *Environmental Noise*, is that an attacker is involved. As we rely on this threat model for our work, we will take up the attacker-centric view of Klein et al. (2022).

№	Category and Description	Threat
1	**Spoofing** *Assuming a wrong identity*	Impersonation Attack
2		Sybil Attack
3		GPS Spoofing
4		Illusion Attack
5	**Tampering** *Manipulation of data*	Bogus Information
6		Stored Data Modification
7		Message Alteration
8		Message Suspension
9		Replay Attack
10		Manipulation of Network Topology
11	**Repudiation** *Denial of actions*	Liability Avoidance
12		False Presence
13		Activity Hiding
14	**Information disclosure** *Revelation of data*	Sniffing
15		Data Breach
16		Data Breach through Physical Access
17		Eavesdropping
18		Privacy Leakage

№	Category and Description	Threat
19	**Denial of Service** *System disruption*	Jamming
20		Malware
21		Physical compromising
22		Communication Obstacles
23		Bandwidth Overhead
24		Denial of Service Attack
25		Physical Denial of Service
26		Distributed Denial of Service Attack
27		Black Hole
28		Architectural-/ Software Weaknesses
29		Environmental Noise
30	**Elevation of privilege** *Unauthorized access to resource*	Unauthorized Access
31		Improper Resource Allocation
32		Improper Resource Sharing
33		Privilege Escalation

Table 1. Overview of the Security Threats Identified by Klein et al. (2022) and their Assignment to the STRIDE Categories

Method – Conducting a Ranking-type Delphi Study

A ranking-type Delphi study is performed to answer the research question and obtain knowledge on the perception of security threats in VFC. As for now, no insight into the perception of security threats by non-domain experts is considered in the literature. Although there are threat models for VFC, as described in Section 3.2, they are built from an expert point of view and do not depict non-experts' perceptions. Thus, an approach is necessary that allows gathering novel insights from a non-expert point of view and into their perception. The Delphi method aims for the collection and consolidation of expert judgment in complex decision-making processes (Gnatzy et al., 2011; Hsu & Sandford, 2007). Due to the research objective of this work, participants, however, should not be technical experts in VFC as suggested by the literature but non-experts in the field. Such an approach is also reported by Thiebes et al. (2018), who seek for individuals that donated their genome data in the past or decided against it to participate in their study, as their goal is to identify motivating and demotivating factors for individuals to donate their genome data. However, one could consider these individuals experts for the objectives of this research, which holds for our case as well. To indicate the focus of this work on a non-expert point of view, we will name the participants of the study to be non-experts, as they are not experts in the topic area of VFC, although one might consider them experts with regard to the research objective.

Whilst there are different kinds of Delphi studies, we perform a ranking-type Delphi study as outlined by Schmidt (1997) as it is one of the most widely used Delphi variations (Thiebes et al., 2018). Furthermore, it allows for the discovery of new elements and the relative ranking of these, thus gaining valuable insights into the perception of security threats. Following Schmidt (1997), ranking-type Delphi studies consist of three subsequent phases, namely (1) *Brainstorming*, (2) *Selection*, and (3) *Ranking*. Figure 2 depicts our research approach for a ranking-type Delphi study as suggested by Schmidt (1997) and adapted from Thiebes et al. (2018).

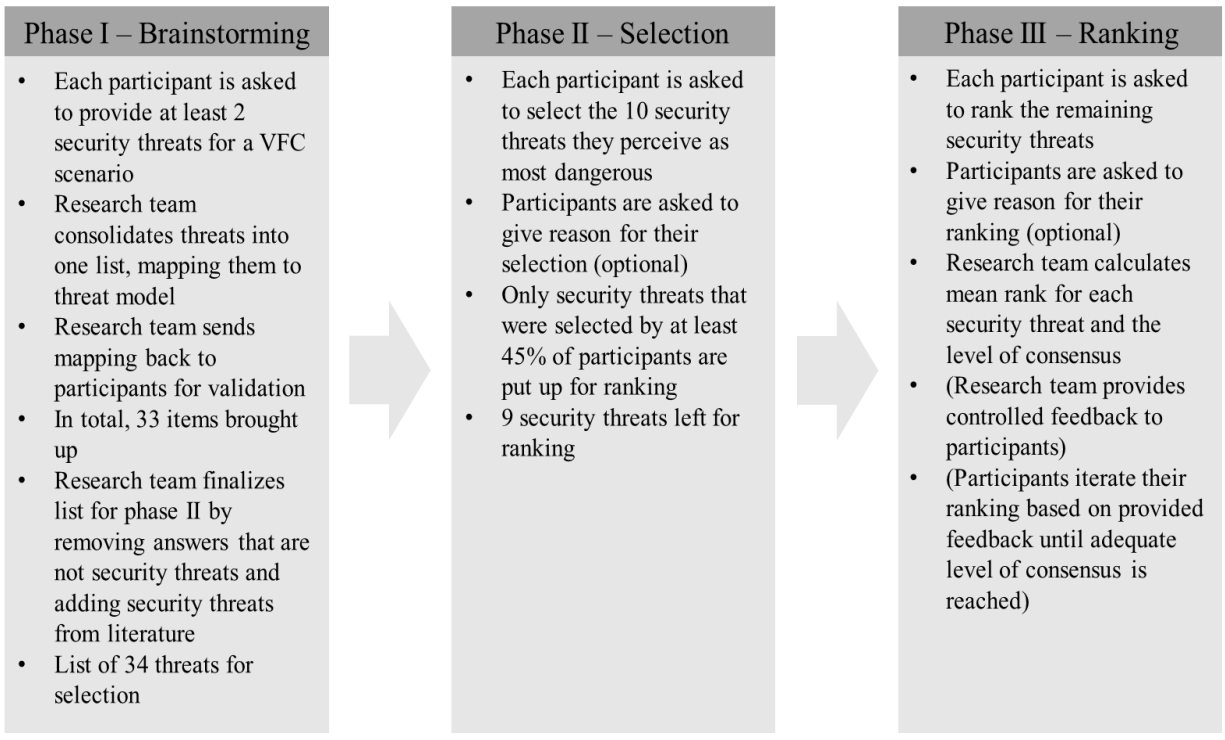

Figure 2. Three-Phase Research Approach Adapted from Thiebes et al. (2018)

Panel Selection and Recruitment

Delphi studies typically do not rely on a representative statistical sample but rather on knowledgeable experts with a deep understanding of the problem domain (Okoli & Pawlowski, 2004). Thus, the selection of experts is a critical factor for Delphi. As the objective of this work is to get an understanding of the perception of security threats in general society, the experts to gather do not need a deep understanding of VFC, but instead should be chosen independently from their knowledge on the topic not to deselect any relevant perceptions in the study. Thus, we did not require the participants to have any prior knowledge of the topic. As VFC is a novel technology and different countries may be differently fast at development and usage of novel technology, having participants from different countries may be difficult. Different residences may come with differently strong interaction or exposition to the topic, as Klein et al. (2022) indicate from a research perspective. For this reason, only German residents were considered part of this study.

To recruit participants, a PDF-flyer was created that covered all essential aspects of the study, including the purpose, method, duration, and medium over which the survey would be conducted. The flyer also contained a link to the survey and referenced an online document with a more in-depth description of the research's motivation, objectives, and method. Please note that the research objective presented in the flyer and the online document supersets this work's objectives. During this research, we could not set up a panel of experts on VFC or related topics to contrast their perception of security threats to the perception of non-experts in the field. After the creation of the flyer, it was distributed via postings on social media platforms (e.g., Facebook, Instagram, LinkedIn) or via direct messaging using WhatsApp or Email. Even though Delphi does not rely on a large participant group, the goal was to reach as many potential participants as the three-phase approach may lead to high dropout rates (Bardecki, 1984). We anticipated that the dropout for our study would be even higher than reported in the literature, as the non-expert participants may not be as inclined to take part in all phases as experts with a personal interest in the topic may be.

The distribution of the survey link resulted in 220 clicks on the survey, including multiple clicks by the same person. Of these 220 potential participants, 28 (12.73%) started the survey, of which 22 (return rate of 10%)

filled out the central question for the first phase. Removing any meaningless answers indicating that the participant only wanted to proceed through the questionnaire left 21 participants' answers.

For the expert panel we tried to set up, 24 clicks were reported, again involving multiple clicks by the same person. Table 2 lists relevant demographic data of our 24 panelists.

Characteristics	Panel Profile[#] (n=24)	
Gender	Female: 29.17% (7) Male: 62.50% (15) Prefer not to say: 8.33% (2)	
Age	Min: 20 years; Max 53 years <25 years: 37.50% (9) 25-30 years: 50.00% (12) > 30 years: 4.17% (1) Prefer not to say: 8.33% (2)	
Current job status	Apprentice: 4.17% (1) Employee: 20.83% (5) Self-employed: 4.17% (1) Un-Employed: 4.17% (1) University Student / Undergraduate: 62.50% (15) Prefer not to say: 4.17% (1)	
Technical knowledge (n=21)*	"I have a lot of technical knowledge":	3.10
	"I want to be up-to-date regarding technological advancements":	3.62
	"In contrast to most other people, I have a lot of technical knowledge":	3.24
	Overall mean:	3.32
VFC knowledge (n=23)[+]	"I know a lot about VFC":	1.83
	"I have a lot of knowledge on VFC":	1.74
	"In contrast to most other people I know a lot about VFC":	2.00
	Overall mean:	1.86

[#] For percentage values, the absolute value is always reported in brackets.

* This question was only posed to the former non-expert panel, as experts were thought to have sufficient technical knowledge. For discussion on this, see Section 4.2. Statements were to answer on a Likert scale from 1 (total disagreement) to 5 (total agreement).

[+] One of the participants did answer, "Prefer not to say." Statements were to answer on a Likert scale from 1 (total disagreement) to 5 (total agreement).

Table 2. Overview of Participants' Demographics

The link to the expert version of the survey was distributed via LinkedIn, Email, and personal networks of the authors. To be considered an expert in VFC or a related field, work, or research experience of at least two years in the expert's field was required. In total, 81 potential experts were contacted, of which four did start the survey, and three (return rate 3.70%) did at least partially fill out the central question of the Phase

I questionnaire. Continuing the expert panel with only 3 participants did not seem feasible. However, the experts answering demographics questions did report not having any experience with VFC. To not lose their input on the topic and strengthen our non-expert panel, we decided to include the participants in the non-expert panel, resulting in 24 participants in total.

Data Collection and Analysis Methods

To enable as many people as possible to participate in the study, we opted for an online survey using the survey tool "SoSci Survey" for the questionnaire rounds and relied on Email and other means of communication, such as WhatsApp, for communication with the participants. An online survey may also reduce the inhibition threshold to participate in the study, as it can be accessed conveniently by the participants and at times they want (Heiervang & Goodman, 2011). Each questionnaire was designed based on a ranking-type Delphi study conducted by Lins et al. (2020), which was provided to us by the supervisor of this seminar thesis, who also pretested our questionnaires. We hoped to gather as many participants as possible and keep the dropout rates as low as possible by employing such an online strategy. Using rapid communication media for contacting participants allows for speeding up the turnaround time between questionnaires, as Delphi may take a large amount of time for data collection without such media (Okoli & Pawlowski, 2004).

Phase I – Brainstorming

After participants interacted with the link to the Phase I questionnaire, which covered the Brainstorming Phase, the participants were welcomed and presented a short introduction to the Delphi methods process. They then were introduced to the study's context, purpose, and objectives. Next, panelists were presented with a scenario in which they should imagine being a participant in a crossroad scene. The scenario was used to aid participants that may not be familiar with VFC to familiarize themselves with the components, their purpose, and the interaction between those. Following the scenario, participants were asked to brainstorm possible security threats they could imagine by naming and describing them (Okoli & Pawlowski, 2004; Schmidt, 1997). While Schmidt (1997) suggests asking each participant for at least six items to identify as many novel items as possible, participants were asked to only list at least two security threats for two reasons. First, participants in our case were not domain experts and thus likely to be overburdened with listing too many security threats, and second, the aim for this phase was not to come up with a complete list of as many security threats as possible but to enhance the security threats already identified in the threat model by Klein et al. (2022) with security threats from a non-expert point of view. The participants could list up to 8 more security threats so that a maximum of ten threats per brainstorming question and participant was possible. The three participants of the former expert-panel were presented the same welcoming screen, introduction, and scenario. In contrast to the non-expert panel, they were asked the brainstorming question three times, each time with regard to other components of the VFC's three-layered architecture, so it is possible that the maximum of 10 answers per non-expert participant is exceeded. The survey was ended with demographics questions, a participant label, and the possibility to provide an Email address for further contact.

In total, 25 participants completed the Phase I brainstorming question(s). One of the non-experts did only answer the first brainstorming question. As we deemed their input relevant, we still included it in the Phase I data. However, we needed to disband one participant's answers, as they were not security threats but meaningless answers to end the survey. Thus only 24 participants' answers were deemed relevant, providing us with 77 security threats (avg. 3.21; med. 2) in total. Every participant did at least list two security threats, and the maximum number of answers by a single participant was 12. The answers were then consolidated by matching them to security threats described in the threat model by Klein et al. (2022) if an answer's threat name and description fits the description of a threat in the model. If no fitting security threat could be found, a new threat was created with a description according to the participant's answer. In the creation of new threats, we tried to unify terminology in case several participants raised the same threat in different wording. As participants' answers mostly were in non-specialists' terms and vague or raised several distinct threats in one answer, we allowed the assignment of one answer to multiple threats. Consolidation was performed by two coders, followed by a discussion and consolidation iteration in the group of all authors, resulting in 33 security threats that were brought up in the brainstorming. As the consolidation step involves much judgment by the coders, it is necessary to get the consolidation validated

by the participants to ensure that their answers have been adequately mapped and fairly represented (Schmidt, 1997). To allow for validation, we sent a list of all participants' answers and the mapping to the security threats, including a short explanation for the mapping, to the participants. Ending the questionnaire, only 14 participants had left an Email address for further contact. Thus, the validation mail only reached 14 of the 24 participants. We decided not to send the list via other communication channels to not give the list to people not involved with the study. No changes to the consolidated list of security threats were necessary. Combined with the not mentioned security threats of the threat model, Phase I resulted in 39 security threats. However, not all of the security threats mentioned by the participants qualify as security threats with regard to the threat model by Klein et al. (2022). Thus, we deleted five threats from the consolidated list, a decision discussed in the results section alongside further insights into the security threats (not) reported.

Phase II – Selection

As too many items may hinder the ranking exercise (Schmidt, 1997; Schmidt et al., 2001), participants were asked to narrow down the list of 34 security threats. It is important that not the researcher decide the number of items deemed relevant, but the participants do so themselves (Schmidt, 1997; Schmidt et al., 2001).

After being welcomed to the second phase and stating that they had already participated in the first phase's questionnaire, participants were reminded of the research background and objectives. They then were presented a randomly shuffled list of all threats and asked to choose, but not rank, the ten security threats they perceived as most dangerous. This number of threats is supported by literature (Okoli & Pawlowski, 2004; Schmidt et al., 2001; Thiebes et al., 2018) and a pretest, which deemed 10 to be a reasonable number of threats. However, in contrast to the mentioned literature, we did not ask for at least ten items, but exactly ten to not overburden the non-domain-experts with the decision of how many items exactly to include. It also prevents the participants from not choosing the most dangerous items but nearly all of them. Participants were also asked to give a reason for their selection, which was, however, only optional. Ending the survey, participants were again asked for their participant label and an Email address for further contact. Sixteen participants responded to the second phase invitation (dropout rate: 33.33% with respect to Phase I) by filling out the questionnaire, of which 15 participants provided a selection and one reporting not being able to perform the selection task. One of the 15 participants did select 11 threats, whereas all other second phase participants did stick to the specification of only choosing ten items. To not lose any input and with regard to literature allowing at least ten items, we decided to keep this answer.

To form the final set of selected security threats, Schmidt (1997) suggests a simple majority vote, thus keeping every item selected by at least half of the participants. However, in the case of this research, this would have resulted in only keeping four security threats for the Ranking Phase. As this does not seem reasonable and results in losing too many items, we did decide to experiment with lower cut-off values resulting in a cut-off at 45%, yielding the most promising results with nine items being kept for ranking. This step is again supported by extant literature reporting different cut-off values (Thiebes et al., 2018). As shown in Figure 3, which provides an overview of the number of items for different cut-off values in steps of 5 percentage points, any value from 35% to 45% results in a list of 9 items. Lowering the cut-off to 30% yields 15 items selected for ranking, which is still nearly half of the items and deemed too many by the authors.

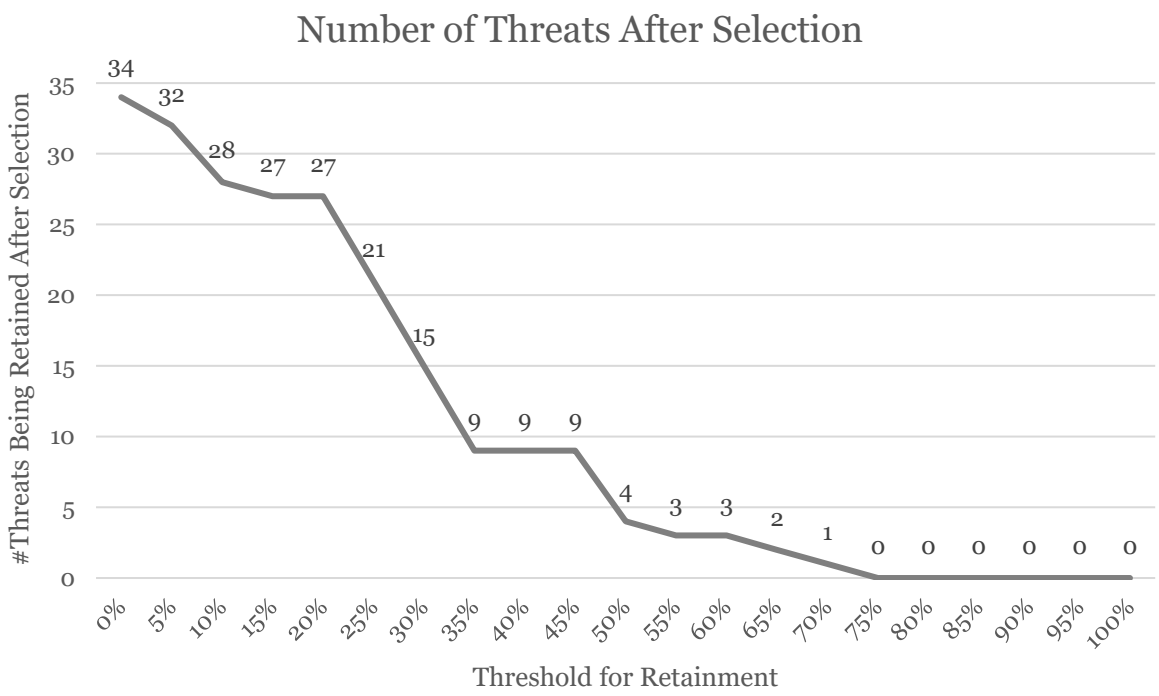

Figure 3. Graph of the Number of Security Threats After Selection for Different Cut-Off Values

Phase III – Ranking

After narrowing down the list of threats to the nine selected ones, participants were invited to the third and last phase of the study, the Ranking Phase. The questionnaire was structured in the same way as the second phase's, except for the selection question, which was switched out for the ranking question, and the statement in the beginning now also involving the participation in the second phase. To perform the ranking, Participants were asked to rank the nine security threats, which were randomly shuffled for each participant, in order of perceived dangerousness from 1 (most dangerous) to 9 (least dangerous). Participants were provided with an overview of the selection percentage for all 34 security threats and a description of the threats to support their decision. After the ranking exercise, participants were again asked to give a reason for their ranking.

In the end, 12 participants took part in the last questionnaire (dropout rate: 50% with respect to the first phase). As suggested by Schmidt (1997), we computed Kendall's W as a measure of consensus, reaching from 0 (no consensus) to 1 (perfect consensus) (Thiebes et al., 2018). Further, mean ranks were computed for each of the nine security threats to reach a final consolidated ranking.

The 12 participants reached a consensus of W=0.2322, resulting in a weak level of consensus (Schmidt, 1997). The ranking and a discussion of the level of consensus are provided in the results section. Due to time restrictions on this work, another iteration to increase the level of consensus was not performed.

Results

Phase I

For the Brainstorming Phase, the participants reported a total of 77 answers, which led to 149 multi-assignments to 33 threats. On average, participants mentioned 3.21 security threats while the median is 2. As 2 was the requested minimum, this circumstance could indicate that participants may not be familiar with the subject of VFC or that this task was perceived as challenging. Table 3 shows the mentioned security

threats and categories ordered in descending number of references that were reported by the participants for Phase I, which could be assigned to the categories of the STRIDE model. As some security threats were not reported by the participants, they are not listed in Table 3. The same table, ordered by category, can be seen in Appendix A, including the security threats not mentioned.

Category	Threat	Times Mentioned	Threat Description
Denial of Service	Architectural-/ Software Weaknesses	18	Weakness resulting from powercuts or erroneous source code
Information Disclosure	Data Breach	16	Stealing, disclosing data
Information Disclosure	Privacy Leakage	16	Disclosing confidential information
Tampering	Bogus Information	8	Transmitting wrong information
Spoofing	Impersonation Attack	7	Imitation of a legitimate user
Elevation of Privilege	Unauthorized Access	7	Gaining access without authorization
Tampering	Stored Data Modification	6	Logs or stored data are modified
Information Disclosure	Eavesdropping	5	Secretly listening to private communication
Denial of Service	Bandwidth Overhead	5	Limited capacity for communication
Denial of Service	Denial of Service Attack	5	Fake requests overload the system
Tampering	Message Alteration	4	Altering a sent message in transmission
Information Disclosure	Sniffing	4	Reading communication between entities
Denial of Service	Communication Obstacles	4	Physical objects prohibit communication
Denial of Service	Environmental Noise	4	Natural noise disturbing the signal
Denial of Service	Malware	3	Injecting disruptive code
Denial of Service	Physical compromising	3	System disruption through physical capture of components
Denial of Service	Physical Denial of Service	3	Incapacitating system components via physical actions
Denial of Service	Distributed Denial of Service Attack	3	Fake requests from different entities overload the system
Tampering	Manipulation of Network Topology	2	Modifying the topology information

Category	Threat	Times Mentioned	Threat Description
Spoofing	Sybil Attack	1	Introduction of ghost identities by one attacker
Spoofing	Illusion Attack	1	Broadcasting a traffic warning message to create illusion
Tampering	Replay Attack	1	Replaying a previously transmitted message
Information Disclosure	Data Breach through Physical Access	1	Data breach achieved via physical actions
Denial of Service	Jamming	1	Disturbing the signal
Denial of Service	Black Hole	1	Disrupted node swallowing information
Elevation of Privilege	Improper Resource Allocation	1	Gaining more resources than the fair share
Elevation of Privilege	Privilege Escalation	1	Abusing legitimate privilege

Table 3. Overview of Threats from Threat Model by Klein et al. (2022) That Were Reported in Phase I

In the following, some of the reported security threats are discussed ordered by their affiliation to the respective category. The category "Denial of Service" with a total of 50 assignments, including the threat *Architectural/ Software Weakness,* which was mentioned 18 times, was reported more often than any other category. Participants stated that the software as a whole or its components might fail or lack technical innovations over time. The threat *Bandwidth Overhead* includes answers about the limited communication capabilities of the system and, therefore, the problem of failure. Especially the concern about the vast amount of data produced by the traffic participants and therefore the bandwidth capacity and the associated difficulties for the processing units seem to bother the participants mentioning this security threat. *Denial of Service Attack* was stated with the question of whether the system could be overloaded with sending many requests to different components of the system at a time. *Environmental Noise* includes answers about connection problems of different components or a failure of connection in general. Participants who reported *Malware* explained the possibility of hacking into the system and injecting harmful code that could lead to failure of components or the system as a whole. *Physical Compromising* includes answers about attackers gaining access to physical objects and manipulating them to achieve failure or malfunction of components. Participants who mentioned *Communication Obstacles* were concerned about problems when components, for example, cars, lose connection to the network. *Physical Denial of Service* like a flat tire, as a participant stated, could be interpreted by the system wrongly, resulting in a malfunction of the system, for example, calculating wrong GPS coordinates. *Distributed Denial of Service* was described similar to and named together with *Denial of Service Attack*.

From the category "Information Disclosure", the threats *Data Breach* and *Privacy Leakage* were both mentioned 16 times each. Stealing the (personal) data and selling it or manipulating the network with the stolen data were mentioned for *Data Breach*. *Privacy Leakage* was brought up along with concerns about personal data that might fall into the wrong hands and the resulting consequences. For example, attackers knowing the exact coordinates of the vehicles and therefore knowing when the house owners are not at home seems to be a concern of the participants as it facilitates theft. *Eavesdropping* was reported as listening to information secretly and in one case, along with poorly encrypted information transmission enabling third parties to decrypt the information. The threat *Sniffing* was stated as listening to the communication between vehicles and fog nodes. *Data Breach through Physical Access* was reported as

accessing hacked nodes and other manipulated components that might leak data to unauthorized third parties.

The category "Tampering", which was mentioned 21 times in total, includes answers of the threat *Bogus Information* about the concerns of receiving intentionally manipulated data or data that doesn't represent the actual state of the system. *Stored Data Modification* cohered with the concern that companies or attackers could influence the data stored. The security threat *Message Alteration* was stated by the participants as manipulated communication between edge devices and fog nodes. Participants who mentioned the *Manipulation of Network Topology* stated that fog nodes or other components could be manipulated with the goal of changing network links. The answer for the security threat *Message Suspension* suggests that a delayed calculation operation by the system might lead to collisions of vehicles. None of the participants mentioned any threat from the category "Repudiation", which means the denial of actions (with the threats *Liability Avoidance*, *False Presence*, and *Activity Hiding*). It could be that this security threat category is seen as unlikely to happen by non-experts or they did not know how to describe these threats. Additionally, the threat *GPS Spoofing* from the category "Spoofing", and the threat *Improper Resource Sharing* from the category "Elevation of Privilege" were not mentioned by the participants.

Threat Name	Times Mentioned	Decision (Short Description)
Trust in System or Service Providers	7	No security threat. Has the potential to become a security threat if the trust is misused (e.g., by service providers that leak data to security agencies or misuse their privileges). These aspects are already covered in STRIDE threats.
Wrong Information	3	Security Threat because wrong information may have escalating effects and may affect system availability or functionality in a critical way.
Exhausted Vehicle Resources	3	No security threat*
Exhausted Fog Service / Fog Node Resources	2	No security threat*
Exhausted Cloud Resources	2	No security threat*
Strange Human Behaviour	1	No security threat. Only safety-related.
* In general, more safety-related. It has the potential to become a security threat in case of DoS, DDoS, or in case of missing redundancy. These aspects are already covered in STRIDE threats.		

Table 4. Overview of Threats not in Threat Model by Klein et al. (2022) That Were Reported in Phase I

After the assignment of the answers from the participants, six threat names couldn't be properly allocated to the categories of the STRIDE model. Table 4 shows the threat name, the number of answers that were assigned to the respective threat, and a short description of the decision why we didn't include the answer into the following analysis of Phase II and Phase III.

Most of them were safety-related without a link to security, which is indeed a challenge for VFC as it is for every other information system, but it was not the respective topic of this work and therefore was not included in the final coding. There weren't named any security threats whose security-related issues weren't already covered by the STRIDE model, despite the security threat *Wrong Information*. We kept *Wrong Information* because of the impact on the system's security that may be greater than the impact of an attacker and because its security-related aspects were not covered in the model.

Phase II and III

By analyzing the answers from the Selection Phase, we find that the security threats belonging to the categories "Information Disclosure" and "Denial of Service" are perceived as most important by many participants. *Privacy Leakage*, *Data Breach* and *Eavesdropping,* belong to the category "Information Disclosure" and *Malware*, *Denial of Service Attack*, *Architectural-/Software Weaknesses* and *Jamming* classified in the category "Denial of Service" are disproportionately often represented in the top nine security threats selected by the participants. This may indicate that those are either most intuitive or that non-experts perceive those as most relevant for VFC. This allocation is similar to the Brainstorming Phase where "Denial of Service" and "Information Disclosure" were reported most frequently. The least selected threats were *Message Suspension*, *Replay Attack*, *Liability Avoidance*, and *Environmental Noise*, with 7% each, while *Improper Resource Allocation* and *Improper Resource Sharing* were not selected at all. From the free comment section with explanations for their selection by the participants (n=7), two clusters of participants may be distinguished. Those who choose privacy first and others who put physical integrity first. For instance, a participant stated that saving human life is most important to them, while others reported that, in their opinion threats, which may affect the operating vehicles are the most dangerous. Therefore, those security threats have a direct impact on traffic security. However, other participants put privacy concerns first with the argument that traffic participants may be blackmailed, or their personal data could be used in a harmful way. For example, burglary from the exact GPS position or identity theft would be facilitated by leaks in the system's privacy.

The ranked aggregation for the Selection Phase can be seen in Table 5. For Phase III, the top nine chosen security threats were selected with a cut-off of c=45%, as discussed in section 4.2.

Rank	Category	Security Threat	Phase II	Phase III
			Percentage of Selection	Mean Ranking
1	Elevation of Privilege	Unauthorized Access	66.67%	2.5
2	Denial of Service Attack	Malware	60.00%	3.66667
3	Information Disclosure	Data Breach	53.33%	4.25
4	Denial of Service Attack	Architectural-/Software Weaknesses	46.67%	4.91667
5	Denial of Service Attack	Denial of Service Attack	46.67%	5.33333
6	Denial of Service Attack	Jamming	46.67%	5.66667
7	Spoofing	Sybil Attack	46.67%	6
7	Information Disclosure	Privacy Leakage	73.33%	6
9	Information Disclosure	Eavesdropping	46.67%	6.66667

Table 5. Overview of the Percentage of Selection for Phase II and the Mean Rankings for Phase III

After conducting Phase III, we end up with *Unauthorized Access* being ranked the most dangerous, followed by *Malware* and *Data Breach*, which could be expected when looking at the frequency of threat selection from Phase II. Interesting is the fact that although *Privacy Leakage* was selected by about 73.33% of the participants in the Selection Phase, it is not rated as the most dangerous but occurs in the second last place. It should be noted that the seventh place was assigned twice, and therefore there is no eighth place. All of these security threats have in common that data is leaked or manipulated in a malicious manner. This may

be an obvious security threat for non-experts when asked about the security threats associated with big data technologies, which may be the reason that they were stated and chosen as the most dangerous ones. *Unauthorized Access* was ranked as the most dangerous and had a selection rate of 66.67% in Phase II, which makes it the second most frequently mentioned threat. Overall, this indicates a shift from privacy-related concerns along with manipulation or theft of personal data in Phase II to more security-related concerns in Phase III when ranking the most dangerous security threats. This was also stated by one participant in the free comment section of the survey, stating that danger to the physical world, such as cars and humans, seems to be more dangerous to them than leaked personal data. Security threats posed by hackers that intentionally want to manipulate the system are more dangerous, as reported by another participant. Problems occurring when the system fails are perceived as dangerous when looking at the free comment section, which reflects the ranking. In total, four free-text answers were given by the participants.

Discussion

Principal Findings

After conducting a Delphi study, we ended up with a ranked list of the nine security threats in VFC perceived as the most dangerous by the participants. They belong to four of the STRIDE categories. *Unauthorized Access* is the threat perceived as most dangerous and falls into the category "Elevation of Privilege". Besides authorization to the system, participants also perceive threats such as *Malware*, *Architectural-/Software Weaknesses*, *Denial of Service Attack,* and *Jamming* out of the "Denial of Service" category, and "Spoofing" category's *Sybil Attack*, as dangerous, indicating the importance of service availability and safety. Furthermore, *Data Breach*, *Privacy Leakage*, and *Eavesdropping* related to the "Information Disclosure" category are in the ranking, suggesting that participants also value their data privacy. During the Brainstorming Phase, participants recovered almost every category and threat of the STRIDE threat model proposed by Klein et al. (2022), except for the "Repudiation" category.

Implications for Research and Practice

The implication for research and practice includes prioritizing security threats by potential users based on the STRIDE model. VFC companies, developers, and researchers are enabled to focus on security threats and solutions that are important to ordinary users, which could positively affect the adoption and usage of the new technology.

Most of the threats and categories from the STRIDE model were reported by the participants during the Brainstorming Phase, implying that the model can be applied to describe what security threats users see in a VFC context, at least to a certain degree.

Limitations

However, this study does not come without limitations. Due to time constraints, we did not have the opportunity to repeat the third stage to achieve a higher level of stable opinion among survey participants. In addition, the dropout of participants in each phase of the online survey may have been lower if the surveys had been online for longer (Phase I: two weeks; Phase II and III, respectively: one week).

Most of the limitations came after the first phase of the survey. It was found that participants reported threats that we did not consider security threats. For example, we did not include participants' responses, grouped under the name *Strange Human Behaviour* (Strange Human Behaviour might not be accounted for in the decision-making process of the systems components), into the threat list for Phase II. Some information that participants gave may have been discarded in this process.

Also, after the Brainstorming Phase, we found that participants did not finish the questionnaire to the end. This could be due to the need to participate in all 3 phases and hence the investment of time for answers, as well as the fact that the topic of the VFC is complex and unknown to most.

Another limitation is that only 14 participants had the opportunity to confirm their answers after the brainstorming, as the other participants had not left their email addresses. Thus, it may be that answers of participants that could not validate the assignment are assigned or interpreted wrongly.

Moreover, the ranking was only according to dangerousness, without ranking by the probability of occurrence and criticality, to not overburden the participants. This, however, is not in line with risk quantification (e.g., Baier et al. (2014), Meier et al. (2003)).

Furthermore, the representativity of our panel is limited, as the demographics suggest. Most of our participants are about the same age (under 27) and university students due to the sampling technique. Thus, we may not have a complete view of society's attitudes towards technology with regard to the demographics of our panelists (e.g., different age groups, different countries).

Lastly, the survey was in English, and it may have been difficult for the German panel to distinguish between safety and security, as participants' answers indicate.

Future Research

One possible and obvious direction for future research could be ranking of security threats by experts as they have more general knowledge about security threats and VFC technology and can identify threats that non-experts do not see clearly or do not know about. Their opinions may differ significantly in their perception of security threats, showing that non-experts value other aspects than the experts.

To this end, performing rankings by criticality and probability of occurrence does seem more appropriate for an expert panel and may uncover interesting findings.

This study would be worth repeating with a more balanced/representative sample (e.g., age, gender, employment structure), with participants from different countries with different levels of VFC development. This would allow for exploring the public perception of security threats in the VFC from different sides.

In addition to the STRIDE model, other existing threat models (e.g., PASTA, CVSS) may be used in combination with various VFC architectures (e.g., architecture by Hou et al. (2016)) and interfaces such as WiFi or radio frequency bands (Klein et al., 2022). By employing such studies, one could compare the perceived threats, advantages, and disadvantages of combinations, to get a complete picture of societal perception and find the most appropriate design for implementation and usage.

Conclusion

This work set out to shed light on the perception of security threats in VFC within society. Conducting a ranking-type Delphi study with non-domain experts and taking security threats already identified in the literature into account, we identified the nine security threats perceived as most dangerous. Brainstorming amongst our participants yielded 33 potential security threats, which combined with threats from literature and after removal of non-security-related threats resulted in a list of 34 threats. These were then narrowed down to the nine threats perceived as most dangerous by the participants and ranked according to their perceived dangerousness in the last step. Although only reaching a low level of consensus (W=0.2322) amongst the young panel, our findings indicate that non-domain experts are concerned about the privacy of their data that is collected, stored, and used for decision making in the VFC system and about threats that may result in system failure and potentially in harm to traffic participants.

Taking a non-expert point of view, our study contributes to the evolving body of knowledge on VFC security threats by providing insights into VFC security threats' perceived dangerousness. Such an understanding is essential, as the perceived risk novel technology comes with is a factor for its adoption by society, and the importance of solutions like VFC may rise considering the growing amount of data produced by vehicles and their increasing interconnectivity.

References

Alrawais, A., Alhothaily, A., Mei, B., Song, T., & Cheng, X. (2018). An efficient revocation scheme for Vehicular Ad-Hoc Networks. Procedia Computer Science, 129, pp. 312–318. https://doi.org/10.1016/j.procs.2018.03.081

Bähr-Seppelfricke, U. (2000). Die Wirkung von Produkteigenschaften auf die Diffusion von Produktgruppen: Empirische Überprüfung in einem aggregierten Diffusionsmodell (Manuskripte

aus den Instituten für Betriebswirtschaftslehre der Universität Kiel No. 525). Kiel. Universität Kiel, Institut für Betriebswirtschaftslehre. https://www.econstor.eu/handle/10419/147605

Baier, H., Braun, M., Busch, C., Heinemann, A., Margraf, M., Moore, R. C., Rathgeb, C., Schütte, A., & Stiemerling, M. (2014). *IT-Sicherheit*. Hochschule Darmstadt. https://www.dasec.h-da.de/wp-content/uploads/2014/10/its_skript1.pdf

Bardecki, M. J. (1984). Participants' response to the Delphi method: An attitudinal perspective. *Technological Forecasting and Social Change*, *25*(3), pp. 281–292. https://doi.org/10.1016/0040-1625(84)90006-4

Gnatzy, T., Warth, J., von der Gracht, H., & Darkow, I.-L. (2011). Validating an innovative real-time Delphi approach - A methodological comparison between real-time and conventional Delphi studies. *Technological Forecasting and Social Change*, *78*(9), pp. 1681–1694. https://doi.org/10.1016/j.techfore.2011.04.006

Hamad, M., Nolte, M., & Prevelakis, V. (2016). Towards comprehensive threat modeling for vehicles. In M. Vülp, Esteves-Veríssimo Paulo, A. Casimiro, & Pellizzoni Rodolfo (Eds.), *Proceedings of CERTS 2016: the first Workshop on Security and Dependability of Critical Embedded Real-Time Systems* (pp. 31–36). CERTS.

Hasan, R., Myagmar, S., Lee, A. J., & Yurcik, W. (2005). Toward a threat model for storage systems. In V. Atluri, P. Samarati, W. Yurcik, L. Brumbaugh, & Y. Zhou (Eds.), *Proceedings of the 2005 ACM Workshop on Storage Security and Survivability* (pp. 94–102). ACM. https://doi.org/10.1145/1103780.1103795

Heiervang, E., & Goodman, R. (2011). Advantages and limitations of web-based surveys: Evidence from a child mental health survey. *Social Psychiatry and Psychiatric Epidemiology*, *46*(1), pp. 69–76. https://doi.org/10.1007/s00127-009-0171-9

Hoque, M. A., & Hasan, R. (2019). Towards a threat model for vehicular fog computing. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 1051–1057). IEEE. https://doi.org/10.1109/uemcon47517.2019.8993064

Hou, X., Li, Y., Chen, M., Di Wu, Jin, D., & Chen, S. (2016). Vehicular fog computing: A viewpoint of vehicles as the infrastructures. *IEEE Transactions on Vehicular Technology*, *65*(6), pp. 3860–3873. https://doi.org/10.1109/TVT.2016.2532863

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research, and Evaluation*, *12*, Article 10, pp. 1–8. https://doi.org/10.7275/pdz9-th90

Huang, C., Lu, R., & Choo, K.-K. R. (2017). Vehicular fog computing: Architecture, use case, and security and forensic challenges. *IEEE Communications Magazine*, *55*(11), pp. 105–111. https://doi.org/10.1109/mcom.2017.1700322

Iorga, M., Feldman, L., Barton, R., Martin, M. J., Goren, N., & Mahmoudi, C. (2018). *Fog computing conceptual model*. https://doi.org/10.6028/NIST.SP.500-325

Khan, R., McLaughlin, K., Laverty, D., & Sezer, S. (2017). STRIDE-based threat modeling for cyber-physical systems. In *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)* (pp. 1–6). IEEE. https://doi.org/10.1109/isgteurope.2017.8260283

Klein, T., Fenn, T., Katzenbach, A., Teigeler, H., Lins, S., & Sunyaev, A. (2022). *A threat model for vehicular fog computing (Working Paper)*.

Lins, S., Kromat, T., Löbbers, J., Benlian, A., & Sunyaev, A. (2020). Why don't you join in? A typology of information system certification adopters. *Decision Sciences*, pp. 1–34. https://doi.org/10.1111/deci.12488

Marshall, P., & Cases, P. (2021). *Enabling the connected vehicle market to thrive*. TOPIO Networks. https://aecc.org/wp-content/uploads/2021/03/Enabling_the_Connected_Vehicle-White_Paper_V5.pdf

Meier, J. D., Mackman, A., Dunner, M., Vasireddy, S., Escamilla, R., & Murukan, A. (2003). *Improving web application security: Threats and countermeasures: Chapter 3 - Threat Modeling*. Microsoft Corporation. https://docs.microsoft.com/en-us/previous-versions/msp-n-p/ff648644(v=pandp.10)?redirectedfrom=MSDN

Microsoft Corporation. (2009). The *STRIDE Threat Model*. Microsoft Corporation. https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)?redirectedfrom=MSDN

Ning, Z., Huang, J., & Wang, X. (2019). Vehicular fog computing: enabling real-time traffic management for smart cities. *IEEE Wireless Communications*, *26*(1), pp. 87–93. https://doi.org/10.1109/mwc.2019.1700441

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, *42*(1), pp. 15–29. https://doi.org/10.1016/j.im.2003.11.002

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences*, *28*(3), pp. 763–774. https://doi.org/10.1111/j.1540-5915.1997.tb01330.x

Schmidt, R. C., Lyytinen, K., Keil, M., & Cule, P. (2001). Identifying software project risks: An international Delphi study. *Journal of Management Information Systems*, *17*(4), pp. 5–36. https://doi.org/10.1080/07421222.2001.11045662

Shevchenko, N., Chick, T. A., O'Riordan, P., Scanlon, T. P., & Woody, C. (2018). *Threat modeling: A summary of available methods*. Carnegie Mellon University. https://resources.sei.cmu.edu/asset_files/WhitePaper/2018_019_001_524597.pdf

Shirley, R. (2000). *Internet security glossary*. Internet Engineering Task Force. https://datatracker.ietf.org/doc/html/rfc4949

Thiebes, S., Lyytinen, K., & Sunyaev, A. (2018). Sharing is about caring? Motivating and discouraging factors in sharing individual genomic data. In *38th International Conference on Information Systems (ICIS 2017)* (pp. 1–20). Curran Associates Inc. https://www.researchgate.net/profile/scott-thiebes/publication/319990200_sharing_is_about_caring_motivating_and_discouraging_factors_in_sharing_individual_genomic_data/links/59c930330f7e9bbfdc32e62a/sharing-is-about-caring-motivating-and-discouraging-factors-in-sharing-individual-genomic-data.pdf

Appendix A

№	Category	Times Mentioned (Category)	Threat	Times Mentioned (Threat)
1	**Spoofing**	9	Impersonation Attack	7
2			Sybil Attack	1
3			GPS Spoofing	0
4			Illusion Attack	1
5	**Tampering**	21	Bogus Information	8
6			Stored Data Modification	6
7			Message Alteration	4
8			Message Suspension	0
9			Replay Attack	1
10			Manipulation of Network Topology	2
11	**Repudiation**	0	Liability Avoidance	0
12			False Presence	0
13			Activity Hiding	0
14	**Information Disclosure**	42	Sniffing	4
15			Data Breach	16
16			Data Breach through Physical Access	1
17			Eavesdropping	5
18			Privacy Leakage	16
19	**Denial of Service**	50	Jamming	1
20			Malware	3
21			Physical compromising	3
22			Communication Obstacles	4
23			Bandwidth Overhead	5
24			Denial of Service Attack	5
25			Physical Denial of Service	3
26			Distributed Denial of Service Attack	3
27			Black Hole	1
28			Architectural-/ Software Weaknesses	18
29			Environmental Noise	4
30	**Elevation of Privilege**	9	Unauthorized Access	7
31			Improper Resource Allocation	1
32			Improper Resource Sharing	0
33			Privilege Escalation	1

Table A-1. Overview of Threats from the Threat Model by Klein et al. (2022) and the Number of Times They Were Mentioned in Phase I (Categorized by STRIDE Category)

Trade-Offs in Provisioning Artificial Intelligence as a Service

Critical Information Infrastructures, Winter Term 21/22

Johannes Jestram
Master Student
Karlsruhe Institute of Technology
johannes.jestram@student.kit.edu

Alexander Pérez
Master Student
Karlsruhe Institute of Technology
ugetf@student.kit.edu

Isabela Bragaglia Cartus
Master Student
Karlsruhe Institute of Technology
isabela.cartus@student.kit.edu

Jan Decker
Master Student
Karlsruhe Institute of Technology
jan.decker@student.kit.edu

Abstract

Background: *The market for Artificial Intelligence as a Service is increasing. With rising adoption of this cloud paradigm, the service design becomes more and more important. Currently, setscrews and corresponding trade-offs in Artificial Intelligence service design are not well understood.*

Objective: *This work aims to deepen the understanding of Artificial Intelligence service design and the trade-offs that follow specific design decisions. The research questions are:* What are the main setscrews for the design of AI as a Service products*? What trade-offs do AI as a Service providers face, when choosing different designs of their AI as a Service products?*

Methods: *To answer the research questions we conducted ten semi-structured interviews with domain experts in Artificial Intelligence as a Service. Based on the interviews we identified setscrews and trade-offs in Artificial Intelligence as a Service design.*

Results: *There are five groups of setscrews for for Artificial Intelligence as a Service products: reliability and security, performance, relationship to competing products, ease of use, and functionalities. Because performance and ease of use are often affected by changes to setscrews of other groups, they are part of many of the trade-offs that we have identified.*

Conclusion: *Artificial Intelligence as a Service has similar service design trade-offs to other cloud computing paradigms such as Software as a Service. However, new challenges arise such as demand peaks for computing power caused by deep learning services.*

Keywords: Artificial Intelligence as a Service, cloud, machine learning, service design, critical infrastructures

Introduction

In recent years, the surge in computational power, data volume, and networking speed has led to the rise of many new computing concepts. Outsourcing storage, compute, and software has been an ongoing trend. The concept of *cloud computing* serves as an umbrella term for those efforts. Using cloud computing resources enables, among others, rapid scalability of IT services, pay-as-you-go subscription models, and high reliability (Mell & Grance, 2011). Depending on the abstraction level of the cloud service, *Hardware-*, *Platform-*, and *Software as a Service* can be distinguished. With the number of cloud computing users continuously increasing, most companies will use some form of cloud computing in the near future (Statista, 2021).

Further, advances in Artificial Intelligence (AI) have led to multifarious use cases for businesses across industries, with 69% of companies using or evaluating the adoption of AI for their business (Loukides, 2022). Additionally, the already large market for AI products is expected to continue to grow (Statista, 2022). Recently, the cloud paradigm has been applied to the AI domain, thereby extending the set of cloud services by Artificial Intelligence as a Service (AIaaS). For businesses, adopting AIaaS and thereby outsourcing AI-related tasks can lower the entrance barrier for using AI, because no own AI infrastructure is required (Pandl et al., 2021).

Due to their size or the type of their customers, certain cloud services pose critical infrastructure (CI). In Germany, CIs are defined as facilities, installations or parts thereof in the sectors energy, information technology and telecommunications, transportation and traffic, health, water, nutrition, as well as the finance and insurance industries, which are of great importance to the functioning of the community, because their failure or impairment would lead to significant supply shortfalls or a threat to public safety (Gesetz Über Das Bundesamt Für Sicherheit in Der Informationstechnik, 2015). With the number of cloud computing users continuously increasing, most companies will use some form of cloud computing in the near future (Statista, 2021). An outage or security breach can have profound consequences, affecting a vast amount of data, many organizations, and a large number of citizens at the same time. Therefore, cloud services are becoming so important within the society and economy that they themselves pose critical infrastructures. With many cloud providers starting to offer AIaaS, some these services and their providers may be considered CIs in the future as well.

Contrary to the established cloud paradigms, it is not yet clear how AIaaS providers best design their services. In this emerging field, service providers face various issues when designing their products. Customers have diverse AI knowledge, use cases, and requirements. Due to the diverse applications of AI, there exists no one-size-fits-all solution for all needs of the customers. To cope with this, AIaaS providers must adapt their service portfolio accordingly. Product designers and system architects have a multitude of *setscrews* for their service design. Such setscrews are key design design areas, where one of several design options can be chosen which influence the product in a specific and relevant dimension. Examples include the pricing model, availability guarantees, and the types of offered AI services. Here, the question arises which setscrews are most important for AIaaS products. By identifying and correctly implementing these setscrews providers can achieve a competitive advantage. However, implementing setscrews might induce *trade-offs* between specific setscrews of their service. For instance, a well-known trade-off from the distributed storage domain is the CAP-theorem (S. Gilbert & Lynch, 2002). According to the CAP-theorem, only two of the three setscrews consistency, availability, and partition tolerance can be achieved at once. This causes a trade-off between the CAP setscrews. Such trade-offs must be resolved in AIaaS as well, in a way that benefits the specific service.

Besides these pracital considerations, there is little research into AIaaS design. For instance, it is unclear which challenges arise from specific service designs and how to tackle these. Therefore, this work investigates the setscrews and trade-offs that AIaaS providers face when building their products. Our research questions are: What are the main setscrews for the design of AI as a Service products? What trade-offs do AI as a Service providers face when choosing different designs of their AI as a Service products?

To answer the research questions we conducted ten interviews with AIaaS experts. By answering the research questions, this study supports practitioners during design decisions for their products, especially by enumerating design trade-offs and explaining ways to alleviate them. Further, the results help guiding future research efforts by highlighting current challenges in AI service design.

This study is structured as follows. First, Section *Background* explains the AIaaS paradigm. Afterwards, Section *Method* details our research approach. The Sections *Setscrews for AIaaS Providers* and *Trade-Offs in AIaaS Service Design* consist of the main research results and their structured analysis. Lastly, we summarize the results and propose future work in Section *Conclusion*.

Background

Artificial Intelligence as a Service

Artificial Intelligence as a Service are cloud-based systems that provide organizations with on-demand services for the deployment, development, training, and management of machine learning (ML) models (Lins et al., 2021). Similar to cloud computing, AIaaS products may be distinguished by their level of abstraction. These service levels include *AI Software Services* that are ready to use AI applications, e.g., inference and pre-trained models, *AI Developer Services*, such as labeling tools and ML frameworks, and *AI Infrastructure Services*, namely data storage and compute. Cloud computing and AI complement each other in multiple ways. Because recent advances in ML lead to models with millions of parameters (Devlin et al., 2018), training as well as deployment of such models become increasingly expensive and more difficult for companies without dedicated server infrastructures. Here, AIaaS helps companies overcome such model- and data-related challenges. Figure 1 displays the three AIaaS levels.

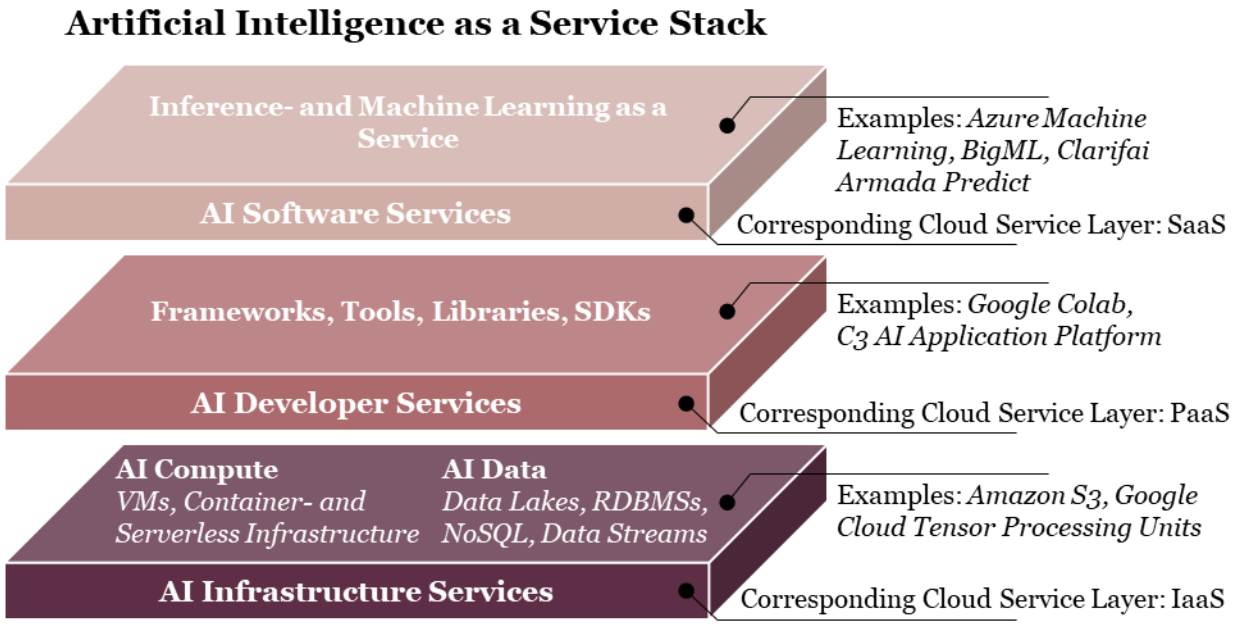

Figure 1. AIaaS Service Levels, Adapted from Lins et al. (2021)

Challenges in Providing AIaaS

All three service levels can be related to the conventional cloud models: AI Software Services relates to Software as a Service (SaaS), AI Developer services relates to Platform as a Service (PaaS), and AI Infrastructure Services relates to Infrastructure as a Service (IaaS). Therefore, AIaaS raises issues and challenges related to AI and cloud computing in general. However, there are various new AIaaS specific challenges and issues, both technological and economic, that can significantly impact their value proposition if not adequately addressed. A very important but insufficiently explored topic concerns trade-offs in provisioning AIaaS, in particular in designing cloud-based AI services.

For example, when deploying pre-trained models, an AIaaS provider must ensure that the datasets used for training are balanced in order to guarantee fairness. Fairness and accuracy are trade-offs, as improving the AIaaS characteristic of fairness affects another characteristic such as accuracy. In 2014, the cloud provider

Amazon Web Services (AWS) had an algorithm that analyzed resumes to decide whether person was a "hire" or "non-hire". The algorithm performed very well on training data, and the use of this experimental tool was widespread in the company. However, after using the tool for a while, AWS noticed that for technical job descriptions, changing the gender in the resume and keeping all other factors the same changed the score. The training data was biased, because it consisted mainly of resumes from male employees, which was in line with the trend of male dominance in the company and the technology industry at the time (Kodiyan, 2019).

So far, these trade-offs and their impact on the configuration and use of AIaaS have not been sufficiently explored. Therefore, this work analyses the setscrews that a provider has with regards to AI-service design, e.g., information privacy and inference latency. Further, based on these setscrews we analyse the trade-offs an AIaaS provider faces for different service levels and designs.

Methodology

To answer the research questions, we chose a qualitative research approach. Since there are very few scientific results around the topic of our research question, this approach is suitable to obtain detailed and new knowledge. Further, because our research question is not based on the assumption of a single truth, but is open to multiple truths, a qualitative approach is more suitable than a quantitative one (Moriarty, 2011). Within the possibilities of a qualitative research approach, we decided to use semi-structured interviews. According to Adams (2015), these are particularly suitable for our work, because we want to learn the thoughts of individuals who are employees of frontline service providers. By recording the experiences of the interview partners, it becomes possible to understand their behavior and their reaction to certain questions and stimuli. Because our research question is embedded in a little researched territory, we currently do not have the knowledge base to structure an interview entirely. Instead, it should be possible for interview partners to report freely and as unbiased as possible. Despite the flexibility to respond to individual answers, the basic structure of a semi-structured interview guide makes sure that the results are complete and comparable with each other (Myers, 2019).

We follow the interview structure proposed by Kaiser (2021). It consists of the ten following steps. The individual steps have to be understood as a chronological processing method on the level of one interview, not of the entire project.

Development of the Interview Guide

Before being able to conduct the first interview, an interview guide was developed (Kaiser, 2021). To benefit from the semi-structured interview approach, the guide must offer a balanced relationship between flexibility and guidance through the conversation. Therefore, it consists of an agenda and the outline of planned topics as well as questions to be addressed, but not a fixed structure or fixed questions (Adams, 2015). Because we interviewed employees of AIaaS provider performing different roles, the guide was tailored to each group. The combination of guiding questions posed in different ways and follow-up questions arising from the course of the interview led to a variety of interesting results based on different perspectives. Despite intensive preparation, the interview guide was continuously modified until the last interview to reflect our newly gained insights and feedback from previous interviews.

Pre-Test of the Interview Guide

In order to avoid misunderstanding in the interviews, a pretest needs to be conducted. For this purpose Kaiser (2021) suggests a pre-test-interview, which can later be used as a basis for the results of the research, if the interview guide does not need to be significantly adjusted afterwards. Since we only had to make minor changes after the first interview, this approach could be adopted.

Selection and Contacting of Interview Partners

In order to find experts in the field of AIaaS who are willing to participate in an interview, we used *LinkedIn*. First, we prepared a flyer that summarized all important information about our research and the planned interviews. In the next step, we identified and contacted 100 suitable AIaaS experts via the LinkedIn chat. In addition to a prepared message and the attached flyer, we sent a link to the appointment setting website

Doodle. This simplified the appointment setting process for contacts, as lengthy appointment setting via the LinkedIn chat was avoided. Instead, numerous free dates and times could be selected, whereupon team members were automatically informed by mail about the appointment booking. Of the 100 experts contacted, six agreed to be interviewed. In order to arrange further interviews, we contacted another 100 candidates on LinkedIn, of whom four agreed to participate.

Table 1 lists all 10 interviewees in chronological order. During the search for interview partners, explicit attention was paid to diversity in terms of the company and the role of the contact person in order to be able to include as many different perspectives as possible. While the employer of the contact person is not shown in the table for reasons of anonymization, the diversity is particularly apparent in relation to the role.

Interview Partner	Role	Years of Experience	Company Industry
i01	Cloud Solution Architect	3	Cloud services
i02	Manager AI, Data Science and Machine Learning	4.5	IT consulting
i03	Technical Consultant	3	IT consulting
i04	ML Engineer	7	Cloud data analytics
i05	ML Practice Lead	3.5	Cloud services
i06	Engineering Lead	12	AI research engineering
i07	Account Manager	2	Cloud services
i08	Solutions Architect	4	Cloud services
i09	Cloud Infrastructure Architect	3	Cloud services
i10	Cloud Consultant	4	IT consulting

Table 1. Information About the Interview Partners

Conduction of the Expert Interviews

After all preparations for the first interview were finished, the phase of conducting the interviews began. To conduct the interviews *Microsoft Teams* was used as the communication channel. The interview partners received a link to the respective meeting the day before the interview. In order to be prepared for all eventualities when conducting the interview, all interviews were held by two or more members of the team. This ensured, for example, that technical problems with one team member did not necessarily lead to the cancellation of the interview. To not impede the structure and the course of the interview by constantly changing speakers, an interview leader was determined before each interview. This person took over the introduction, the main part of the questions during the interview, as well as the outro and farewell. The other team members made sure that the course of the interview did not deviate significantly from the topic of the research question. In such cases, the interview leader was supported by the team with intermediate questions. In the course of the interview, special attention was paid to critically questioning the statements of the interviewee. For this purpose, follow-up questions were mostly used, which deviated from the interview guide. Critical questioning ensures that the expert actually discloses his own relevance structures and detaches himself from formal resolutions of his organization (Kaiser, 2021). This way, information that was actually relevant to the research project could be collected. After the outro and the farewell to the interview partner, the team met directly afterwards to debrief. Here, the feedback just gathered from the interview partner was discussed and, if useful, integrated into the interview guide.

Recording of the Expert Interviews

Depending on the interviewee's preference, either a handwritten transcript was prepared during the interview, or the interview was recorded as audio. With one exception, all interview partners gave their consent to the recording. The audio recording allowed the researchers to focus completely on the conversation. It also automatically ensured the completeness of the information recording. In one case, the expert preferred the preparation of a protocol, since a recording was not permitted by the employer. In this case, two members of the team wrote down the most important findings of the interview in two separate protocols (Kaiser, 2021). These were consolidated and merged directly after the interview. Additionally, after every interview the relevant information about the expert's role was collected in Table 1 to maintain an overview.

Securing the Results

After the preparation and conduction phase of the interviews was finished, this first step of securing the results introduces the evaluation phase. Regardless of the form of interview documentation chosen, a text emerges from the expert interview that forms the basis for the subsequent analysis (Kaiser, 2021). We used *Microsoft Azure* to auto-transcribe the interviews. In most cases, the auto transcript was still heavily error-prone, and was thus only a basis for manual transciption. To correct the errors and add a conversation structure to the script, a manual adjustment followed, using the *f4transcript* software. We listened to the entire audio again, in order to adapt the script as accurately as possible to the audio. Here, the basic rules of transcription according to Kaiser (2021) were followed. For the one interview that was not recorded, we consolidated the recorded results, and sent them to the interviewee for review.

Coding of the Text Material

In qualitative research, coding is commonly understood as a process in which a specific expression is assigned to an explanatory variable . Corresponding variables can be identified in different ways and then assigned to the most important findings of a study. Since to the best of our knowledge there is no categorization of trade-offs in the provision of AIaaS products in extant scientific literature, selective coding could not be applied. Instead, we used open coding in the first run, in which the most important findings are first freely assigned to newly created variables. After this first iteration of open coding, further iterations were needed to establish a consistent coding structure and variables. In this process, we supplemented the open codes variables by subcategories through the application of axial coding. In a final iteration, we reviewed all transcripts a second time by a different researcher. After the last coding iteration a total of 82 text passages with 33 codes were coded for information on setscrews. 14 of the 33 codes were open codes and 19 were axial codes. In the case of trade-offs, 33 text passages were marked with 23 codes. In this case, 3 codes emerged from the initial open coding and 20 codes from the axial coding. Finally, all coding was aggregated and consolidated into a uniform naming and structure. All codes which are relevant to this work are shown in Table 2. For the coding of all interviews we used the tool *f4analyse*.

Groups	ID	Subcodes	Explanation
Reliability and Security	s01	Access Control	The option to assign different rights of access for different users or teams. This includes encapsulated environments that protect a team from critical errors made by another team.
	s02	Encryption	The method that is used to encrypt data in conjunction with services that can handle encrypted data.
	s03	Compliance and Certification	Certification by an independent third party insures that requirements for security and compliance are fulfilled.
	s04	Data Replication and Backup	Creation of backup copies of customer data, code, and models, including replication across multiple locations.

Groups	ID	Subcodes	Explanation
Performance	s05	Execution Speed	The runtime of a service until it provides a result. Mainly driven by inference time.
	s06	Model Performance	The quality of the model output. Measured in accuracy, F1 score, or others.
	s07	Geo-Spatial Location of Data	The geographic locations where the data reside or the services run.
	s08	Scalability	The ability for a service to increase or decrease its usage of computational power based on the current usage. Includes deploying the service to more machines.
Relation to Competing Products	s09	Degree of Integration with other Services	The integration of third party services into the provider's products.
	s10	Treatment of Downstream Services	The treatment of services that build upon the own provided services.
	s11	Flexibility of Product Choice	The different ways a customer can accomplish a task when using a specific AIaaS product or product family.
	s12	Publication and Licensing of Products	The method of publication and licensing of products. For instance, offering products that are built upon open-source libraries or open sourcing product itself.
Ease of Use	s13	Entry Barrier	The hoops that potential customers need to go through before they can use the service.
	s14	Access to Third Party Applications	The possibility for third party developers to provide additional services and plugins complementary to the services of the provider. Often provided via a central marketplace.
	s15	Ease of Application Creation	The ability to create a minimal working application with few clicks. Most settings for such an application are pre-configured.
Functionalities	s16	Offered Model Size	The model size can be measured by the number of parameters, FLOPs/MACs, or the storage size of the parameters.
	s17	Complexity and Variety of Functionalities	The number of possible configurations,the workload types, the types of AI capabilities, such as labelling, MLOps, and data augmentation, and others, of a service.
	s18	General Methods and Specialized Methods	Methods that have very niche use-cases but are highly optimized for those. In contrast to methods that are widely applicable to many tasks.
	s19	Explainability and Transparency	Methods that provide explanations behind outputs of the AI services.

Table 2. Codes Based on the Conducted Interviews

Identification of the Core Statements

Based on the present coding structure, we identified the most important codes that are directly related to the research question in this step of the process. Particularly when using open coding, a large number of different main categories and subcategories are collected, which are thematically relevant, but cannot be used directly to answer the research question. As presented in Table 2, codes of the categories "Setscrews" and "Trade-offs" were selected. We used *Miro* to visualize the codes and relate setscrews to trade-offs.

Expansion of the Data Basis

At this point, we have condensed and structured our material to the point where we could do the analysis and interpretation. According to Kaiser (2021), an expansion of the data basis should take place if not all relevant interview statements can be assigned to the defined coding categories. Since we applied open coding, which was subsequently aggregated and consolidated, there were no relevant interview segments that could not be assigned. Accordingly, an expansion of the data base was not necessary.

Theory-Based Generalization and Interpretation

In the final phase of the analysis, the core statements of the expert interviews are analyzed and interpreted in the light of the theoretical references of the research project (Kaiser, 2021). The corresponding results are presented in the following chapter.

Setscrews for AIaaS Providers

This section details the results of the conducted expert Interviews. Based on our research questions, we first explain the setscrews that a AIaaS provider has when designing its products. Afterwards, we address our second research question and explain the trade-offs based on decisions at different setscrews in Section *Trade-Offs in AIaaS Service Design*.

Based on the insights of the interview partners, we categorized the setscrews for AI service design into five groups: reliability and security, performance, relation to competing products, ease of use, and functionalities. In this section, we explain the specific codes of each group in more detail. Each setscrew is first defined, followed by an explanation of different settings from which a provider can choose. We note that the list of possible settings that we provide for each setscrew is not exhaustive, but rather focuses on the settings we encountered during the interviews.

Reliability and Security

There is a strong overlap between cloud computing and AIaaS regarding reliability and security. However, many of these aspects gain additional design space when providing AIaaS. Therefore, the most important points are discussed below.

Access Control (s01)

IT administrators aim to restrict access to data and code to the users that actually use them. There exist different access permissions, i.e., read, write, delete, and, depending on the type of data, execution. Access can be controlled in different ways. By employing multiple accounts, i.e., one account for each team, IT administrators can very clearly divide access to data between teams or other organizational units (i09, paragraph 38). When allowing multiple accounts for a customer, providers could consider offering a landing zone, where new accounts can easily be created without long waiting times and bureaucracy (i09, paragraphs 52-54). Further, locks on specific objects can be used, which allow changing these objects only with a multi-factor-authentication (i09, paragraph 75). Lastly, access to servers and services can be restricted by distributing them across different sub-networks and controlling access to the respective sub-networks (i07, paragraph 23).

Encryption (s02)

Providers have two options for data encryption: either the encryption happens on the client side (*bring your own key*), or on the provider side. By using client-side data encryption the client can make sure that the provider does not have access to the contents of the stored data. However, in order to use the data as input for ML models they must first be decrypted. The important implication of offering client-side encryption is that most ML methods can not be applied, as they require decrypted data as input (i01, paragraph 91). Therefore, besides the general choice of the offered encryption methods, a provider must consider the compatibility of the encryption level that the customers want or require with their chosen AI service.

Compliance and Certification (s03)

Regulatory compliance is important for most companies, especially regarding data privacy (i01, paragraph 96; i07, paragraph 39; i08, paragraph 46). The use of external AIaaS products may not compromise the regulatory compliance of the customer. In order to prove regulatory compliance, AIaaS providers can choose to get security as well as compliance certifications. As further measures, providers can perform penetration testing and audits (i07, paragraph 47).

Data Replication and Backup (s04)

To prevent the loss of data, providers can choose to physically replicate relevant customer data. Here, AIaaS adds new types of relevant data, namely checkpoints during model training and parameters of trained models, which might be costly to create. As specific measures, providers can choose to replicate the data to multiple availability zones (i07, paragraph 23). Further, versioning of data can be appropriate to ensure the capability to revert unwanted changes (i09, paragraph 75).

Performance

Execution Speed (s05)

In the context of AIaaS, execution speed refers to the speed at which the service can execute AI functionalities. The main factors are training and inference time of ML models. To improve execution speed, providers can offer specialized tools, e.g., for managing end-to-end ML lifecycle (i10, paragraph 87).

Model Performance (s06)

The quality of the model is referred to as model performance. The performance is measured based on the comparison of the model's predictions with the ground truth values of the training data. Measures to improve performance of a model include the possibility to augment the data set and increasing the number of parameters as well as the model size (i06, paragraph 56). For specific use cases that require good model generalization, training multi modal models can improve the performance. Multi modal models can handle not only text, but also images, sound, and other sources from which world knowledge can be learned (i06, paragraphs 29, 95). There are other ways to improve performance that AIaaS providers can use to meet the needs of the customers. Lastly, model performance requirements vary depending on the customer and specific AI use case.

Geo-Spatial Location of Data (s07)

The geo-spatial location of the data influences the propagation delay when accessing the data. In certain use cases where real-time data is crucial, ultra-low latency is required. This is the case, for example, with stock trading (i01, paragraph 169). By employing edge clouds, i.e., placing servers close to the customer and thereby reducing the physical distance to the data source, providers can improve latency. However, this is costly and therefore difficult to implement for smaller providers (i02, paragraph 102).

Scalability (s08)

We refer to scalability as the ability for a service to increase or decrease its usage of computational power based on the current usage of the service. While for cloud applications scalability is needed during workload peaks, AI services require additional scalability for complex computations that have to be executed in short periods of time (i09, paragraphs 96-97). In this case, providers must offer the ability to scale computational power and data storage resources according to customer demand. AIaaS providers can limit and adjust the resource allocation for a customer by adjusting the backplane sizes. Backplane size is a limit for ressource allocation for a customer. To increase this limit, the customer is set to a higher backplane size (i09, paragraphs 107-108).

Relation to Competing Products

Degree of Integration with other Services (s09)

Depending on the provider's range of products, customers may depend on other products for up- or downstream tasks. Further, even if a company offers end-to-end AIaaS solutions, customers might want to use competing products for parts of their AI pipeline. This behavior is closely related to the customer-side trend of applying a multi-cloud strategy. Therefore, providers must decide which specific competing services they want to integrate. Integration of other products and technologies is especially relevant for data sources (i03, paragraph 98; i10, paragraph 144), and for moving ML models between providers (i09, paragraphs 211, 217, 218). However, one interview partner argued that most customers tend to rely on only one AIaaS provider for most of their tasks. Therefore, employing multiple providers within the same workload "is usually a mini portion of everything you do" (i08, paragraph 36).

Treatment of Downstream Services (s10)

With increasing complexity and diversity of the product portfolios of AIaaS providers, new business models arise *on top* of these AI services. There is a whole industry centered around setting up and deploying cloud services for customers. Here, providers face the question of how to build their relation to those downstream providers (i05, paragraph 54). The spectrum of relations lies between ignoring the downstream services and making them an essential part of the cloud platform. One way of encouraging new business models is to offer an central platform for services around provider's products. Examples include the Microsoft Azure Marketplace and the Google Cloud Platform Marketplace (i01, paragraph 199).

Flexibility of Product Choice (s11)

We refer to flexibility as the different ways a customer can accomplish a task when using a specific AIaaS product or product family. One manifestation is the choice of programming languages when implementing ML models and the required packaging for deploying the models on the provider's infrastructure (i05, paragraph 50). As a provider, using open standards for service deployment and pipeline orchestration is one way of increasing flexibility (i05, paragraph 50, 54). The flexibility of integrating it with other products changes based on the abstraction level of the respective service. End-to-end AIaaS products tend to allow for lower flexibility than single building blocks (i09, paragraph 60, 61).

Publication and Licensing of Products (12)

Companies must choose whether to provide their products as closed or open source software. Open sourcing enables third parties, including customers, to customize the product to their needs. Further, providers can choose to allow community-driven contributions to their open source software. The decision to open-source is strongly influenced by the broader company strategy (i05, paragraph 60).

Ease of Use

Entry Barrier (s13)

A low entry barrier is crucial when it comes to ease of use and acquisition of new customers. Therefore, the time between the customer's decision to use a particular AIaaS product and the actual access to that product

must be kept to a minimum. Optimally, the customer can access the desired products within a few minutes through a fast registration process. For registration, only the most important data is requested, such as the email address and a credit card number, which are required for invoicing (i07, paragraph 65). The contractual conditions are automatically accepted via the terms and conditions opt-in selection. Even though AIaaS products are offered by rather complex price structures such as pay-per-use, the purchasing process hardly differs from a purchase at a retail store on the Internet.

Access to Third Party Applications (s14)

Provision of a marketplace can be another important differentiator in terms of provided ease of use. Marketplaces are mostly operated by large companies and hyperscalers like Google, Microsoft and AWS. By creating an ecosystem, customers are given easy access to many AIaaS products and services that are mostly compatible with the main products of the hyperscalers (i01, paragraph 199). This way small companies and startups have the opportunity to reach customers with their downstream services. The providers in a marketplace are often not tied to just ecosystem. For example, the company *Databricks* offers its services on the Google, AWS, and Microsoft Azure Marketplace.

Ease of Application Creation (s15)

Providing services that are easy to use even for non-experts can be the main differentiator when it comes to mass market adoption of AIaaS products. For example, a customer who has no experience with training ML models can access default algorithms that are easy to apply (i08, paragraph 36). As a provider it is crucial to offer a large amount of these ready-made algorithms to cover as many customer use cases as possible.

Functionalities

A major distinguishing factor between different providers of AIaaS are the variety and offered number of services (i08, paragraph 36) as well as functionalities and features of those services.

Offered Model Size (s16)

The model size determines the amount of parameters and layers in a ML model. There is a recent trend to create models with a large amount of parameters, i.e., millions to billions (Brown et al., 2020; Devlin et al., 2018), based on massive datasets to achieve higher performance. This comes at the cost of longer training time and slower inference as well as higher required computing power. They are more expensive and time costly but offer state of the art results (i06, paragraph 96). Furthermore, larger models often have capabilities to fulfill more than one simple task. For instance, in natural language processing a specified language model can be used to translate from one language into another, while a bigger model with few adaptations could be used for many languages (i06, paragraph 64).

Complexity and Variety of Functionalities (s17)

The number of possible configurations, the workload types, the types of AI capabilities, such as labelling, MLOps, and data augmentation of a service, among others, influence its complexity and variety of functionalities. The variety of offered services is a strong differentiating factor for customers, when choosing a AIaaS provider (i08, paragraph 36). These options can for example change the behavior of the service, the interaction with the service and the way that the service is integrated into the systems of the customer. On the other hand, providers can make a deliberate decision to offer only a reduced or more focused selection of functionalities (i03, paragraph 79). The aim of provided functionalities is to be highly customizable and applicable to even niche tasks. At the same time a large amount of functionalities will generally increase the complexity for all users.

Specialized Methods and General Methods (s18)

Specialized methods are methods that have very niche use cases but are highly optimized for those. In contrast, general methods are methods that are widely applicable to many tasks. One expert noted that more specialized the method generally have better results, but providers usually do not provide solutions for very specific problems (i02, paragraph 118). However, offering specialized solutions for every use case

would be very elaborate and not scalable (i02, paragraph 118). One opportunity for AIaaS providers arises by reusing specialized models from former customer projects and use cases and offering them to other customers.

Explainability and Transparency (s19)

Explainability and Transparency are terms used to subsume methods that provide explanations for the AI services and how the services come to different results. While in many situations a good performing method is favored over explainability and interpretability, recent developments show that customers increasingly care about the reasoning behind an outcome of ML methods. Further, for the use cases of some providers interpretability and explainability of models matter more than the performance (i04, paragraph 47). Providers that focused on other criteria such as performance and ease of use of there services, now see the pent-up demand in this area (i09, paragraph 202).

Trade-Offs in AIaaS Service Design

In this section we highlight the trade-offs derived from the setscrews or mentions in the interviews. Each trade-off can be related to multiple setscrews. For each trade-off, we provide one or more ways to mitigate or alleviate it. The interview partners tended to focus on specific sides of a trade-off. Therefore, we extended their statements by examining other facets of each trade-off.

Security and Reliability vs. Performance

Improving security of the cloud systems increases their reliability. Further, better reliability leads to less down-time, which increases efficiency and performance (i07, paragraph 35).

However, some mechanisms that improve security, such as client-side encryption (s02), render using the data for ML impossible (i01, paragraph 91). Further, encryption and decryption have negative performance implications on the data processing speeds, which can be relevant in high-performance applications. Therefore, to mitigate compatibility issues, AIaaS providers must match the encryption measures with the services the customer aims to use (s02).

Ensuring data replication and availability is expected by customers. However, it does not directly create revenue for the provider and costs precious compute and storage resources of the provider. One solution is for providers to research ways to increase cost-efficiency of the backups. Additionally, providers can make backups subject to charge.

Because in AIaaS the ML methods run at the cloud provider's infrastructure, decrypting data there might compromise the customer's data security. Depending on the legislature in the customer's country of operation, using the cloud provider's services might only be allowed with encryption. While this problem can be mitigated or alleviated when only *storing* data, it becomes critical when using AIaaS. Therefore, not providing resources that comply with the regulations of the customer's legislature potentially costs the provider a whole region of customers. Providers can mitigate these issues by aligning their business practices with the respective legislation or by getting certifications (s03). This might, however, influence the providers' business model negatively. Therefore, the potential loss of customers has to be compared to the loss of adhering to the legislature in question.

Security vs. Ease of Use

Providing multiple accounts for access control (s01) can improve data security and allows for a faster creation of accounts, thereby reducing waiting time for the customer. However, its implementation is not trivial for providers (i09, paragraphs 52-54).

When providers choose to offer close integration of third party services, e.g., via a marketplace, they increase the ease of use for the customer, thereby making the product more attractive (s14). The downside is that such close integration of third party services might decrease security for the customer, because the AIaaS provider cannot audit the third party software and services like his own. To alleviate these issues, providers can, among others, restrict the type of products and services in the marketplace, employ a review system, and require audits for third party applications. For further literature we refer to studies about

security of centralized application-marketplace (Anderson et al., 2010; P. Gilbert et al., 2011; Martin et al., 2017).

Performance vs. Ease of Use

Cloud providers offer no-code or low-code solutions that allow less experienced users to create entire ML pipelines without having to dive deep into programming. One such tool is for example SageMaker Canvas from AWS.These solutions are practical and easy to use for exploration purposes. However, they are not suitable for production, because these pre-configured pipelines tend to be very slow (i10, paragraph 85). To improve execution speed (s05), providers can offer specialized tools for managing the end-to-end ML lifecycle (i10, paragraph 87).

The better the model performance, the more providers can charge customers (i06, paragraph 56). However, measures to increase performance (s06) are strongly associated with effort and costs. Further, offering more complex models requires better explainability methods (i04, paragraph 47). Providers can opt to offer different functionalities (s15,s16) with regard to the size of the model and explainability methods.

Performance vs. Data Locality

Ultra-low latency is important in real-time scenarios. If the data center is far away, latency suffers (i10, paragraph 146). There are also regional constraints that providers must overcome when operating worldwide (i08, paragraph 56). High latency leads to a poor customer experience and a reduction in system performance. Therefore, the location where the service is hosted (s07) is an important decision for a provider. Latency can be reduced by placing solutions close to the customer and then distributing them across the network, so that the respective capacities are available where demand is largest (i02, paragraph 102). Another option is to place servers on premise, which reduces the physical distance to these nodes (i01, paragraph 169).

Performance vs. Scalability

When training a very large model, scalability is important. However, it can be difficult to procure computational power for a continued period of time (i06, paragraph 108). If an AIaaS provider does not host its own services and the resources are not available, the provider has to decide to either compromise the performance of the model or wait longer to train the model. This trade-off is about the ability to scale (s08). Cloud providers can increase the resource allocation limit for a customer by, for example, granting the customer a higher backplane size (i09, paragraphs 108-110).

Openness vs. Revenue

As a provider, opening up your services to competing products (s09, s10) is a double-edged sword. On one hand, easy integration with other services and flexibility about which services and technologies to use might increase customer satisfaction and therefore revenue. For customers, this non-exclusivity is a great advantage, since they do not have to switch the ecosystem because of the absence individual services. On the other hand, one interview partner mentioned that providers might also lose revenue by opening up their services, because customers might choose more competing products (i05, paragraph 54). As there seems to be no one-size-fits-all solution regarding the best way to handle this trade-off, we refer to it as a strategic business decision an AIaaS provider must make.

Functionality vs. Ease of Use

One factor for ease of use is simplicity. Providing many functionalities (s17) inherently increases complexity. This increased complexity results in too many additional decisions for a user to effectively use the service. A main distinguishing factor between different AIaaS providers is the variety and number of services (i08, paragraph 36), thus for a provider it is necessary to provide a larger number of functionalities. Integrating these functionalities into the provided service is a challenge that needs to be addressed by clear communication, e.g., through the user interface or the documentation. In contrast, a provider may choose to provide a reduced and focused set of functionalities. This greatly benefits the usability for even inexperienced users. But a service with too few functions can quickly become impractical and inefficient

(i03, paragraph 20; paragraph 53). By investing into abstracting complexity from the user, AIaaS providers can get the best of both worlds. Naturally, this requires large efforts and is in part an open research task.

A common way to offer such focused services is called *managed service*, where a clear use for the service is provided and the user must do little setup before running it (i05, paragraph 74; i08, paragraph 16-20; i09, paragraph 61-63). This allows non-experts to create applications for their specific needs (s15). Here, for the provider one challenge is the balance of functionality and ease of use. A focused subset of options appeals to a specific user group, while other user groups are deterred by the lack of options. To appeal to a wide range of possible customers a good mix of general and specified methods is necessary (s18).

Conclusion

In this work we conducted ten expert interviews and identified setscrews and trade-offs when designing AI as a Service products. For the setscrews we identified five groups of aspects that influence the decision making during AI service design: reliability and security, performance, relation to competing products, ease of use and functionality. Based on these groups, we discussed trade-offs that result from the interaction of contrary effects of AIaaS product designs, thereby highlighting the viewpoint of the provider. Interestingly, even for the experts, a strict demarcation of provider specific trade-offs turned out to be difficult.

We found out that there is a strong overlap between conventional cloud computing services and AIaaS. Many of the same service design factors apply directly. Further, the challenges and opportunities provided by the use of AI such as explainability, fairness, data quality, were often discussed during the interviews. However, in addition to the challenges faced in cloud computing, new challenges arise with the emergence of cloud-based AI. For instance, AIaaS providers now face the new dimension of demand peaks for computing power that is introduced by deep learning. This demand usually can only be fulfilled by using server clusters. This renders cloud-based AI the only feasible solution for tasks above a certain magnitude. Therefore, it is likely that AIaaS will become the standard way of conducting AI efforts by businesses.

There are several open direction for future research. First, this study focuses on the provider side of AIaaS. However, customer needs strongly influence service design as well. Therefore, we would like to investigate the customer point of view in future work. Second, as the regulation of AI becomes more important, researching its implications on AI service design might become necessary. Lastly, synthesizing research into drivers of AIaaS adoption with our findings regarding AI service design poses a promising research avenue.

References

Adams, W. C. (2015). Conducting Semi-Structured Interviews. In *Handbook of Practical Program Evaluation* (pp. 492–505). John Wiley & Sons, Ltd.

Anderson, J., Bonneau, J., & Stajano, F. (2010). Inglorious Installers: Security in the Application Marketplace. *9th Annual Workshop on the Economics of Information Security, WEIS 2010, Harvard University, Cambridge, MA, USA, June 7-8, 2010*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*.

Gesetz über das Bundesamt für Sicherheit in der Informationstechnik, (2015). Sect. 2 para. 10 cl. 1. In German.

Gilbert, P., Chun, B.-G., Cox, L. P., & Jung, J. (2011). Vision: Automated Security Validation of Mobile Apps at App Markets. In *Proceedings of the Second International Workshop on Mobile Cloud Computing and Services*, 21–26.

Gilbert, S., & Lynch, N. (2002). Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services. *SIGACT News*, *33*(2), 51–59.

Kaiser, R. (2021). *Qualitative Experteninterviews: Konzeptionelle Grundlagen und praktische Durchführung*. Springer Fachmedien Wiesbaden. In German.

Kodiyan, A. A. (2019). An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. *Researchgate Preprint*.

Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2021). Artificial Intelligence as a Service. *Business & Information Systems Engineering*, *63*(4), 441–456.

Loukides, M. (2022). *AI Adoption in the Enterprise 2022*. O'Rilley. https://www.oreilly.com/radar/ai-adoption-in-the-enterprise-2022/

Martin, W., Sarro, F., Jia, Y., Zhang, Y., & Harman, M. (2017). A Survey of App Store Analysis for Software Engineering. *IEEE Transactions on Software Engineering*, *43*(9), 817–847.

Mell, P., & Grance, T. (2011). The NIST Definition of Cloud Computing. *NIST Special Publication 800-145*, 1–7.

Moriarty, J. (2011). *Qualitative Methods Overview*. National Institute for Health Research School for Social Care.

Myers, M. D. (2019). *Qualitative research in business and management*. Sage.

Pandl, K. D., Teigeler, H., Lins, S., Thiebes, S., & Sunyaev, A. (2021). Drivers and inhibitors for organizations' intention to adopt artificial intelligence as a service. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1769.

Statista. (2021). *Umfrage zur Nutzung von Cloud Computing in deutschen Unternehmen bis 2020*. Statista. https://de.statista.com/statistik/daten/studie/177484/umfrage/einsatz-von-cloud-computing-in-deutschen-unternehmen-2011/.

Statista. (2022). *Revenues from the artificial intelligence (AI) software market worldwide from 2018 to 2025*. Statista. https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/

Continuous Learning and Applying

Selected Issues in Critical Information Infrastructures, Winter Term 21/22

Eva Vanessa Appenzeller
Bachelor Student
Karlsruhe Institute of Technology
eva.appenzeller@student.kit.edu

Dominik Martus
Master Student
Karlsruhe Institute of Technology
dominik.martus@student.kit.edu

Elif Yüsra Özcelik
Master Student
Karlsruhe Institute of Technology
elif.oezcelik@student.kit.edu

Betül Özdemir
Master Student
Karlsruhe Institute of Technology
betuel.oezdemir@student.kit.edu

Kerem Okay
Master Student
Karlsruhe Institute of Technology
kerem.okay@student.kit.edu

Abstract

Background: *Lecturer-centered teaching is common at universities. Here, the focus is on theoretical information transfer by the lecturer and there is little interaction between the students. Students have no incentive to continuously engage with the course content due to a single performance test at the end of the semester. In addition, the lack of practical references makes it difficult for students to apply the taught content in real-life situations.*

Objective: *The aim of this thesis is therefore to develop a teaching concept that supports continuous learning and application. The individual learning speed of the students and the interaction between the students should be promoted and thus a student-centered teaching should be achieved.*

Methods: *For this purpose, the teaching concept "Continuous Learning and Applying" (KoLA) was developed iteratively on the basis of Flipped Classroom. The iterative improvement and design of the teaching concept is based on literature research and brainstorming sessions.*

Results: *At its core, KoLA is composed of alternating in-class and out-of-class parts, and the individual components were specified as part of this work.*

Conclusion: *By shifting the delivery of information outside of class time, learning is made possible regardless of time or location, and individual learning pace is encouraged. In addition, student-centeredness is achieved by focusing the face-to-face sessions on the implementation of interactive learning activities. Since KoLA is suitable for a small group of participants, a generally applicable teaching concept should be further developed on the basis of this work.*

Keywords: flipped classroom, in-class, out-of-class, teaching, learning, applying, practice, course, instruction, lecture, student, satisfaction, skills, problem-solving, online

Zusammenfassung

Die dozierendenzentrierte Lehre ist an Universitäten geläufig. Dabei steht die theoretische Informationsvermittlung durch den Lehrenden im Vordergrund und es findet nur wenig Interaktion zwischen den Studierenden statt. Studierende haben aufgrund einer einzelnen Leistungsprüfung am Ende des Semesters, keinen Anreiz sich kontinuierlich mit den Lehrinhalten auseinanderzusetzen. Zudem erschweren fehlende Praxisbezüge den Studierenden die vermittelten Inhalte in realen Anwendungsfällen anzuwenden. Ziel der vorliegenden Arbeit ist es deshalb, ein Lehrkonzept zu erarbeiten, welches das kontinuierliche Lernen und Anwenden unterstützt. Dabei soll das individuelle Lerntempo der Studierenden und die Interaktion zwischen den Studierenden gefördert werden und somit eine studierendenzentrierte Lehre erreicht werden. Hierfür wurde auf Grundlage von Flipped Classroom iterativ das Lehrkonzept „Kontinuierliches Lernen und Anwenden“ (KoLA) entwickelt. Die iterative Verbesserung sowie Ausgestaltung des Lehrkonzeptes stützen sich auf Literaturrecherchen und Brainstorming-Sessions. KoLA setzt sich im Kern aus abwechselnden In- und Out-of-Class-Teilen zusammen, wobei die einzelnen Bestandteile im Rahmen dieser Arbeit konkretisiert wurden. Durch die Verlagerung der Informationsvermittlung außerhalb der Präsenzzeit wird ein zeit- und ortsunabhängiges Lernen ermöglicht und das individuelle Lerntempo gefördert. Außerdem wird eine Studierendenzentrierung erreicht, indem der Fokus der Präsenztermine auf der Durchführung interaktiver Lernaktivitäten gelegt wird. Da KoLA für eine kleine Teilnehmergruppe geeignet ist, sollte auf Basis dieser Arbeit ein allgemeingültiges Lehrkonzept weiterentwickelt werden.

Einleitung

Problemstellung

An Universitäten ist die dozierendenzentrierte Lehre die häufigste Lehrform, welche dem Lehrenden den größeren Redeanteil überlässt (Winteler, 2011). Aufgrund der mangelnden Interaktion erhalten die Lehrenden kein Feedback zu den übermittelten Lehrinhalten und Studierende nehmen über 90 Minuten eine passive Rolle ein. Infolgedessen nimmt die Aufnahmefähigkeit der Studierenden wegen sinkender Konzentration ab (Winteler, 2011). Hinzu kommt, dass bei der dozierendenzentrierten Lehre der Fokus auf der theoretischen Informationsvermittlung liegt und kaum Praxisanwendungen durchgeführt werden. Dabei könnten mithilfe von Praxisanwendungen verbesserte Problemlösekompetenzen erzielt werden (Kern, 2002).

Darüber hinaus kennzeichnen sich Vorlesungen der dozierendenzentrierten Lehre durch feste Vorlesungstermine und Veranstaltungsorte, wodurch Studierende zeitlich und örtlich gebunden sind. Da Studierende unterschiedliche biologische Rhythmen aufweisen, eignet sich der vorgegebene Vorlesungstermin nicht für den optimalen Lernerfolg jedes Studierenden. Beispielsweise ist die Leistung von Studierenden mit frühem Biorhythmus am Morgen höher (Goldin et al., 2020). Zudem kollidieren andere Verpflichtungen, wie feste Arbeitszeiten, mit Lehrveranstaltungen. Wenn keine Vorlesungsaufzeichnungen zur Verfügung gestellt werden, besteht keine Möglichkeit, verpasste Vorlesungstermine oder bei Unklarheiten die Erläuterungen des Lehrenden zu wiederholen. Entsprechend können Studierende die Lehrinhalte nicht zeitnah nacharbeiten. So geraten sie in Rückstand und befassen sich nicht kontinuierlich mit den Lehrinhalten.

Ein weiteres Problem stellt die Prüfungsleistung dar, denn diese setzt sich meist aus einer einzelnen Note zusammen. Diese Momentaufnahme könnte, bedingt durch äußere Gegebenheiten wie Nervosität oder privaten Umständen, wenig aussagekräftig über das tatsächliche Leistungspotenzial der Studierenden sein. Die Prüfungsleistung wird erst am Ende des Semesters erbracht und unter dem Semester erfolgt keine Rückmeldung zum persönlichen Lernstand. Deshalb prokrastinieren viele Studierende das Lernen auf die letzten Wochen vor der Prüfung (Dehling, Roegner & Winzker, 2014). Dabei führt diskontinuierliches Lernen zu einem schlechteren Beibehalten der Lehrinhalte im Gedächtnis (LaTour & Noel, 2021).

Diese beispielhaften Probleme der Lehre motivieren folgende Forschungsfrage: Was sind mögliche Lösungsansätze, um Studierende zum kontinuierlichen Lernen zu motivieren und um Gelerntes praxisnah

anzuwenden? Als einen möglichen Lösungsansatz bieten bestehende Forschungsergebnisse den Flipped Classroom an (Akçayır & Akçayır, 2018; Strelan, Osborn & Palmer, 2020). Diese Forschungsergebnisse beziehen sich meist nur auf einen Teilaspekt des Flipped Classroom verglichen mit der traditionellen Lehre, wie z.B. der Zufriedenheit unter Studierenden (Strelan, Osborn & Palmer, 2019), möglichen Vorteilen und Herausforderungen (Akçayır & Akçayır, 2018) oder der Leistung der Studierenden (Strelan, Osborn & Palmer, 2020). Bisher fehlt es aber noch an der Entwicklung eines Lehrkonzeptes, welches die oben ausgearbeitete Problemstellungen adressiert und die einzelnen Bestandteile des Flipped Classroom darauf abstimmt. Aufgrund der mangelnden Übertragbarkeit der bestehenden Forschungsergebnisse ist weitere Forschung notwendig. Hierbei gilt es, das kontinuierliche Lernen von Studierenden und eine Praxisanwendung zu fördern und in einem ganzheitlich ausgearbeiteten Lehrkonzept umzusetzen. Diese Arbeit besitzt deshalb den Anspruch, ein Lehrkonzept zur Verminderung der aufgeführten Probleme zu entwickeln.

Zielsetzung

Die Lernmotivation und die Aufnahmefähigkeit der Studierenden kann durch ein abgestimmtes Lehrkonzept gefördert werden. Das übergeordnete Ziel dieser Arbeit ist die Entwicklung eines Lehrkonzeptes, das Studierende motiviert, kontinuierlich zu lernen und das Gelernte mit konkreten Praxisanwendungen zu vertiefen. Das kontinuierliche Lernen sowie die Praxisanwendung der Lehrinhalte unterstützen Studierende dabei, sich regelmäßig mit den Lehrinhalten auseinanderzusetzen und somit langfristig Wissen aufzubauen. Dafür verfolgt das Lehrkonzept das Teilziel, das individuelle Lerntempo der Studierenden zu fördern.

Da die Vorlesungen oft in Präsenz gehalten werden, haben die Studierenden keine Möglichkeit, sich die Vorlesungen zeitunabhängig und im eigenen Lerntempo anzueignen. Das kann dazu führen, dass die Motivation zum Lernen nachlässt, weil man beispielsweise der Vorlesung nicht folgen kann oder einige Inhalte nicht verstanden werden.

Des Weiteren soll die Interaktion zwischen den Studierenden gefördert werden. Das hat zur Folge, dass der Redeanteil des Lehrenden verringert wird. Die Präsenzzeit kann so dafür genutzt werden, dass sich die Studierenden interaktiv über den Vorlesungsinhalt austauschen und voneinander lernen können. So werden auch Softskills, wie beispielsweise die Fähigkeit zur Gruppenarbeit und die soziale Interaktion, gefördert. Hierfür wird im Folgenden das Lehrkonzept „Kontinuierliches Lernen und Anwenden“ (KoLA) entwickelt. Dabei sollen die einzelnen Bestandteile des Lehrkonzeptes die oben genannten Ziele adressieren.

Aufbau der Arbeit

Die vorliegende Seminararbeit gliedert sich in sieben Abschnitte. Nach der Einleitung widmet sich der Abschnitt „Einführung in das Konzept Flipped Classroom“ den theoretischen Grundlagen dieser Arbeit. Dabei wird das Konzept des Flipped Classroom vorgestellt, von dem ausgehend ein Lehrkonzept weiterentwickelt wird. Im Fokus des Abschnitts „Entwicklung des Lehrkonzeptes KoLA“ steht die Vorgehensweise zur Erstellung des Lehrkonzeptes. Insbesondere werden die unternommenen Entwicklungsschritte erläutert. Im nächsten Abschnitt „Überblick über das Lehrkonzept KoLA“ folgt ein Überblick über das Lehrkonzept, wobei auf notwendige Rahmenbedingungen verwiesen und ein schematischer Ablauf präsentiert wird. Der folgende Abschnitt „Detaillierte Beschreibung der Bestandteile“ beschreibt die einzelnen Bestandteile des Lehrkonzeptes und gibt deren Merkmale wieder. Im Zuge dessen wird die Umsetzung der Bewerbungsphase und Einführungsveranstaltung konkretisiert, Aktivitäten für Out-of-Class und In-Class besprochen sowie die Form der Prüfung und Zusammensetzung der Note diskutiert. Auf die Grenzen und Herausforderungen des Lehrkonzepts wird im Abschnitt „Grenzen und Alternativen“ eingegangen, welches zugleich alternative Ausgestaltungsmöglichkeiten der Bestandteile aufzeigt. Die Arbeit wird mit einer Zusammenfassung, einer kritischen Würdigung sowie einem Ausblick im Abschnitt „Fazit“ abgeschlossen.

Einführung in das Konzept Flipped Classroom

Flipped Classroom wird als ein pädagogisches Lehrkonzept verstanden, bei dem ein Wechsel von einer dozierendenzentrierten zu einer studierendenzentrierten Lehre stattfindet (Bergmann & Sams, 2012). Im Vergleich zur traditionellen Lehre, bei der die Lehrinhalte vor Ort durch den Lehrenden präsentiert und

erklärt werden, findet bei Flipped Classroom die Informationsvermittlung außerhalb der Präsenzzeit (Out-of-Class) statt (Strelan, Osborn & Palmer, 2019). Dabei eignen sich die Studierenden ortsunabhängig und im eigenen Lerntempo die Lehrinhalte an (McDonald & Smith, 2013). Hierfür werden verschiedene Möglichkeiten zum Lernen der Lehrinhalte seitens des Lehrenden zur Verfügung gestellt, wie z.B. durch Videos (Akçayır & Akçayır, 2018). Der Austausch unter den Studierenden außerhalb der Präsenzzeit kann durch verschiedene Technologien, beispielsweise mithilfe von Diskussionsforen, gefördert werden.

Die Präsenzzeit (In-Class) wird genutzt, um das eigenständig angeeignete Wissen anhand von Praxisanwendungen aktiv und gemeinsam mit anderen Kommilitonen zu vertiefen (Topping & Ehly, 1998). Der Fokus liegt während der Präsenzzeit auf dem Austausch und dem gemeinschaftlichen Lernen unter den Studierenden. Hierbei nimmt der Lehrende die Rolle des Lernbegleiters ein, der für aufkommende Fragen und Unklarheiten den Studierenden zur Seite steht. Damit die Studierende in den In-Class-Aktivitäten diskutieren und diverse Problemstellungen gemeinsam lösen können, wird vorausgesetzt, dass sie sich vorab regelmäßig mit den Lehrinhalten Out-of-Class auseinandersetzen und vorbereitet sind (Al-Zahrani, 2015).

Entwicklung des Lehrkonzeptes KoLA

Zur Entwicklung eines Lehrkonzeptes wurden zunächst die Probleme der dozierendenzentrierten Lehre mithilfe von Brainstorming zusammengetragen. Basierend auf der Wahrnehmung von Studierenden wurden die Kernprobleme, die im Kapitel „Problemstellung“ erläutert wurden, identifiziert. Aus den identifizierten Kernproblemen wurden Verbesserungspotenziale ausgearbeitet. Daraus wurden die im Kapitel „Zielsetzung“ vorgestellten Haupt- und Teilziele für das zu entwickelnde Lehrkonzept abgeleitet. Aufgrund des Fokus auf das kontinuierliche Lernen und Anwenden wurde der Name des neuen Lehrkonzepts als Akronym „KoLA“ etabliert. Da die Teilziele durch die studierendenzentrierte Ausrichtung des Konzeptes Flipped Classroom adressiert werden, wurde Flipped Classroom, welches im Rahmen des Seminarkurses „Selected Issues in Critical Information Infrastructures” vorgestellt wurde, als Grundlage für das neu zu entwickelnde Lehrkonzept festgelegt. Anschließend wurden Literaturrecherchen in Google Scholar mit dem Begriff „Flipped Classroom“ durchgeführt. Dies verhalf zu einer besseren Übersicht über die Bestandteile und Ziele, auf die „Flipped Classroom“ eingeht.

Auf Basis vorhandener Forschungsergebnisse wurde iterativ ein Grobkonzept ausgearbeitet. Im Grobkonzept waren zunächst bewertete Online-Tests vorgesehen. Aufgrund des Feedbacks zum hohen Manipulationsrisiko im Online-Format wurde das Lehrkonzept dahingehend angepasst, dass die Tests in Präsenz stattfinden. Weiterhin wurden mithilfe vertiefter Literaturrecherchen zu „Out-of-Class“ und „In-Class“ die Bestandteile des Lehrkonzepts verfeinert. Hierbei hat sich das Feedback an Studierende als wichtigen Bestandteil von In-Class herausgestellt, wodurch am Ende jedes In-Class-Termins eine Feedbackrunde eingebunden wurde.

Das verfeinerte Lehrkonzept wurde in einer Zwischenpräsentation vorgestellt. Für die abschließende Leistungsprüfung war zunächst eine mündliche Prüfungsform vorgesehen. Jedoch präferierten die bei der Zwischenpräsentation anwesenden Studierenden eine schriftliche Prüfung gegenüber einer mündlichen Prüfung. Daraufhin wurden in dem Lehrkonzept beide Prüfungsformen berücksichtigt. Außerdem wurde Rückmeldung bezüglich der Gewichtung der Leistungsprüfungen gegeben. Die Gewichtung, die von der Mehrheit der Anwesenden unterstützt wurde, wurde in das Lehrkonzept eingearbeitet.

Mithilfe der vorherigen Schritte wurde eine Lehrprobe aufgebaut, in welcher der Fokus auf der Umsetzung des In-Class-Teils lag. Dies wurde dadurch begründet, dass die Hauptvorteile des Flipped Classroom durch die interaktiven Lernaktivitäten vor Ort zustande kommen (Jensen, Kummer & Godoy, 2015). Hierfür wurde ein beispielhafter Präsenztest vorgestellt und eine exemplarische Gruppenarbeit durchgeführt. Dabei wurde darauf aufmerksam gemacht, den Präsenztest und die Gruppenarbeit thematisch gut aufeinander abzustimmen. Ansonsten besteht die Gefahr, dass sich die Studierenden nur auf den Bestandteil des In-Class-Teils vorbereiten, der höher in der Benotung gewichtet wird. Zudem wurde bestätigt, dass durch die Präsenztests das Lernen in kleinere Abschnitte aufgeteilt wird, wodurch ein Anreiz zum kontinuierlichen Lernen besteht. Abschließend wurden offene Lücken mit Literaturrecherche nach gezielten Begriffen geschlossen, unter anderem wie die Gruppeneinteilung In-Class erfolgt. Anhand der durchgeführten Schritte konnte das Lehrkonzept iterativ weiterentwickelt werden. Hieraus resultierte das

Lehrkonzept, welches einen Ansatz darstellt, die ausgearbeiteten Probleme der dozierendenzentrierten Lehre zu lösen.

Überblick über das Lehrkonzept KoLA

Rahmenbedingungen

Das Lehrkonzept KoLA basiert auf dem Flipped Classroom, weswegen die Veranstaltung einen Out-of-Class-Teil in Remote und einen In-Class-Teil in Präsenz enthält. Die Teilnehmerzahl wird auf 20-30 Personen begrenzt, um kleine Gruppen mit maximal fünf Studierenden im In-Class-Teil zu realisieren und Betreuer für je zwei Gruppen bereitzustellen. Kleine Gruppen bieten für den In-Class-Teil diverse Vorteile, auf die im Kapitel „In-Class“ eingegangen wird. Bei einer größeren Teilnehmerzahl fällt zusätzlicher Personalaufwand an und weitere Räumlichkeiten werden benötigt. Wiederum kann die Gruppenarbeit nur zustande kommen, wenn genügend Teilnehmer angemeldet sind, d.h. es werden mindestens Studierende für ein bis zwei Gruppen benötigt. KoLA eignet sich unter diesen Rahmenbedingungen vor allem für kleinere Veranstaltungen mit Seminar- bzw. Praktikumscharakter.

Übersicht Zentraler Bestandteile & Ablauf

Aufgrund der Teilnehmerbegrenzung beginnt die Durchführung des Lehrkonzeptes mit einer Bewerbungsphase zur Auswahl der Studierenden (siehe Abbildung 1). Nachdem die Bewerbungsphase abgeschlossen wurde, werden die Teilnehmer zur Einführungsveranstaltung eingeladen. Dabei liegt der Fokus auf dem gegenseitigen Kennenlernen, der Vorstellung des Lehrkonzeptes, der Darstellung der Leistungsprüfung sowie einer Einführung in die Lehrveranstaltung.

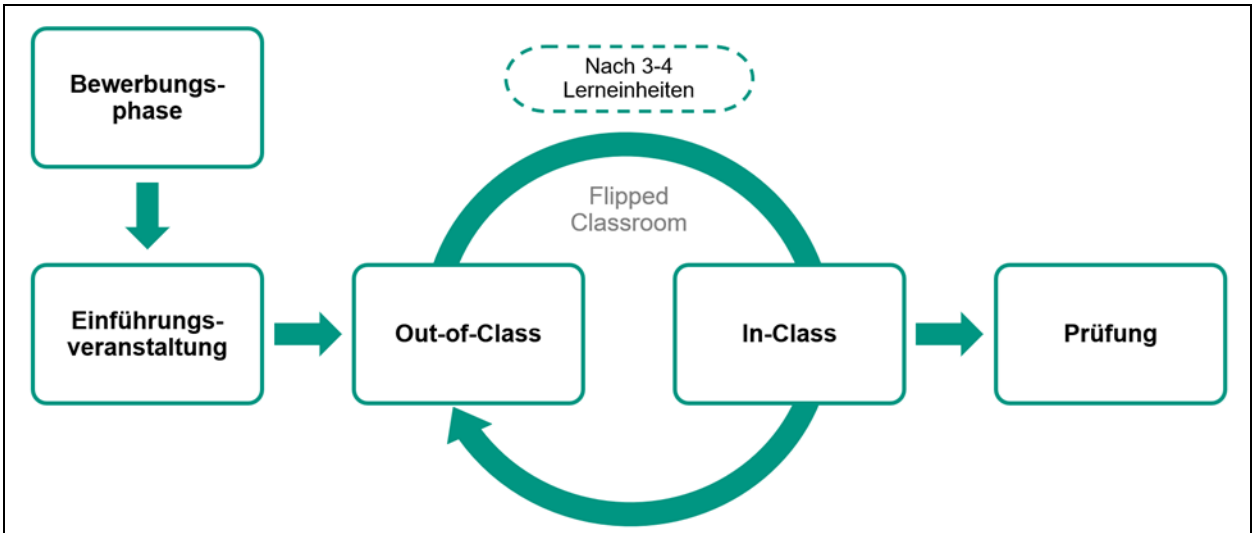

Abbildung 1. Ablauf einer Lehrveranstaltung nach dem Lehrkonzept KoLA

Unter dem Semester wechseln sich Out-of-Class- und In-Class-Teile ab. Zu Beginn eignen sich die Studierenden selbstständig die Lehrinhalte Out-of-Class an. Dazu wird eine Plattform, genutzt, auf dem alle Lehrinhalte On-Demand abrufbar sind. Anschließend haben die Studierenden die Möglichkeit, über auf der Plattform bereitgestellte Online-Tests ihren Lernstand zu ermitteln. Fragen und Verständnisprobleme können in einem Forum oder Chat gestellt werden, sodass sowohl Lehrende als auch andere Studierende antworten können. Nach drei bis vier Lerneinheiten Out-of-Class, wobei eine Lerneinheit dem Umfang einer Vorlesungswoche entspricht, findet jeweils ein In-Class-Termin in Präsenz statt. Zu Beginn des In-Class-Termins führen die Studierenden einen benoteten Präsenztest zur Leistungsprüfung durch. Nachfolgend wird das Gelernte in einer benoteten Gruppenarbeit praktisch angewendet und dadurch ein Austausch zwischen den Studierenden angeregt. Der Wechsel zwischen Out-of-Class und In-Class wiederholt sich in regelmäßigen Abständen bis zum Ende des Vorlesungszeitraumes,

sodass insgesamt vier In-Class-Termine vorgesehen sind. Von diesen vier Terminen sind drei verpflichtend wahrzunehmen. Abschließend nehmen die Studierenden an einer Prüfung teil, damit das Gesamtverständnis sowie die Zusammenhänge zwischen den Lehrinhalten überprüft wird.

Detaillierte Beschreibung der Bestandteile

Bewerbungsphase

Vor Beginn des Semesters findet eine Bewerbungsphase statt. Das Karlsruher Institut für Technologie (KIT) stellt für Bewerbungen den "YouSubscribe"-Dienst über das Wiwi-Portal bereit, welches für das Lehrkonzept KoLA verwendet werden kann. Der Dienst bietet die Möglichkeit, den Studierenden einen Überblick über die Lehrveranstaltung zu geben. Dadurch werden Studierende vorab mit den Inhalten und dem Ablauf vertraut gemacht. Weiterhin werden die Präsenztermine für die Einführungsveranstaltung und für die In-Class-Teile angegeben, damit alle Teilnehmenden Planungssicherheit haben und Lehrende frühzeitig Räumlichkeiten buchen können. Als Teilnehmerkreis eignen sich besonders Masterstudierende oder Bachelorstudierende in Vertiefungsveranstaltungen, da diese bereits über grundlegende Fachkenntnisse verfügen und diese vertiefen können. Die Studierenden bewerben sich fristgerecht mit einem Motivationsschreiben und einem aktuellen Notenauszug. Für die Auswahl der Teilnehmenden spielt die Motivation der Studierenden eine zentrale Rolle. Daneben kann im Motivationsschreiben auf erlangte Fachkenntnisse durch Praxiserfahrungen eingegangen werden. Sofern Vorkenntnisse für die Lehrveranstaltung notwendig sind oder weitere Kriterien durch den Lehrenden festgelegt wurden, wird darauf in YouSubscribe hingewiesen. Der Notenauszug dient als Nachweis für die Belegung notwendiger Lehrveranstaltungen.

Einführungsveranstaltung

Zu Beginn des Semesters wird nach dem Lehrkonzept KoLA eine Einführungsveranstaltung angesetzt. Diese Einführungsveranstaltung dient der Vorstellung von Lehrinhalten und dem Aufbau der Lehrveranstaltung. Da die Lehrinhalte je nach Lehrveranstaltung variieren können, wird auf diese an dieser Stelle nicht weiter eingegangen. Dabei wird insbesondere der Ablauf der In- und Out-of-Class-Teile, deren Termine sowie die Leistungsprüfungen, die im Kapitel „Prüfungsleistung bei KoLA" ausgeführt werden, erläutert.

Neben der Kommunikation der organisatorischen Rahmenbedingungen verfolgt die Einführungsveranstaltung das Ziel des gegenseitigen Kennenlernens. Dies erfolgt in einer Vorstellungsrunde, in der sich alle Teilnehmenden sowie Lehrenden vorstellen. Zur Durchführung der Einführungsveranstaltung eignet sich ein Präsenztermin. Da die Studierenden sich außerhalb der Veranstaltungszeit, beispielweise in Pausen, austauschen können, wird die Interaktion gefördert und der Teamgeist gestärkt.

Alternativ kann die Einführungsveranstaltung digital über eine Plattform wie Zoom oder Microsoft Teams gehalten werden, um Studierenden eine ortsunabhängige Veranstaltung zu ermöglichen. Falls die Räumlichkeit vor Ort eine digitale Zuschaltung von Studierenden ermöglicht, beispielsweise über Beamer und Kamera, kann als eine weitere Möglichkeit eine Hybridveranstaltung angeboten werden.

Nach der Einführungsveranstaltung haben die Studierenden die Möglichkeit, innerhalb einer kommunizierten Frist von der Lehrveranstaltung zurückzutreten. Sofern Abmeldungen erfolgen, können Studierende aus der Warteliste letztmalig der Lehrveranstaltung beitreten, sodass sie rechtzeitig integriert werden können. Im Falle von Nachrückern kann eine erneute kurze Vorstellungsrunde im ersten In-Class-Termin angesetzt werden, um das Zusammengehörigkeitsgefühl zum Team weiterhin zu stärken.

Out-of-Class

Informationsvermittlung

Der Out-of-Class-Teil dient der Vermittlung des Lehrinhaltes (siehe Abbildung 2). Hierzu werden die Inhalte den Studierenden mithilfe von Skripten und dazugehörigen Videos bereitgestellt. Dadurch können die Inhalte jederzeit erarbeitet und wiederholt werden, um Verständnislücken zu schließen (Ulrich, 2016).

Studierende präferieren dabei das Anschauen von Videos gegenüber dem Lesen von Texten (Akçayır & Akçayır, 2018). Diese Präferenz wird dadurch begründet, dass mithilfe von Videos mehrere Sinneskanäle angesprochen werden. Durch das Lesen werden Informationen nur visuell aufgenommen, während Videos zusätzlichen auditiven Input ermöglichen (Dorgerloh & Wolf, 2020). Wenn Studierende Informationen gleichzeitig auditiv und visuell im Gedächtnis speichern, können sie sich besser an diese erinnern (Dorgerloh & Wolf, 2020). Ein weiterer Vorteil der Videos besteht darin, dass die Lehrinhalte im individuellen Lerntempo bearbeitet werden. Videos können jederzeit pausiert, zurück- oder vorgespult werden. Zur Gewährleistung des Lernerfolgs sind die Videolängen ein ausschlaggebendes Kriterium. Daher empfiehlt sich eine Videolänge von maximal 20 Minuten (Akçayır & Akçayır, 2018).

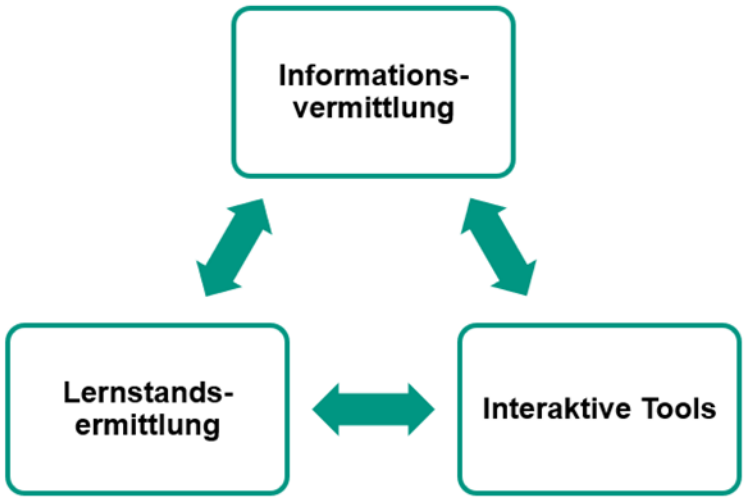

Abbildung 2. Idealtypischer Ablauf der Out-of-Class-Teile

Lernstandsermittlung

Die Lernstandsermittlung bietet Studierenden die Möglichkeit, sich auf den Präsenztest vorzubereiten, ihren Lernprozess zu evaluieren und Wissenslücken zu erkennen (FNL, 2014; Hamdan & McKnight, 2013).

KoLA sieht hierfür freiwillige Online-Tests vor, die von den Lehrenden kapitelweise zur Verfügung gestellt und nicht benotet werden. Ergänzend zu den Online-Tests kann eine (pseudonymisierte) Rangliste zum Vergleich der Studierenden untereinander eingeführt werden. Ein alternatives Format zum Online-Test sind Aufgabenblätter mit Lösungen, die die Studierenden selbst auswerten. Die Online-Tests fragen, beispielsweise in Form von Multiple-Choice-Fragen, grundlegende Lehrinhalte ab. Nach der Durchführung eines Online-Tests werden die Ergebnisse automatisch ausgewertet und die Studierende erhalten eine Rückmeldung zu ihrem persönlichen Lernstand. Bei Bedarf können entsprechende Inhalte wiederholt und vertieft werden (Strelan, Osborn & Palmer, 2020).

Interaktive Tools

Eine zentrale Lernplattform ermöglicht den Studierenden, alle Inhalte des Out-of-Class an einem Ort abzurufen. Als essenzielle Funktion sollen auf dieser Plattform Lehrinhalte in Form von diversen Medienformaten eingebunden werden können. Außerdem soll die Durchführbarkeit von Online-Tests gewährleistet werden. Um Interaktion Out-of-Class zu ermöglichen, sind interaktive Tools notwendig. Darüber können Fragen, die während des Out-of-Class-Teils aufkommen, diskutiert werden. Lehrende und Studierende können sich zeitlich unabhängig, z.B. über ein Diskussionsforum oder eine Chat-Funktion, austauschen (Akçayır & Akçayır, 2018). Für einen unmittelbaren Dialog bieten sich zusätzlich Online-Sprechstunden an. In einem gemeinsamen Gespräch können Fragen sowie potentielle Rückfragen geklärt werden. Lehrende erhalten über die Menge der gestellten Fragen indirekt Feedback, inwieweit die Lehrinhalte erfolgreich vermittelt wurden. Bei Bedarf kann die Ausgestaltung der entsprechenden Lehrinhalte iterativ verbessert werden.

Im Rahmen dieser Arbeit wurden zwei Lernplattformen verglichen: Integriertes Lern-, Informations- und Arbeitskooperations-System (ILIAS) und Microsoft Teams (MS Teams). Beide Lernplattformen ermöglichen das Einbinden unterschiedlicher Medienformate. Die Gestaltungsfunktion für Online-Tests ist bei ILIAS im Vergleich zu MS Teams deutlich umfangreicher, wobei beide Lernplattformen eine automatisierte Auswertung der Online-Tests anbieten. Im Bereich interaktive Tools stellt ILIAS Diskussionsforen zur Verfügung, sodass Themen gezielt gruppiert und diskutiert werden. Im Gegensatz dazu können bei MS Teams Chat-Nachrichten geschrieben, Audio- und Videokonferenzen abgehalten und dabei der eigene Bildschirm mit anderen Personen geteilt werden. In den weiteren Funktionen sind MS Teams und ILIAS vergleichbar (ZML, 2020). Es lässt sich festhalten, dass die Stärken von MS Teams in den kollaborativen Elementen und in einer synchronen Zusammenarbeit liegen (ZML, 2020). MS Teams fördert insbesondere den Austausch zwischen Studierenden und Lehrkräften. Dagegen hebt sich ILIAS durch die strukturierte Organisation des generellen Ablaufs der Lehrveranstaltung hervor (ZML, 2020). Unter anderem können Übungsaufgaben verwaltet und Gruppen zum Austausch zwischen Studierenden und Lehrenden eröffnet werden (ZML, 2020). Da für KoLA die Online-Tests von zentraler Bedeutung sind, ist ILIAS in diesem Kontext vorzuziehen.

In-Class

Wie im Kapitel „Übersicht zentraler Bestandteile & Ablauf“ erläutert, finden in regelmäßigen Abständen In-Class-Termine in Präsenz statt. Dabei gliedert sich der Ablauf eines In-Class-Termins in fünf Phasen (siehe Abbildung 3): Präsenztest, Austausch, Wiederholung und Aufgabenstellung, Gruppenarbeit sowie Abschluss.

Abbildung 3. Ablauf der In-Class-Teile

Präsenztest

Zu Beginn jedes In-Class-Termins findet ein schriftlicher Präsenztest statt, der jeweils die neu erarbeiteten Lehrinhalte aus dem Out-of-Class-Teil abfragt. Zum einen wird das kontinuierliche Lernen durch die Benotung des Präsenztests gefördert, wodurch Studierende einen höheren Anreiz zum Lernen haben. Zum anderen wird sichergestellt, dass sich die Studierenden im Vorfeld mit den Lehrinhalten vertraut gemacht haben. Dies wird vorausgesetzt, um die Anwendungsaufgaben in der Gruppenarbeit effizient zu lösen (Awidi & Paynter, 2019). Ein exemplarischer Fragebogen für den Präsenztest wurde ausgearbeitet und ist im Anhang A einsehbar.

Austausch

Nach dem Präsenztest besteht die Möglichkeit für eine Austauschphase, in der Fragen an den Lehrenden gerichtet werden. Hierunter fallen beispielsweise Fragen, die durch mehrere Studierende in den Foren gestellt wurden und die der Lehrende erneut in Präsenz aufgreift, um über bestehende Verständnisprobleme zu diskutieren. Des Weiteren besteht die Möglichkeit auf Fragen einzugehen, die im Zusammenhang mit dem Präsenztest aufgekommen sind.

Wiederholung und Aufgabenstellung

In der Phase der Wiederholung rekapituliert der Lehrende relevante Lehrinhalte zur Vorbereitung auf die Gruppenarbeit. Der Lehrende kann an dieser Stelle beispielsweise ein bereits aus den Lehrinhalten gelerntes Konzept aufgreifen, welches in der Gruppenarbeit angewendet wird, oder einen kurzen Exkurs vorstellen. Zweck der Wiederholung ist es nicht, den Studierenden die Lehrinhalte zu vermitteln, denn die direkte Vermittlung sollte bereits Out-of-Class geschehen sein.

Im Anschluss an die Wiederholung wird die Aufgabenstellung für die Gruppenarbeit vorgestellt. Für KoLA bieten sich als Aufgabenstellung Case Studies an, da diese praktische Problemlösekompetenzen fördern (Popil, 2011) und den Austausch zwischen den Studierenden anregen (Yadav et al., 2007). Die Case Studies werden ausgesucht oder selbst entwickelt, wobei ein Kooperationspartner aus der Industrie Daten für einen Case liefern könnte, um mehr Praxisnähe zu erreichen. Falls die Lehrinhalte softwarebasierte Konzepte umfassen, können diese mit der entsprechenden Software eingeführt und angewendet werden. Studierende können so an praxisnahe Abläufe herangeführt werden und diese eigenständig durchführen. Als Alternative zu Case Study und Softwareeinführung empfiehlt sich die Ausarbeitung von wissenschaftlichen Artikeln, um forschungsnahes Lernen zu fördern. Dabei gibt der Lehrende wissenschaftliche Artikel vor, die von den Gruppen separat analysiert und am Ende zusammen diskutiert werden.

Gruppenarbeit

Im Kern des In-Class-Teils wird die Gruppenarbeit durchgeführt, welche die Anwendung und Vertiefung der Lehrinhalte zum Ziel hat. Außerdem soll die Teamkompetenz sowie das Verständnis der Lehrinhalte durch den Austausch mit den Gruppenmitgliedern verbessert werden. Vorgeschaltet zur Gruppenarbeit erfolgt die Einteilung der Studierenden in Gruppen. Die Gruppeneinteilung gilt nur für den jeweiligen Termin, damit die In-Class-Aktivitäten unabhängig voneinander durchführbar ist. Einzelne Gruppen werden so nicht durch fehlende Mitglieder in Folgeterminen benachteiligt. Bezüglich der Gruppengröße eignen sich Gruppen mit drei bis fünf Studierenden für die gemeinsame Gruppenarbeit (AbuSeileek & Ali, 2012; Kriflik & Mullan, 2007). Kleinere Teams entwickeln schneller einen engeren Gruppenzusammenhalt und verbessern dadurch ihre anfängliche Teamleistung (Thompson et al., 2015). Damit geht der Ringelmann-Effekt einher, welcher besagt, dass Gruppenmitglieder weniger produktiv werden, wenn die Gruppengröße zunimmt (Ingham et al., 1974).

Um die Studierenden in Gruppen einzuteilen, gibt es unterschiedliche Methoden: nach Zufallsprinzip, Studierende suchen sich selbst in Gruppen zusammen oder der Lehrende teilt die Studierenden nach eigenen Kriterien ein. Das Zufallsprinzip beansprucht kaum Zeit und Arbeit, allerdings können solche Gruppen unausgewogen und ineffektiv zusammengestellt sein (Srba & Bielikova, 2014). Studierende präferieren stattdessen, ihre Gruppenmitglieder selbst auszuwählen (El Massah, 2018). Jedoch besteht dabei die Tendenz zur Bildung von homogenen Gruppen (Srba & Bielikova, 2014). Beispielsweise können Gruppen mit Studierenden gebildet werden, die einen ähnlichen Wissensstand besitzen oder sich bereits vor der Lehrveranstaltung kannten. Zum einen kann es dadurch vorkommen, dass ein Teil der Studierende ausgegrenzt wird. Zum anderen kann der erwünschte Wissens- und Ideenaustausch begrenzt werden (Srba & Bielikova, 2014). Eine geeignetere Methode die Gruppeneinteilung seitens des Lehrenden dar. Dieser hat die Möglichkeit, heterogene Gruppen anhand folgender Charakteristiken einzuteilen: Wissensstand, Kommunikations- und Führungskompetenzen (Moreno, Ovalle & Vicari, 2012). Da solch eine Einteilung komplex und zeitaufwendig für den Lehrenden sein kann, lohnt sich die Anwendung von computergestützten Algorithmen (Srba & Bielikova, 2014). An dieser Stelle wird auf einen beispielhaften Algorithmus von Moreno, Ovalle & Vicari (2012) verwiesen, da weitere Ausführungen den Rahmen dieser Arbeit überschreiten würden. Hierfür kann die Datenerhebung in die Bewerbungsphase mithilfe eines Fragebogens integriert werden. Der Algorithmus kann ad-hoc mit der Anwesenheitsliste am In-Class-Termin durchgeführt werden.

Sobald die Gruppen eingeteilt sind, können sich die Studierenden in verschiedenen Räumlichkeiten zusammensetzen. Zur Förderung einer kreativen Zusammenarbeit innerhalb der Gruppe können beispielsweise Flipcharts oder Whiteboards, Schreibstifte und Post-its zur Verfügung gestellt werden. Falls zur Lösung der Aufgabenstellung eine Closed Source Software, d.h. eine Software, die nicht kostenfrei erwerblich ist, zum Einsatz kommt, ist deren Bereitstellung durch das Institut vorgesehen. Die Ergebnisse und der Arbeitsfortschritt der Gruppe können beispielsweise in Microsoft PowerPoint oder in einem Miro-Board erfasst werden. Durch die gemeinsame Erarbeitung der Lösungen im gleichen Tool wird zudem die Kollaboration zwischen den Studierenden durch die erhöhte Anzahl an Abstimmungen gesteigert.

Während der Gruppenarbeit nehmen die Lehrenden eine zentrale Rolle ein. Sie dienen neben der Rolle als Lernbegleiter und Beobachter auch als Moderator. Bei Fragen, die während der Bearbeitung der Aufgabe entstehen, oder bei falscher Herangehensweise unterstützen Lehrende mit Ratschlägen und Hinweisen den Problemlösungs- bzw. Lernprozess der Gruppe. Dabei können Lehrende Methoden oder Tools empfehlen sowie auf bestimmte Lehrinhalte verweisen, aber sollen nicht die Lösung der Aufgaben bekannt geben. Die

Rolle des Beobachters sieht vor, einzelne Gruppenmitglieder und ihren Beitrag zur Gruppenarbeit zu beobachten und zu dokumentieren, denn die Mitarbeit und die Beiträge der einzelnen Studierenden fließen in deren individuelle Note mit ein (vgl. Kapitel „Prüfungsleistung bei KoLA“). Sowohl die Benotung der individuellen Mitarbeit als auch die Anwesenheit des Lehrenden kann zur Reduzierung von Trittbrettfahrern (Schenk & Schwabe, 2001) beitragen. Als Trittbrettfahrer werden dabei Gruppenmitglieder bezeichnet, die sich an der Gruppenarbeit sowie dem Problemlösungsprozess wenig bis gar nicht beteiligen, jedoch von der Arbeit der weiteren Gruppenmitglieder profitieren. Des Weiteren agiert der Lehrende im Falle von Unstimmigkeiten oder persönlichen Spannungen innerhalb der Gruppe als Moderator, sodass durch seine Koordination die Arbeitsfähigkeit der Gruppe aufrechterhalten wird (Schenk & Schwabe, 2001). Außerdem dient die Moderation zur Aufstellung und Beachtung von gemeinsamen Gruppenregeln sowie Rollen innerhalb der Gruppe. Einfache Regeln, wie gegenseitige Akzeptanz der Meinung, tragen zu einer angenehmeren Gruppenharmonie bei. Durch die Definition konkreter Rollen innerhalb der Gruppen werden mögliche Rollenkonflikte in der Gruppenarbeit vermindert (Jöns, 2016). Als Rolle können beispielsweise Zeitwächter, Gruppensprecher und Protokollant definiert werden. Der Lehrende sorgt für die Einhaltung der Rollen und erinnert die Studierende bei Bedarf an ihre Rolle. Trotz Unterstützung und Moderation der Gruppe durch den Lehrenden kann die Gruppe unter gewissen Umständen den gewünschten Output nicht erbringen. Denn „[d]er Erfolg der Gruppenarbeit hängt entscheidend von der Einstellung, der Motivation und der aktiven Mitarbeit der Gruppenmitglieder ab“ (Jöns, 2016). Die Motivation der einzelnen Gruppenmitglieder ist wiederum vom individuellen Interesse am Lehrinhalt sowie persönlichen Zielen, wie beispielsweise das Erlangen von einer guten Note oder von neuem Wissen, abhängig (Niegemann et al., 2008). Aus diesem Grund nehmen die Bewerbungsphase und das Motivationsschreiben (vgl. Kapitel „Bewerbungsphase“) zur Auswahl der interessierten Studierenden eine zentrale Bedeutung ein, um den Erfolg der Lehrveranstaltung zu gewährleisten.

Nach der Aufgabenbearbeitung werden die Ergebnisse im Plenum vorgestellt[1]. Alternativ können die Ergebnisse der einzelnen Gruppen abgegeben und erst im darauffolgenden In-Class-Termin besprochen werden. Der Vorteil besteht darin, dass die Lehrenden die Aufgaben ganzheitlich bewerten und für jeden Aufgabenteil eine Gruppe zur Vorstellung auswählen können. Durch die lange Zeitspanne zwischen den In-Class-Teilen besteht jedoch die Gefahr, dass die Aufgaben bei den Studierenden nicht mehr präsent sind und das Interesse für die zuvor behandelte Thematik niedriger ist.

Die Aufgaben und die darauffolgende Feedbackrunde zu den Ergebnissen werden vom Lehrenden anmoderiert. Aufgrund der zeitlichen Beschränkung werden nicht die Ergebnisse von allen Gruppen präsentiert. Jeder Gruppe wird vom Lehrenden ein Aufgabenteil zugeordnet, welchen sie vorstellen. Durch die Vorstellung der Ergebnisse wird zum einen die Präsentationskompetenz der Studierenden gefördert. Zum anderen erhalten die anderen Gruppen die Möglichkeit, ihre Ergebnisse zu vergleichen und ggfs. alternative Lösungsmöglichkeiten kennenzulernen.

Abschluss

Nach der Vorstellung der Ergebnisse im Plenum sieht KoLA ein ausführliches Feedback für die erarbeiteten Aufgaben vor. Durch das Feedback werden Verbesserungspotentiale der einzelnen Gruppen bzw. -mitglieder aufgedeckt. Hierdurch erhalten die Studierende die Möglichkeit, ihre Leistung sowie ihre Fähigkeiten kontinuierlich zu verbessern. Das Feedback zu ihren vorgestellten Ergebnissen erhalten die Studierenden über zwei Phasen. Zunächst findet ein Peer-to-Peer-Feedback in Form eines Gruppenfeedbacks statt. Beispielweise gibt Gruppe A Feedback an Gruppe B, wobei Gruppe B ihr Feedback von Gruppe C erhält. Unter Peer-to-Peer-Feedback wird dabei die Evaluation der Studierenden durch ihre Mitstudierenden verstanden (Hanstein & Lanig, 2020). Das Gruppenfeedback fördert neben der Interaktion zwischen den Studierenden die Beteiligungsquote an den Diskussionen, da die Gruppen-mitglieder durch die dedizierte Zuteilung einer anderen Gruppe zur Teilnahme an der Evaluation

[1] Alternativ können die Ergebnisse der einzelnen Gruppen abgegeben und erst im darauffolgenden In-Class-Termin besprochen werden. Der Vorteil besteht darin, dass die Lehrenden die Aufgaben ganzheitlich bewerten und für jeden Aufgabenteil eine Gruppe zur Vorstellung auswählen können. Durch die lange Zeitspanne zwischen den In-Class-Teilen besteht jedoch die Gefahr, dass die Aufgaben bei den Studierenden nicht mehr präsent sind und das Interesse für die zuvor behandelte Thematik niedriger ist.

aufgefordert werden. Durch die Bereitstellung von Feedbackregeln unterstützen Lehrende Studierende, wie sie gezieltes und konstruktives Feedback äußern können. Zum Beispiel soll mit positivem Feedback, das das Selbstwertgefühl des Individuums stärkt, begonnen werden und im Anschluss auf die Verbesserungspotentiale eingegangen werden (Ki Chan & Pawlina, 2020). Im Anschluss an das Gruppenfeedback werden ergänzende Rückmeldungen und Kommentare von dem Lehrenden an die Gruppe bzw. Gruppenmitglieder gegeben. Für ein besseres Zeitmanagement sowie für eine faire Nachbesprechung der gruppenindividuellen Ergebnisse wird eine zeitliche Restriktion pro Gruppe empfohlen.

Am Ende jedes In-Class-Teils findet zusätzlich eine Feedbackrunde zum Lehrkonzept statt. Durch das Feedback der Studierenden wird eine fortdauernde Verbesserung und Weiterentwicklung des Lehrkonzepts gewährleistet. Die Lehrenden können die Feedbackrunde über eine Reflektion des Lehrtages in Form einer offenen Diskussion einleiten. Jedoch kann die Hemmschwelle für ein ehrliches Feedback in offenen Diskussionsrunden hoch sein, da sich Studierende zurückhaltend verhalten können. Um ein ehrliches Feedback zu erhalten, eignen sich anonyme Online-Umfragen, die der Lehrende bereitstellt. Für die Erstellung von anonymen Online-Umfragen können Audience Response Tools wie Mentimeter oder Kahoot! in Betracht gezogen werden. Über Bewertungsskalen können Studierende beispielsweise den Schwierigkeitsgrad der Aufgaben, den empfundenen Arbeitsaufwand sowie ihren Spaßfaktor in der Gruppenarbeit bewerten. Weitere Verbesserungsvorschläge können über offene Fragestellungen, wie beispielsweise „Mir hat weniger gut gefallen“ oder „Mir hat gefehlt“, eingeholt werden. Der Einsatz von einer Likert-Skala zum individuellen Lernfortschritt kann die Online-Umfrage abrunden.

Mündliche Prüfung

Mit der Prüfung am Ende der Lehrveranstaltung soll das Gesamtverständnis überprüft werden. Die Prüfung kann mündlich oder schriftlich durchgeführt werden. Bei einer mündlichen Prüfung ist der zeitliche Aufwand zu berücksichtigen. Ab einer gewissen Anzahl an Teilnehmenden unterschreitet der Korrekturaufwand einer schriftlichen Prüfung dem zeitlichen Aufwand einer mündlichen Prüfung. Deshalb ist bei einer hohen Teilnehmeranzahl die schriftliche Prüfungsform zu bevorzugen. Da bei KoLA die Teilnehmeranzahl auf 30 beschränkt ist, ist der Aufwand für die mündlichen Prüfungen vertretbar. Des Weiteren spricht für die mündlichen Prüfungsform, dass eine breite und tiefe Wissensabfrage individuell möglich ist. Jedoch ist zu bedenken, dass der Anspruch mündlicher Prüfungen bei allen zu Prüfenden möglichst gleich sein sollte, damit die Gerechtigkeit und die Vergleichbarkeit gewahrt wird. Zur Erreichung dieses Ziels sollte vorab ein einheitlicher Bewertungsmaßstab festgelegt werden.

Zur Durchführung der mündlichen Prüfung kommen zwei mögliche Varianten in Frage: Einzel- oder Gruppenprüfung. Eine mündliche Einzelprüfung würde in der Regel zwischen 30 und 45 Minuten dauern (Gerick, Sommer & Zimmermann, 2017). Die mündliche Gruppenprüfung dauert je nach Gruppengröße 60 Minuten und mehr (Gerick, Sommer & Zimmermann, 2017). Letztere Variante eignet sich für Lehrveranstaltungen, bei denen interdisziplinär und problembasiert gelernt werden muss, da das Prüfungssetting anwendungsbezogen ist (Gerick, Sommer & Zimmermann, 2017). Da KoLA neben dem kontinuierlichen Lernen die Praxisanwendung zum Ziel hat, wird eine Gruppenprüfung empfohlen. Ein exemplarischer Ablauf könnte wie folgt aufgebaut sein: Die Gruppe erhält ein komplexes, interdisziplinäres Problem und hat 15 Minuten Zeit zum Durchlesen der Aufgabenstellung (Gerick, Sommer & Zimmermann, 2017). Bei Bedarf dürfen Notizen gemacht werden. Anschließend findet die mündliche Prüfung als Gruppendiskussion bzw. Brainstorming für mögliche Lösungsansätze statt.

Prüfungsleistung bei KoLA

Im Vergleich zur lehrendenzentrierten Lehre, bei der die Gesamtnote in einer einzelnen Abschlussprüfung ermittelt wird, soll KoLA sicherstellen, dass sich die Prüfungsleistung nicht auf eine einzelne Leistung beschränkt. Bei KoLA setzt sich die Prüfungsleistung deshalb aus drei Noten zusammen: die Präsenztests, die Gruppenarbeiten und die mündliche Prüfung.

Zur Bewertung der Präsenztests werden Punkte vergeben. Die maximale Anzahl an Punkten pro Teilaufgabe ist zur Wahrung der Vergleichbarkeit einheitlich festgelegt. Die Punkteverteilung für die jeweiligen Teilaufgaben ergibt sich deshalb aus einer Musterlösung. Diese stellt die Grundlage für die Korrektur dar. Die Gruppenarbeit wird von den Lehrenden nach festgelegten Kriterien bewertet. So soll

gewährleistet werden, dass die Bewertung objektiv erfolgt. Ansätze für mögliche Bewertungskriterien sind Interaktion zwischen den Gruppenmitgliedern, Ergebnisse der Gruppenarbeit und die Teamfähigkeit. Die mündliche Prüfung sollte ebenfalls nach einem festgelegten Bewertungsmaßstab erfolgen. Detaillierte Hinweise zur Umsetzung finden sich im Kapitel „Mündliche Prüfung".

Die Prüfungsleistung wird durch die Bewertung der einzelnen Noten mit Gewichtungen berechnet. Für KoLA empfiehlt sich folgende Verteilung: 25% Präsenztest, 25% Gruppenarbeit, 50% mündliche Prüfung (vgl. Abschnitt „Entwicklung des Lehrkonzeptes KoLA").

Grenzen und Alternativen

Voraussetzungen und Anwendbarkeit von KoLA

Für die reibungslose Durchführung der Out-of-Class-Bestandteile bestehen einige technische Voraussetzungen: Teilnehmende müssen über eine Internetverbindung und ein mobiles Endgerät verfügen, um die Lehrinhalte orts- und zeitunabhängig abzurufen. Zudem muss durch die Lehrenden eine Lernplattform mit Cloudspeicher aufgesetzt, konfiguriert und betreut werden (Akçayır & Akçayır, 2018).

Organisatorisch ist ein höherer Zeitbedarf für die Betreuung wie auch für die Teilnahme anzusetzen (Akçayır & Akçayır, 2018). Beispielsweise sollten die Lehrenden aufgrund der technischen Herausforderungen für ein gutes Onboarding der Studierenden sorgen (Werner & Ebel, 2018). Außerdem ist die Bereitstellung und das Korrigieren der Tests zeitintensiv. Die Teilnehmenden wiederum bedürfen pro Präsenztermin mehr Zeit aufgrund der Bearbeitung des Präsenztests und der Aufgaben während der Gruppenarbeit.

Bei einer großen Teilnehmeranzahl (ab 40 bis 50 Studierende) KoLA schwer anzuwenden. Hier besteht die Problematik, dass der Organisationsaufwand für die Präsenzveranstaltungen sehr hoch ist (Nederveld & Berge, 2015). In der Folge muss mehr Zeit eingeplant werden, weil die Ergebnisbesprechung und die Fragerunden voraussichtlich länger dauern. Weil KoLA für eine Gesamt-Teilnehmeranzahl von 20 bis 30 Studierenden konzipiert ist, wäre ein naheliegender Lösungsansatz für dieses Problem, zwei parallele Veranstaltungen für jeweils etwa die Hälfte der Teilnehmenden anzubieten.

Herausforderungen und Grenzen von KoLA

Der Arbeitsaufwand beim Flipped Classroom ist sowohl für Studierende als auch Lehrende höher als bei traditionellen Lehrveranstaltungen (Akçayır & Akçayır, 2018). Auf der einen Seite wird eine aktive Mitarbeit durch die Studierenden vorausgesetzt: Wenn Studierende die Lehrinhalte nicht Out-of-Class gelernt haben, kann es sein, dass sie bei den In-Class-Aktivitäten schlechter abschneiden (Sayeski, Hamilton-Jones & Oh, 2015). Dadurch werden die Vorteile des Flipped Classroom vermindert. Auf der anderen Seite müssen die Lehrenden die Lehrinhalte neu für den Out-of-Class-Teil aufbereiten und ihr Konzept von 90-minütigen Vorlesungen auf kurze Videos umstrukturieren. Die Vorbereitung ist dadurch insbesondere für die erste Durchführung sehr arbeitsintensiv (Akçayır & Akçayır, 2018). Weiterhin ist der Korrekturaufwand der Tests und der Betreuungsaufwand aufgrund der Diskussionsforen und Sprechstunden hoch.

Im Vergleich zu anderen Lehrkonzepten kann KoLA zu Ablehnung aufgrund der Neuartigkeit bei Lehrenden und Studierenden führen. Der erhöhte Zeit- und Arbeitsaufwand kann Ängste, Widerstände und Umsetzungsprobleme erzeugen (Akçayır & Akçayır, 2018).

Studierende sind aufgrund der In-Class-Termine unter dem Semester zeitlich gebunden und ein Fehlen bei einem der Termine hat Auswirkungen auf die Note. Etwas abgefangen wird dieser Aspekt durch die Tatsache, dass die Teilnahme an nur drei von vier Präsenzterminen verpflichtend ist. Hierdurch sollen die Folgen des einmaligen Fehlens durch Krankheit, höhere Gewalt oder andere Gründe kompensiert werden.

Die Konsequenzen, die aus einem Teilnahmeabbruch die Studierenden tragen, sollten vom Lehrenden im Vorfeld kommuniziert werden. Je nach Fortschritt der Veranstaltung wird ein Rücktritt zugelassen oder das Nichtbestehen der Gesamtprüfungsleistung und die Note 5,0 resultieren. Zudem sollte geregelt werden, bis zu welchem Anteil an Fehzeiten bedingt durch Krankheit oder höhere Gewalt ein Abschluss der Veranstaltung (eventuell durch Ersatzleistungen, -termine, -aufgaben) möglich ist und ab wann eine

Abmeldung durch die Lehrende die einzige Option bleibt. Bei all dem sollte zusätzlich berücksichtigt werden, dass der Erfolg von KoLA maßgeblich davon abhängt, dass die einzelnen Gruppen nicht aufgrund von Abbrüchen in sich zusammenfallen und Vorkehrungen getroffen werden, welche dieses Szenario abfedern.

Die maximale Anzahl an Teilnehmenden ist stärker begrenzt als bei konventionellen Lehrveranstaltungen, die durch KoLA ersetzt werden sollen und gleicht eher derer von Seminaren. Insbesondere umfasst die Idee und Motivation von KoLA auch den Ersatz großer Vorlesungen, sodass bedacht werden muss, dass der Durchführungsaufwand und die Teilnahmebeschränkungen weitaus höher sein können als bei den traditionellen Veranstaltungen, die dadurch ersetzt werden sollen. Möglicherweise sind sogar mehrere parallele Lehrveranstaltungen notwendig, um dem gerecht zu werden.

Die Auswahl qualitativer Case Studies für die Gruppenarbeiten kann je nach gewünschtem Themengebiet ebenfalls eine Herausforderung darstellen. Möglicherweise ist es sogar notwendig, dass die Case Studies extra dafür formuliert werden müssen.

Fazit

Im Rahmen dieser Arbeit wurde die Frage beantwortet, wie kontinuierliches Lernen und Anwenden mithilfe eines Lehrkonzeptes bei Studierenden gefördert werden kann. Als Grundlage diente hierbei das Konzept des Flipped Classroom. Durch Kombination von wissenschaftlichen Forschungsergebnissen zum Flipped Classroom und Erfahrungen von Studierenden wurde das Lehrkonzept KoLA ausgearbeitet. KoLA sieht einen Wechsel von der dozierendenzentrierten zur studierendenzentrierten Lehre vor. Dafür wurde gemäß dem Flipped Classroom die Informationsvermittlung Out-of-Class verlagert, sodass die Zeit In-Class zur Vertiefung und Anwendung der Lehrinhalte genutzt werden kann. Dementsprechend eignet sich KoLA besonders für Lehrveranstaltungen mit Seminar- bzw. Praktikumscharakter.

Aufgrund dessen ist KoLA nicht allgemein anwendbar, sondern nur für Lehrveranstaltungen unter den angegebenen Rahmenbedingungen, wie z.B. einer kleinen Teilnehmeranzahl von 20-30 Studierenden. Eine weitere Limitation dieser Arbeit liegt in der begrenzten Gestaltungsmöglichkeiten der Online-Tests. Diese können je nach Fachgebiet und Lehrinhalt angepasst werden, um die Lernstandsermittlung aussagekräftiger zu machen. Auf die Gestaltungsmöglichkeiten der Online-Tests wurde in dieser Arbeit nicht weiter eingegangen. Des Weiteren wurde nicht auf die Aufbereitung der Lehrinhalte im Out-of-Class-Teil eingegangen. Im Vergleich zur traditionellen Lehre ist hierbei die starke Komprimierung der Erklärzeit in den Videos herausfordert, da die Qualität der Ausführungen gleichbleiben soll. Der Fokus dieser Arbeit lag auf der Ausarbeitung der In-Class-Elemente, da die Hauptvorteile des Flipped Classroom durch die interaktiven Lernaktivitäten vor Ort zustande kommen (Jensen, Kummer & Godoy, 2015). Bei der Gruppeneinteilung wurde nur ein beispielhafter Algorithmus vorgeschlagen. Die praktische Anwendung während eines In-Class-Termins wurde nicht weiter geprüft. Weiterhin fehlt eine Fairnessprüfung geeigneter Bewertungskriterien der Gruppenarbeit. Unter anderem wurde die Interaktion zwischen den Gruppenmitgliedern aufgeführt, wobei offenbleibt, ob dies nachteilig für introvertierte Persönlichkeiten ausfällt.

Da die Informationsvermittlung gemäß dem Flipped Classroom Out-of-Class verlagert wurde, wird ein zeit- und ortsunabhängiges Lernen mit KoLA ermöglicht. Dadurch wird das Teilziel des individuellen Lerntempos gefördert, denn die Studierenden können die Lehrinhalte so oft und beliebig schnell bearbeiten, wie es ihrem Lerntyp entspricht. Eine vollständige örtliche und zeitliche Unabhängigkeit wurde im Rahmen dieser Arbeit nicht umgesetzt, da In-Class-Termine zur Vertiefung und Anwendung benötigt werden. Verglichen mit einem klassischen Semester mit durchschnittlich zwölf Vorlesungsterminen konnte die zeitliche Abhängigkeit auf drei bis vier In-Class-Termine deutlich reduziert werden. Weiterhin wurde der Fokus in den wenigen Präsenztermine auf die Studierenden gelegt, indem beispielsweise Gruppenarbeiten durchgeführt werden. Im Out-of-Class-Teil wird der Aspekt der Interaktion mithilfe von interaktiven Tools und Sprechstunden aufgegriffen. Somit wird ganzheitlich das Teilziel der Interaktionsförderung erfüllt. Die beiden Teilziele unterstützen das übergeordnete Ziel der Arbeit, dem kontinuierlichen Lernen und Anwenden. Das übergeordnete Ziel wird ebenfalls durch den Bestandteil der benoteten Präsenztests adressiert. Die kapitelweise durchgeführten Leistungsprüfungen mit Präsenztests unter dem Semester unterstützen das kontinuierliche Auseinandersetzen mit den Lehrinhalten. Ein Bulimielernen kurz vor der

Prüfung wird dadurch entgegengewirkt, da sich die Leistungsprüfung aus mehreren Noten zusammengesetzt. Die Ziele, die für diese Arbeit angesetzt wurden, konnten somit erfüllt werden.

Lehrenden liegt mit dem Lehrkonzept KoLA ein Lösungsvorschlag vor, um das kontinuierliche Lernen und Anwenden bei Studierenden zu fördern. Sofern Lehrveranstaltungen den Rahmenbedingungen des Lehrkonzepts entsprechen, kann es in der universitären Lehre angewendet werden. Durch Feedback aus dem Einsatz in der universitären Lehre kann das Lehrkonzept iterativ verbessert werden.

Forschende können aus dem Lehrkonzept KoLA entnehmen, wie die Bestandteile von Flipped Classroom anhand einer Problemstellung ganzheitlich aufeinander abgestimmt werden können. Damit das Lehrkonzept auch auf Lehrveranstaltungen außerhalb der definierten Rahmenbedingungen angewendet werden kann, sollte auf Grundlage dieser Arbeit ein allgemeingültiges Lehrkonzeptes ausgearbeitet werden. Einzelne Bestandteile, wie z.B. die Gestaltung der Online-Tests, könnten dabei noch verfeinert werden. Weiterer Forschungsbedarf besteht beispielsweise bei der Aufbereitung der Lehrinhalte. Hier sollte auf die Frage eingegangen werden, wie die Lehrinhalte Out-of-Class aufbereitet werden sollen, um eine Vorlesung zu ersetzen oder die Lehrinhalte im Rahmen eines kurzen Videos wiederzugeben. In Bezug auf Anreizsysteme sollte auch betrachtet werden, wie die Online-Tests sowie die Videos durch die Einbindung von Gamification-Elementen profitieren könnten. Einen weiteren Punkt stellen die benoteten Präsenztests dar, die zum kontinuierlichen Lernen anregen sollen. Andere Anreizsysteme für kontinuierliches Lernen außer den Notendruck durch Tests könnten noch erforscht werden.

Literaturverzeichnis

AbuSeileek, A. F. (2012). The effect of computer-assisted cooperative learning methods and group size on the EFL learners' achievement in communication skills. Computers & Education, 58(1), 231–239. https://doi.org/10.1016/j.compedu.2011.07.011

Akçayır, G. & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. Computers & Education, 126, 334–345. https://doi.org/10.1016/j.compedu.2018.07.021

Al-Zahrani, A. M. (2015). From passive to active: The impact of the flipped classroom through social learning platforms on higher education students' creative thinking. British Journal of Educational Technology, 46(6), 1133–1148. https://doi.org/10.1111/bjet.12353

Awidi, I. T. & Paynter, M. (2019). The impact of a flipped classroom approach on student learning experience. Computers & Education, 128, 269–283. https://doi.org/10.1016/j.compedu.2018.09.013

Bergmann, J. & Sams, A. (2012). Flip your classroom: Reach every student in every class every day (1. ed.). ASCD; International Society for Technology in Education.

Chan, L. K. & Pawlina, W. (Hrsg.). (2020). Springer eBook Collection. Teaching Anatomy: A Practical Guide (2. Aufl.). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-43283-6

Dehling, H., Roegner, K. & Winzker, M. (2014). Transfer von Studienreformprojekten für die Mathematik in der Ingenieurausbildung. BoD – Books on Demand.

Dorgerloh, S. & Wolf, K. D. (Hrsg.). (2020). Lehren und Lernen mit Tutorials und Erklärvideos. Beltz.

El Massah, S. S. (2018). Addressing free riders in collaborative group work. International Journal of Educational Management, 32(7), 1223–1244. https://doi.org/10.1108/IJEM-01-2017-0012

Flipped Learning Network (FLN) (2014), What is flipped learning?, Zugriff am 13.03.2022. Verfügbar unter http://flippedlearning.org/cms/lib07/VA01923112/Centricity/Domain/46/FLIP_handout_FNL_Web.pdf

Gerick, J., Sommer, A. & Zimmermann, G. (Hrsg.). (2018). utb Hochschuldidaktik: Bd. 4840. Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre. Waxmann.

Goldin, A. P., Sigman, M., Braier, G., Golombek, D. A. & Leone, M. J. (2020). Interplay of chronotype and school timing predicts school performance. Nature human behaviour, 4(4), 387–396. https://doi.org/10.1038/s41562-020-0820-2

Hamdan, N. & McKnight, P. (2013). Review of Flipped Learning. Vorab-Onlinepublikation. https://doi.org/10.4236/ce

Hanstein, T. & Lanig, A. K. (2020). Digital lehren: Das Homeschooling-Methodenbuch (1. Auflage). Tectum – ein Verlag in der Nomos Verlagsgesellschaft. http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1865747

Ingham, A. G., Levinger, G., Graves, J. & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. Journal of Experimental Social Psychology, 10(4), 371–384. https://doi.org/10.1016/0022-1031(74)90033-X

Jensen, J. L., Kummer, T. A. & d M Godoy, P. D. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. CBE life sciences education, 14(1), ar5. https://doi.org/10.1187/cbe.14-08-0129

Jöns, I. (Hrsg.). (2016). Springer eBook Collection. Erfolgreiche Gruppenarbeit: Konzepte, Instrumente, Erfahrungen (2. Aufl.). Gabler Verlag. https://doi.org/10.1007/978-3-8349-4762-8

Kern, B. B. (2002). Enhancing accounting students' problem-solving skills: the use of a hands-on conceptual model in an active learning environment. Accounting Education, 11(3), 235–256. https://doi.org/10.1080/09639280210141680

Kriflik, L. & Mullan, J. (2007). Strategies to Improve Student Reaction to Group Work. Journal of University Teaching and Learning Practice, 4(1), 17–32. https://doi.org/10.53761/1.4.1.3

LaTour, K. A. & Noel, H. N. (2021). Self-Directed Learning Online: An Opportunity to Binge. Journal of Marketing Education, 43(2), 174–188. https://doi.org/10.1177/0273475320987295

McDonald, K. & Smith, C. M. (2013). The flipped classroom for professional development: part I. Benefits and strategies. Journal of continuing education in nursing, 44(10), 437–438. https://doi.org/10.3928/00220124-20130925-19

Moreno, J., Ovalle, D. A. & Vicari, R. M. (2012). A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. Computers & Education, 58(1), 560–569. https://doi.org/10.1016/j.compedu.2011.09.011

Nederveld, A. & Berge, Z. L. (2015). Flipped learning in the workplace. Journal of Workplace Learning, 27(2), 162–172. https://doi.org/10.1108/JWL-06-2014-0044

Niegemann, H. M., Domagk, S., Hessel, S., Hein, A., Hupfer, M. & Zobel, A. (2008). Kompendium multimediales Lernen. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-37226-4

Popil, I. (2011). Promotion of critical thinking by using case studies as teaching method. Nurse education today, 31(2), 204–207. https://doi.org/10.1016/j.nedt.2010.06.002

Sayeski, K. L., Hamilton-Jones, B. & Oh, S. (2015). The Efficacy of IRIS STAR Legacy Modules Under Different Instructional Conditions. Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children, 38(4), 291–305. https://doi.org/10.1177/0888406415600770

Schenk, B. & Schwabe, G. (2001). Moderation. https://doi.org/10.5167/UZH-67756

Srba, I. & Bielikova, M. (2014). Dynamic Group Formation as an Approach to Collaborative Learning Support. IEEE Transactions on Learning Technologies, 8(2), 173–186. https://doi.org/10.1109/TLT.2014.2373374

Strelan, P., Osborn, A. & Palmer, E. (2019). Student satisfaction with courses and instructors in a flipped classroom: A meta-analysis. Journal of Computer Assisted Learning, 36(3), 295–314. https://doi.org/10.1111/jcal.12421

Strelan, P., Osborn, A. & Palmer, E. (2020). The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels. Educational Research Review, 30, 100314. https://doi.org/10.1016/j.edurev.2020.100314

Thompson, B. M., Haidet, P., Borges, N. J., Carchedi, L. R., Roman, B. J. B., Townsend, M. H., Butler, A. P., Swanson, D. B., Anderson, M. P. & Levine, R. E. (2015). Team cohesiveness, team size and team performance in team-based learning teams. Medical education, 49(4), 379–385. https://doi.org/10.1111/medu.12636

Topping, K. & Ehly, S. (1998). Peer-assisted Learning. Routledge. https://doi.org/10.4324/9781410603678

Ulrich, I. (2016). Gute Lehre in der Hochschule: Praxistipps zur Planung und Gestaltung von Lehrveranstaltungen. Springer. http://www.springer.com/

Werner, J., Ebel, C., Spannagel, C. & Bayer, S. (Hrsg.). (2018). Flipped Classroom - Zeit für deinen Unterricht: Praxisbeispiele, Erfahrungen und Handlungsempfehlungen. Verlag Bertelsmann Stiftung. http://flipyourclass.christian-spannagel.de/wp-content/uploads/2018/10/9783867938693_Flipped_PDF-Onlineversion.pdf

Winteler, A. (2011). Professionell lehren und lernen. Ein Praxisbuch. https://www.kaththeol.uni-muenchen.de/lehre/multiplikatoren-projekt/virtuelle_bib/downloads_virt-bib/winteler_profess-lehren_2008.pdf

Yadav, A., Lundeberg, M. A., DeSchryver, M., Dirkin, K. H. & Herreid, C. F. (2007). Teaching science with case studies: A national survey of faculty perceptions of the benefits and challenges of using cases. Journal of College Science Teaching, 37(1), 34–38. Zugriff am 18.03.2022. Verfügbar unter: https://www.researchgate.net/profile/aman-yadav-12/publication/262960630_teaching_science_with_case_studies_a_national_survey_of_faculty_perceptions_of_the_benefits_and_challenges_of_using_cases

ZML, Zentrum für Mediales Lernen. (2020). Verwendung von Microsoft Teams in der Lehre. Zugriff am 18.03.2022. Verfügbar unter: https://www.zml.kit.edu/downloads/Anleitung_Lehre_MSTeams.pdf

Anhang

Beispielhafter Präsenztest

In Anlehnung an die Vorlesung „Innovationsmanagement“ von Universitätsprofessorin Dr. Marion A. Weissenberger-Eibl, Inhaberin des Lehrstuhls für Innovations- und TechnologieManagement (iTM) am Institut für Entrepreneurship, Technologie-Management und Innovation (EnTechnon).

Karlsruher Institut für Technologie

Seite 1 von 2

Test 1 – Grundlagen Innovationsmanagement

Name	Matrikelnummer
Studiengang	**Erreichte Punktzahl (von max. x Punkten)**

Aufgabe 1: Richtig oder falsch? Korrigiere die falschen Aussagen. **(x Punkte)**

Aussagen	Richtig	Falsch	Korrektur
a. Innovation ist die erstmalige technische Realisierung einer neuen Problemlösung.	☐	☐	
b. Inkrementelle Innovationen schaffen tiefgreifende Veränderungen.	☐	☐	
c. Als Open Innovation wird der Austausch eines Unternehmens mit externen Akteuren in allen Phasen bezeichnet.	☐	☐	

Aufgabe 2: Nenne... **(x Punkte)**

a. die Merkmale einer Innovation, damit sie von Individuen angenommen werden:
- _____
- _____
- _____
- _____
- _____

b. die Kompetenztypen:
- _____
- _____
- _____
- _____

c. die Phasen der Szenariotechnik:
1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____

Aufgabe 3: Zeichne… **(x Punkte)**

- das idealtypische Verhalten von Akteuren im Verlauf des Diffusionsprozesses. Achte auf die Beschriftung.

Aufgabe 4: Multiple Choice **(x Punkte)**

a. Die Dimensionen des Kundenwissens sind:
- ☐ Wissen von Kunden
- ☐ Wissen über Kunden
- ☐ Wissen durch Kunden
- ☐ Wissen der Kunden
- ☐ Wissen für Kunden

b. Open Innovation zeichnet sich aus durch:
- ☐ Einen kollaborierenden Innovationsprozess
- ☐ Transparente Öffentlichkeitsarbeit
- ☐ Vertrauensvolle Zusammenarbeit
- ☐ Vernetzung von Wissen
- ☐ Outsourcing von Entwicklungsprozessen

c. Die Kategorien der Innovationen sind:
- ☐ Prozessinnovation
- ☐ Produktionsinnovation
- ☐ Marketinginnovation
- ☐ Vertriebsinnovation
- ☐ Organisationale Innovation
- ☐ Personelle Innovation

Aufgabe 5: Offene Frage **(x Punkte)**

- Beschreibe das Umfeld, das die Ausgestaltung des Innovationsmanagement mitbestimmt.

Flipped Classroom 2.0: A New Teaching Concept

Selected Issues in Critical Information Infrastructures, Winter Term 21/22

Christian Wiegand
Master Student
Karlsruhe Institute of Technology
christian.wiegand@student.kit.edu

Tim Konnowski
Master Student
Karlsruhe Institute of Technology
tim.konnowski@student.kit.edu

Dorsaf Ameur
Master Student
Karlsruhe Institute of Technology
dorsaf.ameur@student.kit.edu

Lukas Brecht
Master Student
Karlsruhe Institute of Technology
lukas.brecht@student.kit.edu

Fatih Celik
Master Student
Karlsruhe Institute of Technology
fatih.celik@student.kit.edu

Abstract

Background: *Students lack the motivation to learn continuously during the semester, which has been identified as a widespread problem. Instead, they opt for condensed learning during the exam period. These students contribute less to the teaching process and settle for a passive role within the academic system. As a result, the content learned is forgotten shortly after the exams, and the teaching objectives are not fully achieved.*

Objective: *In this paper, we aim to develop a teaching concept for a university course that increases the involvement of students in the teaching process. For this, the given courses have to be more interactive and practice-oriented. Besides, the concept should allow self-evaluation during the learning phase and not only during the final examination. Due to the Covid19 pandemic, it should also be possible to implement the designed concept online or face-to-face at the university.*

Methods: *The teaching model in this paper is utility-oriented and sets the students´ needs and requirements on top of the concept development process. An analysis of students' experiences combined with a literature search on existing teaching and learning theories has led to a rough concept design. This first sample has been validated through focus group discussions among students and by getting feedback from university lecturers. In addition, we conducted an experimental interactive tutorial session on a group of students. Accordingly, we evaluated the concept's pros and cons to bring the necessary adjustments and realize the final concept version, namely "Flipped Classroom 2.0".*

Results: *The resulting concept, namely "Flipped Classroom 2.0", combines online lecture videos with regular interactive "Bootcamp" meetings to consolidate what has been learned. Learning the theoretical basics is accompanied by a compulsory semester*

project, in which the students transfer their theoretical knowledge into practice. The lecturers provide the uploaded lecture videos, whereas students prepare in advance summaries and quizzes related to the respective chapters of the learning content. Students work in groups on the semester project to solve issues related to subject-specific use cases.

***Conclusion:** "Flipped Classroom 2.0" meets the set objectives regarding the involvement of students in the teaching process. Participation in the "Bootcamps" allows a paradigm shift from a passive to an active student role. Continuous learning during the semester is ensured, and flexible implementation of the concept is possible.*

Keywords: lecture, teaching concept, flipped classroom, remote learning, online teaching, students, university, continuous learning

Einleitung

Die traditionelle Lehrveranstaltung an einer Universität besteht typischerweise aus Vorlesung, Übung sowie Tutorien und einer abschließenden Klausur. Hierbei werden die Inhalte in der Vorlesung vom Dozenten vorgetragen, in der Übung und Tutorium vertieft und durch die Klausur anschließend abgefragt. Dennoch kann dieses Konzept zu Problemen führen. Schließlich belegt eine erfolgreich absolvierte Klausur nicht gleichzeitig einen langfristigen Lernerfolg. Beispielsweise gaben innerhalb einer Umfrage 36,4% der Studierenden von deutschen Universitäten an, erhebliche Probleme bei der Aneignung der Lerninhalte zu haben (Willige et al., 2018). Häufig fehlt es den Studierenden an der Motivation, sich während des Semesters kontinuierlich Lerninhalte anzueignen. Die mittlere Arbeitsbelastung von Bachelorstudierenden während der Prüfungsphase kann bis zu 50% über dem Durschnitt des Semesters liegen (Metzger & Schulmeister, 2020). Ohne fehlende Anreize führt dies zu „Bulimielernen", bei dem sich innerhalb kürzester Zeit vor der Klausur Inhalte angeeignet werden, um sie danach wieder zu vergessen. Auch der Praxisbezug angebotener Veranstaltungen an deutschen Universitäten wurde mehrheitlich als „mittelmäßig" bis „sehr schlecht" bewertet (Willige et al., 2018). Das Problem des mangelhaften Praxisbezugs möchten wir mit der Einführung eines innovativen Lehrkonzepts vermeiden. Zusätzlich möchten wir den Austausch zwischen Studierenden sowie das langfristige Lernen fördern.

Im Rahmen dieser wissenschaftlichen Arbeit möchten wir durch unser Lehrkonzept „Flipped Classroom 2.0" die oben genannten Problemstellungen lösen. Hauptziel des vorgestellten Lehrkonzepts ist es, die Studierenden dazu ermutigen, sich bereits während des Semesters kontinuierlich mit den Vorlesungsinhalten zu befassen und somit über die Klausur hinaus zu lernen. Hierbei gilt es, den Lernaufwand unmittelbar vor einer Klausur zu verringern und stattdessen gleichmäßig auf das Semester zu verteilen. Zudem soll es den Studierenden ermöglichen, ihren individuellen Lernstand evaluieren zu können. Durch die drei Bestandteile, bestehend aus einem Semesterprojekt, einem Bootcamp sowie bereitgestellten Lehrvideos, soll durch unser Lehrkonzept „Flipped Classroom 2.0" der Austausch zwischen Studierenden gefördert sowie durch realitätsnahe Aufgabenstellungen ein Praxisbezug hergestellt werden. Aufgrund der Corona-Pandemie wird zusätzlich ein flexibler Wechsel zwischen virtueller Durchführung und Veranstaltungen in Präsenz sichergestellt. Am Beispiel der Vorlesung „Angewandte Informatik – Internet Computing" des Instituts für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB) des Karlsruher Instituts für Technologie möchten wir diese Ziele mithilfe des pädagogischen Prinzips des „Flipped Classroom" erreichen.

Die vorliegende Arbeit ist wie folgt aufgebaut: Zunächst werden die theoretischen Grundlagen des pädagogischen Konzepts des „Flipped Classroom" erklärt. Anschließend wird die zur Erstellung des vorliegenden Lehrkonzepts angewandte Vorgehensweise beschrieben. Darauf folgt ein Überblick über das finale Konzept „Flipped Classroom 2.0". Dabei werden die Rahmenbedingungen für eine zugehörige Lehrveranstaltung bestimmt sowie die Struktur des Lehrkonzepts erklärt. Hieran anknüpfend werden die Hauptbestandteile des Lehrkonzepts, die bereitgestellte Lehrvideos, ein Semesterprojekt sowie ein Bootcamp umfassen, detailliert beschrieben. Daraufhin wird auf die Grenzen des Lehrkonzepts und auf mögliche Alternativen eingegangen. Abschließend werden die wichtigsten Aspekte des Lehrkonzepts zusammengefasst. Zudem erfolgt auf Basis seiner Vorteile eine Einordnung des Lehrkonzepts im Hinblick auf mögliche Einsatzmöglichkeiten und Ansätze für weitergehende Forschung.

Flipped Classroom

Das Lehrkonzept bedient sich dem Prinzip des „Flipped Classroom“. Bei diesem handelt es sich um ein pädagogisches Konzept, bei dem die traditionelle Vermittlung und der Erwerb von Wissen zeitlich und örtlich unabhängig erfolgt (Werner et al., 2018). Dieses Konzept ermöglich es den Lernenden, sich die Lehrinhalte selbstständig und unabhängig aneignen zu können, in dem die Lehrinhalte in der Regel in digitaler Form bereitgestellt werden. Die klassische Informationsvermittlung einer typischen Vorlesung wird ausgelagert und durch „In-Class Tasks“ und „Pre-/Post-Class Work“ ersetzt (Abeysekera & Dawson, 2015). Anstelle der typischen Vorlesung treten Aufgaben, die gemeinsam mit den Dozierenden gelöst werden. Studierende bereiten sich für diese „In-Class Tasks“ vor, in dem sie den „Pre-Class Work“ durcharbeiten und die Inhalte durch „Post-Class Work“ weiter vertiefen. Im Folgenden werden „Pre/ Post-Class Work“ als Out-of-Class Elemente und “In-Class Tasks” als In-Class Elemente bezeichnet.

Lernerfolge, die aus der Anwendung des Flipped Classroom resultieren, können auf das Prinzip des „aktiven Lernens“ zurückgeführt werden (Jensen et al., 2015). Bei diesem werden die Studierenden aktiv in den Prozess der Wissensvermittlung miteinbezogen (Prince, 2004). Dieses Prinzip besteht aus zwei Hauptbestandteilen. Zum einen werden Aufgaben und Aktivitäten der Studierenden in die gewöhnliche Vorlesung verlagert. Des Weiteren soll als zweiter Hauptbestandteil das Engagement der Studierenden gefördert werden (Prince, 2004). Durch die Implementierung des Flipped Classrooms lässt sich die gemeinsame Zeit von Studierenden und Dozierenden stärker für Aktivitäten des aktiven Lernens nutzen, da die klassische Vorlesung ausgelagert wird (Al-Samarraie et al., 2020). Beispielsweise kann die gewonnene Zeit für die Bearbeitung der Inhalte und die Behebung möglicher Schwierigkeiten verwendet werden.

Die Anwendung des Flipped Classrooms bringt verschiedene Vorteile mit sich. Beispielsweise kann eine Implementierung des Flipped Classrooms gegenüber traditionellen Konzepten zu einer Erhöhung der Zufriedenheit der Studierenden führen (Strelan et al., 2020). Auch eine allgemeine Verbesserung der Lernergebnisse der Studierenden konnte gezeigt werden. Sie ist jedoch stark abhängig vom jeweiligen Fachbereich (Cheng et al., 2019). Weniger geeignet sind Fachbereiche, die häufigen Austausch der Studierende und Dozenten benötigen, sowie Inhalte, die für eine selbständige Aneignung nicht geeignet sind und Studierende überfordern.

Aufgrund der notwendigen Vorbereitung der Studierenden für die In-Class Elemente sowie der aktiven Auseinandersetzung mit den Lehrinhalten sehen wir Flipped Classroom als geeignetes Prinzip an, um das Ziel des kontinuierlichen Lernens zu erreichen. Auch die gemeinsamen Interaktionen innerhalb der In-Class Tasks zielen auf einen vermehrten Austausch zwischen den Studierenden ab. Auch hinsichtlich der Flexibilität bietet das Prinzip des Flipped Classrooms Vorteile. Vorbereitende Materialen wie z.B. Videos können online bereitgestellt werden, ohne dass die Studierenden örtlich gebunden sind. Da „Angewandte Informatik – Internet Computing“ vor allem grundlegendes Wissen vermitteln soll, wird das Prinzip Flipped Classroom als für diese Veranstaltung geeignet angesehen.

Vorgehensweise zur Erstellung des Lehrkonzepts

Wie aus den berichteten Problemen der Studierenden hervorgeht, sind klassische Lehrformate, die auf Vorlesungen basieren, nicht optimal auf die Bedürfnisse der Studierenden abgestimmt. Vorlesungen, in denen die Fähigkeiten und Schwierigkeiten nicht explizit berücksichtigt werden, werden als wenig zielführend angesehen (Dehaene, 2020). Gleiches gilt jedoch für Formate, bei denen sich die Studierenden die Inhalte explorativ vollständig selbst erarbeiten, ohne dass zuvor inhaltliche Grundlagen den Studierenden vermittelt worden sind (Mayer, 2004). Es wird deutlich, dass eine Kombination aus der Vermittlung grundlegender Inhalte und deren aktiver Anwendung durch die Studierenden notwendig ist. Um beide Aspekte miteinander in Einklang zu bringen, wurde bei der Entwicklung daher auf eine ausgeprägte Nutzerorientierung geachtet.

Sicherstellung der Nutzerorientierung

Um die Orientierung an ihren Bedürfnissen sicherzustellen, wurde auf die sogenannte Lean-Startup-Methode zurückgegriffen (Ries, 2011). Bei dieser wird zunächst auf Basis verschiedener Ideen und Hypothesen ein erster Prototyp des Produkts oder Services entwickelt. Anschließend erfolgt dessen Testung.

Die während des Tests gewonnenen Daten werden anschließend verwendet, um den Prototypen und die ihm zugrundeliegenden Hypothesen an die tatsächlichen Bedürfnisse der Nutzenden anzupassen. Auf diese Weise ergibt sich ein kontinuierlicher Prozess, der zur stetigen Verbesserung des Angebots führt (Ries 2011). Diese Vorgehensweise ist somit besonders nutzerorientiert (Blank, 2013).

Entwicklung des Grobkonzepts

Universitäten und Dozierende haben in der Vergangenheit wenig größere Veränderungen an ihren Lehrformaten vorgenommen (Christensen & Eyring, 2011). Um einer möglichen Überforderung des Lehrpersonals durch zu schnelle Veränderungen in größerem Umfang vorzubeugen, die eine ablehnende Haltung gegenüber dem Lehrkonzept mit sich bringen könnte, galt es dies im weiteren Verlauf zu berücksichtigen. Zudem war es denkbar, dass bestimmte Wünsche der Studierenden bezüglich des Lehrkonzepts aus pädagogischer Sicht nicht zielführend sind. Daher wurde eine modifizierte Variante der Lean-Startup-Methode zur Entwicklung des Lehrkonzepts angewendet. Zu Beginn des iterativen Vorgehens sind Probleme identifiziert worden, mit denen die derzeitig angewendeten Lehrformate verbunden sind. Dazu wurde innerhalb des Projektteams ein Brainstorming durchgeführt, in welches neben Forschungsergebnissen aus der Pädagogik universitäts- und lehrstuhlspezifische Anforderungen sowie eigene Erfahrungswerte mit Lehrveranstaltungen einflossen. Durch Gespräche mit weiteren Studierenden und Dozierenden wurden die Brainstorming-Ergebnisse im Anschluss auf ihre Aussagekraft untersucht. Dabei zeichnete sich eine besondere Bedeutung der Probleme ab, die bereits in der Einleitung dieser Arbeit angeführt worden sind. Aus ihnen wurden die dort ebenfalls vorgestellten Ziele des neu zu entwickelndem Lehrkonzepts abgeleitet. Jene Ziele stellten die Grundlage für das weitere Vorgehen dar. Denn auf ihnen basierend wurde zunächst ein Grobkonzept erarbeitet. Bereits das entworfene Grobkonzept sah die Anwendung eines Flipped Classrooms vor. Dieses Prinzip spielte im weiteren Entwicklungsprozess eine entscheidende Rolle. Ausgehend von den grundlegenden Merkmalen eines Flipped Classrooms wurden iterativ Anpassungen vorgenommen, um das Lehrkonzept optimal auf die Anforderungen der betrachteten Lehrveranstaltung abzustimmen. Neben eigenen Ansätzen, die in wöchentlichen Besprechungen und unter Berücksichtigung wissenschaftlicher Erkenntnisse vorgenommen worden sind, wurde regelmäßig das Feedback von Studierenden und Dozierenden berücksichtigt. Auf diese Weise sollte sichergestellt werden, dass das entworfene Lehrkonzept wissenschaftlich fundiert ist und gleichzeitig die speziellen Anforderungen der zu überarbeitenden Lehrveranstaltung beachtet werden. Der beschriebene Prozess der kontinuierlichen Verbesserung lässt sich in zwei zeitliche Abschnitte unterteilen, die im Folgenden näher betrachtet werden.

Validierung des Grobkonzepts

Nachdem ein erstes Grobkonzept entworfen worden war, galt es, dieses zu validieren. Hierzu wurde es Studierenden und Dozierenden vorgestellt. Sowohl die Ersetzung der bisherigen Live-Vorlesungen durch bereitgestellte Videos als auch die Anwendung des Gelernten im Rahmen eines semesterbegleitenden Projekts wurden von allen Seiten begrüßt. Ein diverseres Bild zeigte sich mit Blick auf die In-Class-Elemente des Flipped Classrooms. Gleiches galt für eine mögliche Erweiterung der Veranstaltung durch zugehörige Tutorien. Hervorzuheben sind hierbei die konträren Wünsche bezüglich des Ausmaßes der aktiven Mitarbeit der Studierenden während der In-Class-Elemente. Während einige Studierende sich eine aktive Rolle wünschten, bevorzugten andere es, wenn die Dozierenden diese ohne die Mitarbeit der Studierenden durchführen. Es wird deutlich, dass ein Lehrkonzept in diesem Fall nicht den individuellen Bedürfnissen aller Studierenden gerecht werden kann. Dennoch wurde ein Kompromiss angestrebt, der für einen möglichst großen Teil der Studierenden zufriedenstellend ist. Jenen Kompromiss sollte eine Q-and-A-Session bieten. In ihr sollten zuvor eingereichte Fragen zu den Vorlesungsinhalten geklärt werden. Ergänzend waren hier Gastvorträge sowie tiefergehende Einblicke in die aktuelle Forschung durch den Dozierenden vorgesehen. Auf diese Weise sollte sowohl für Studierende mit Verständnisproblemen als auch für solche mit weiterführendem Interesse an den Vorlesungsinhalten ein Anreiz zur Teilnahme an den Q-and-A-Sessions geschaffen werden. Die gleichzeitige Teilnahme dieser beiden Arten an Studierenden ist vor allem für die Art entscheidend gewesen, nach der die eingereichten Fragen beantwortet werden sollten. Denn die Dozierenden sollten dabei primär eine unterstützende Rolle übernehmen. Die Antworten auf die Fragen sollten dagegen von den Studierenden in Kleingruppen eigenständig erarbeitet und anschließend im Plenum diskutiert werden. Auf diese Weise sollten sie sich untereinander unterstützen und die Inhalte

sich weiter verfestigen. Inwieweit die Q-and-A-Session tatsächlich ein adäquater Kompromiss zwischen den verschiedenen Wünschen ist, war in der zweiten Entwicklungsphase die prägende Leitfrage.

Lehrprobe und finale Überarbeitung

Den Auftakt der zweiten Entwicklungsphase bildete eine Lehrprobe. In dieser sollten einzelne Bestandteile des Lehrkonzepts mittels einer Simulation getestet werden. Da die anderen Bestandteile weitestgehend auf Zustimmung gestoßen sind, wurde die interaktive Q-and-A-Session während der Lehrprobe simuliert. Bei ihr bestand die größte Unsicherheit, ob sie den Erwartungen der teilnehmenden Studierenden entspricht. Im Verlaufe der durchgeführten Lehrprobe erwies sich das eigenständige und interaktive Arbeiten als nicht zufriedenstellend. Denn an der simulierten Q-and-A-Session beteiligte sich nur eine geringe Anzahl der Teilnehmenden mit aktiven Beiträgen. Die restlichen Teilnehmenden blieben sowohl während der Arbeit in den Kleingruppen als auch im Plenum passiv. Dies hatte zur Folge, dass der Dozent deutlich stärker in die Diskussion eingreifen musste, als es ursprünglich vorgesehen gewesen ist. Aus diesem Grund erschien das bisher geplante Format wenig zielführend für zukünftige Lehrveranstaltungen. Dies deckte sich mit dem Feedback der teilnehmenden Studierenden sowie des Gruppenmitglieds, das in der Lehrprobe die Rolle des Dozenten übernommen hatte. Während die Studierenden von einer gewissen Überforderung sprachen, wurde die Lehrprobe durch den Dozenten als zäh und demotivierend wahrgenommen. Kommt keine wirkliche Interaktion zwischen den teilnehmenden Studierenden zustande, ist daher davon auszugehen, dass die Motivation zur Teilnahme und der Durchführung sowohl für Studierende als auch für Dozierende sinkt. Aufgrund der gemachten Erfahrungen wurden weitere Änderungen am Lehrkonzept vorgenommen. An die Stelle der Q-and-A-Session ist das Bootcamp gerückt, bei dem eine stärkere Interaktion der Studierenden untereinander erwartet wird. Inwiefern sich das neugestaltete Bootcamp von einer klassischen Q-and-A-Session unterscheidet, wird im nächsten Kapitel ersichtlich, in dem die einzelnen Bestandteile des Lehrkonzepts ausführlich beschrieben werden. Wichtig zu betonen ist hierbei, dass eine Q-and-A-Session nicht immer so verlaufen muss, wie es in der durchgeführten Lehrprobe der Fall gewesen ist. Vielmehr kann unter Umständen auch die gewünschte Interaktion eintreten und die Q-and-A-Sessions für alle Beteiligten ein voller Erfolg sein. Denkbar ist, dass für die Q-and-A-Session gewählte Thema, Non-Fungible Tokens, dazu unpassend gewesen ist und die Inhalte von „Angewandte Informatik-Internet Computing“ unzureichend repräsentiert. Gleiches gilt möglicherweise für den Modus, in dem die Q-and-A-Session durchgeführt wird. Anders als im Lehrkonzept vorgesehen, konnte sie pandemiebedingt nur online und nicht in hybrider Form stattfinden. Zudem wich sowohl die Anzahl als auch der Studienfortschritt von den Merkmalen der Teilnehmenden der Lehrveranstaltung, die mit Hilfe des vorliegenden Lehrkonzepts überarbeitet werden soll, ab. Dennoch wird davon ausgegangen, dass innerhalb des Bootcamps, mit dem die Q-and-A-Session in diesem Lehrkonzept ersetzt wurde, die Studierenden stärker miteinander interagieren und so ein längerfristiger Lernerfolg eintritt sowie ihre Soft Skills intensiver trainiert werden.

Überblick über “Flipped Classroom 2.0“

Das finale Lehrkonzept „Flipped-Classroom 2.0“ wird auf die Veranstaltung „Angewandte Informatik – Internet Computing“ angewendet. Die gewählte Veranstaltung des AIFBs richtet sich primär an Bachelor- und Master-Studierende des Studiengangs Wirtschaftsingenieurwesen. Zudem ist die Teilnahme an der Veranstaltung für Studierende der Wirtschaftsinformatik im Bachelor verpflichtend. Gerechnet wird für diese Veranstaltung mit circa 90 teilnehmenden Studierenden. Basierend auf dieser Schätzung lässt sich der benötigte Personaleinsatz ableiten. Es ist geplant, dass jeder Tutor drei Gruppen betreut, die aus jeweils fünf Studierenden bestehen. Insgesamt werden somit ungefähr sechs Tutoren für die Veranstaltung benötigt. Darüber hinaus lässt sich zudem der Bedarf an Räumen ableiten. Das Konzept kann grundsätzlich sowohl online als auch in Präsenz stattfinden. Die Umsetzung als hybride Veranstaltung wird jedoch empfohlen. Sollten alle Studierenden vor Ort teilnehmen, werden entsprechend ausreichend Räume benötigt, in denen sich circa 15 Studierenden mit ihrem Tutor in einem Drei-Wochen-Rhytmus für 90 Minuten treffen können. Für die Studierenden wird ein eigener Computer mit einer stabilen Internetverbindung zur Teilnahme an den Treffen mit den Tutoren vorausgesetzt.

Struktur und Bewertung

„Flipped Classroom 2.0“ als Lehrkonzept sollte definierte Ziele erfüllen. Im Fokus dieser Ziele liegt die Umwandlung der Rolle der Studierende von der passiven Teilnahme zur aktiven. Damit wird die Eigenverantwortung einzelner Studierender für ein erfolgreiches Lernergebnis erheblich erhöht. Beabsichtigt wird dadurch, die oberflächliche Aufnahme der vermittelten Inhalte durch die Studierenden zu verhindern und gleichzeitig eine tiefergehende und vielschichtigere Auseinandersetzung mit den Inhalten zu fördern. Somit wird aus Sicht der Studierenden das langfristige Lernen gefördert, indem das engagierte Arbeiten während des Semesters belohnt wird. Zudem soll mehr Fokus auf das Verständnis und die Fähigkeit zur Anwendung des Gelernten statt auf reines Auswendiglernen gelegt werden. Außerdem ermöglicht das entwickelte Konzept den Studierenden ihr Projektmanagement und andere Soft Skills zu verbessern.

Um die genannten Ziele zu erreichen, wurde das Lehrkonzept basierend auf drei Hauptbestandteilen entwickelt: die hochgeladenen Lehrvideos als Lernmaterial, ein Bootcamp als Alternative zum klassischen Übungskonzept und ein Semesterprojekt, das von den Studierenden zu erarbeiten ist. Nach Ende der Vorlesungszeit müssen die Studierenden zusätzlich eine schriftliche Prüfung schreiben, um eine individuelle Evaluierung ihrer Fähigkeiten zu ermöglichen und das sogenannte Trittbrettfahrer-Problem durch eine ausschließliche Beurteilung auf Basis der Gruppenarbeit zu vermeiden.

Insgesamt umfasst die Veranstaltung 13 Wochen. Wie sich die einzelnen Termine auf diese Wochen verteilen, ist Tabelle 1 zu entnehmen. Es wird mit 120 für Wirtschaftsinformatiker bzw. 135 Stunden für Wirtschaftsingenieure als Arbeitsaufwand gerechnet. Diese beinhalten zehn Präsenzstunden (Bootcamp: 1,5h x 3, Abschlusspräsentation des Semesterprojekts: 4h und Live-Kickoff: 1,5h). Weitere 60 Stunden sind für die Arbeit am Semesterprojekt vorgesehen. Die restlichen 60 bzw. 75 Stunden können die Studierenden für das Selbststudium nutzen, um sich die Videos anzusehen und sich auf die schriftliche Prüfung vorzubereiten. Da sich die Inhalte der Veranstaltung mit denen anderer Lehrveranstaltungen überschneiden, wird für Wirtschaftsinformatik-Studierende weniger Zeit zum Selbststudium eingeplant. Durch das Bestehen der gesamten Prüfung, welche die Pflichtteilnahme am Semesterprojekt sowie an den Bootcamp-Terminen voraussetzt, werden 4 bzw. 4,5 Leistungspunkte erlangt. 50% der Endnote besteht dabei aus der Bewertung des in Gruppen erarbeiteten Semesterprojekts. Die andere Hälfte der Note wird durch die schriftliche Prüfung ermittelt. Der Umfang der schriftlichen Prüfung sollte entsprechend angepasst werden.

	1	2	3	4	5	6	7	8	9	10	11	12	13
Lehrvideos	Live-Kickoff	VL 1 - 4			VL 5 - 7			VL 8 - 11					
Bootcamp					VL 2 - 4			VL 5 - 7			VL 8 - 10		
Semester-projekt	Gruppen-bildung											Finale Präsenta-tion	

Tabelle 1. Ablaufplan

Detaillierte Beschreibung der Bestandteile

Die drei Kernbestandteile des Lehrkonzepts, Lehrvideos, Semesterprojekt und Bootcamp, sind für den gemeinsamen Einsatz konzipiert worden. Die Anwendung einzelner Bestandteile allein ist ebenfalls denkbar. In diesem Fall sollten allerdings an den im Folgenden vorgestellten Abläufen der Bestandteile Modifikationen vorgenommen werden.

Bereitgestellte Lehrvideos

Der erste Kernbestandteil des Lehrkonzepts sind bereitgestellte Lehrvideos. Diese vermitteln den Studierenden das benötigte Grundlagenwissen. Auf diese Weise werden die Inhalte vermittelt, die für das weitergehende eigenständige Lernen erforderlich sind (Mayer, 2004).

Format der Videos

Die bereitgestellten Videos ermöglichen im Vergleich zum alten Vorlesungsmodell eine zeit- und ortsunabhängigere Vermittlung der Inhalte durch den Dozierenden. Sie unterscheiden sich von Vorlesungen im klassischen Sinne, bei welchen den anwesenden Studierenden 90 Minuten lang die Inhalte erläutert werden. So sind die insgesamt elf Vorlesungseinheiten der neu gestalteten Veranstaltung „Angewandte Informatik – Internet Computing" nochmals untergliedert in drei bis vier Videos mit einer Dauer von ca. 15-20 Minuten. Somit sind zu den einzelnen Vorlesungseinheiten im Regelfall ungefähr 60 Minuten Videomaterial abrufbar. Die abrufbaren Inhalte können zudem um weitere Videos oder Materialien anderer Art ergänzt werden, die zum einen von Studierenden mit Verständnisschwierigkeiten und zum anderen von solchen mit tiefergehendem Interesse als freiwillige Zusatzmaterialien bearbeitet werden können. Es sei jedoch ausdrücklich auf die Freiwilligkeit der Bearbeitung jener Inhalte hingewiesen. Denn die Bereitstellung der Zusatzmaterialien dient primär zur Erhaltung des Interesses der Studierenden an den Inhalten der Veranstaltung, indem diese sich weder unter- noch überfordert fühlen (Dehaene, 2020).

Die Lehrvideos im Umfang von ca. 60 Minuten je Vorlesungseinheit sollten so gestaltet werden, dass sie zum vollständigen Verständnis des Kapitels genügen. Dabei ist das zeitliche Limit von 20 - 30 Minuten pro Lehrvideo zwingend einzuhalten. Dies ist auch empfehlenswert, obwohl die Aufmerksamkeit der Studierenden nicht zwingend auf diesen Zeitraum beschränkt ist (Gerbig-Calcagni, 2010). Denn wie im Austausch mit dem Dozenten der Lehrveranstaltung „Angewandte Informatik- Internet Computing" deutlich wurde, werden diese und andere Veranstaltungen seines Lehrstuhls tendenziell als überdurchschnittlich arbeitsintensiv wahrgenommen. Dies wird von ihm vor allem darauf zurückgeführt, dass von den Studierenden in seinen Lehrveranstaltungen während des Semesters eine intensivere Mitarbeit als an anderen Lehrstühlen erwartet wird. Das neuentwickelte Lehrkonzept könnte diese Wahrnehmung weiter verstärken, da es ebenfalls stark auf die aktive Mitarbeit der Studierenden setzt. Um dem vorzubeugen und die mentale Hürde zum Ansehen der Videos zu minimieren, erscheint das vorgegebene zeitliche Limit geeignet. Denn kürzere Videos lassen sich leichter in den Zeitplan der Studierenden integrieren und reduzieren den Anreiz dafür, die Videos einer höheren Geschwindigkeit abzuspielen. Letzteres könnte sonst die Menge der Informationen reduzieren, die Studierende sich nach dem Ansehen merken. Dies überschneidet sich mit den Erfahrungen aus der Umsetzung anderer Lehrkonzepte, die auf einem Flipped Classroom basieren (Hoyer & Mundt, 2014).

Neben der reinen Länge ist auf eine hohe didaktische Qualität der Videos zu achten. Gerade bei Grundlagen-Veranstaltungen wie „Angewandte Informatik - Internet Computing" scheint die Verlockung jedoch groß zu sein, schnell und ohne großartige Vorbereitung Lehrvideos aufzuzeichnen, da es insbesondere bei diesen inhaltlich selten zu Veränderungen kommt. Ein solches Vorgehen würde im Rahmen des entwickelten Lehrkonzepts aller Voraussicht nach gleich bei mehreren Generationen an Studierenden zu geringerer Zufriedenheit und verschlechterten Lernerfolgen führen. Denn aufgrund der eher statischen Inhalte der Veranstaltung ist geplant die Videos über mehrere Semester hinweg zur Wissensvermittlung einzusetzen. Dadurch sinkt auf den ersten Blick der zeitliche Aufwand des Lehrenden gerade in den auf die Neueinführung des Lehrkonzepts folgenden Semestern erheblich. Dies sollte allerdings dadurch ausgeglichen werden, dass zuvor zusätzliche Zeit in die Produktion hochwertiger Videos investiert wird. Zudem ist ein kontinuierlicher Verbesserungsprozess anzuwenden. Feedback durch Studierende gilt es hierbei gezielt einzusetzen. Damit werden zwei Ziele verfolgt. Zum einen erlangt der Dozent durch die Rückmeldungen mit der Zeit ein besseres Verständnis dafür, wie verständlich seine jeweiligen Erklärungen sind. Dieses Verständnis lässt sich dazu einsetzen, die schwer verständliche Abschnitte der Lehrvideos zu ersetzen, wovon vor allem die Teilnehmenden der Lehrveranstaltungen in späteren Semestern profitieren. Zum anderen kann das Feedback der Studierenden bereits im jeweils aktuellen Semester aufgegriffen werden. Sollten bestimmte Lehrvideos von ihnen als schlecht verständlich eingestuft werden, kann der Dozent an dieser Stelle durch eine neue Version des Videos oder weiterführende Erklärungen frühzeitiger eingreifen. Damit wird die Interaktion zwischen Studierenden und Dozent gefördert und der unidirektionale Charakter der Lehrvideos geschmälert. Dies stimmt mit der

Forderung nach der Berücksichtigung der Stärken und Schwächen der an der Lehrveranstaltung teilnehmenden überein (Dehaene, 2020). Um beide Ziele zu erreichen, wird die einmalige Lehrevaluation am Ende des Semesters durch ein Feedback-Tool ersetzt, mithilfe dessen die Studierenden ohne großen Aufwand Rückmeldung zu den bereitgestellten Materialien geben können. Die ergänzende Beratung durch didaktische Fachkräfte während der Erstellung und Anpassung der Lehrmaterialien ist zu empfehlen.

Zeitlicher Ablauf

Das Semester beginnt mit einem Live-Kickoff, der in der ersten Woche in hybrider Form stattfindet. Dort werden die verschiedenen Komponenten der Veranstaltung, der zeitliche Rahmen sowie das Bewertungsverfahren ausführlich vorgestellt. Außerdem werden offene Fragen geklärt. Ab der zweiten Woche erfolgt dann der blockweise Upload der Lehrvideos. Zunächst werden die ersten vier Kapitel freigeschaltet, wobei es sich beim ersten Kapitel um eine kurze inhaltliche Einführung handelt. Alternativ kann die inhaltliche Einführung bereits in den Live-Kickoff integriert werden. Der nächste Upload erfolgt drei Wochen später mit den Einheiten Fünf bis Sieben, während in der achten Woche die Videos zu den vier letzten Kapiteln hochgeladen werden. Wie aus dem Upload-Rhythmus der insgesamt elf Kapitel ersichtlich ist, ist dabei eine Vorlesungseinheit pro Woche eingeplant. Wann genau die Studierenden sich mit den Vorlesungsinhalten auseinandersetzen, bleibt allerdings ihnen selbst überlassen. Durch die blockweise Bereitstellung erhalten sie die Möglichkeit, flexibel die Arbeitszeit auf die einzelnen Bestandteile der Veranstaltung aufzuteilen. Um besser auf die Bedürfnisse der einzelnen Jahrgänge an Teilnehmenden an „Angewandte Informatik- Internet Computing“ eingehen zu können, wird jedoch von der gleichzeitigen Veröffentlichung aller Vorlesungseinheiten abgesehen, damit das Feedback zu den vorherigen Blöcken entsprechend in den weiteren Videos berücksichtigt und diese ggf. angepasst werden können. Gleichzeitig sollen die Studierenden dazu angeregt werden, sich kontinuierlich mit den Inhalten zu beschäftigen, statt diese auf einmal zu konsumieren. Es wird davon ausgegangen, dass bei gleichzeitiger Bereitstellung der Videos die subjektive Wahrnehmung des aufzubringenden Aufwands deutlich höher ausfällt. Dies kann am Anfang des Semesters den ersten Eindruck bezüglich der Veranstaltung prägen und sich entsprechend negativ auf die Zufriedenheit mit dieser auswirken.

Quiz zur Evaluation

Zusätzlich werden zu jeder Vorlesungseinheit ungefähr acht Fragen in einem Ilias-Quiz gestellt. Das Quiz soll den Anreiz dafür senken, ohne Pause die Videos zu schauen. Zusätzlich werden mithilfe des Quizes psychologische Forschungsergebnisse miteinbezogen, die auf eine Verbesserung des langfristigen Lernens durch Tests hindeuten (Roediger et al., 2011). Die Beantwortung der Fragen soll entsprechend bei der Reflexion des Gelernten unterstützen, ist jedoch freiwillig. Denn es ist davon auszugehen, dass sich ein zu geringer Schwierigkeitsgrad der Fragen negativ auf die Motivation zum weiteren Ansehen der Videos auswirkt, wenn ihre Beantwortung verpflichtend ist. Die Wahl eines geeigneten Schwierigkeitsgrads erscheint aufgrund der hohen Diversität der Gruppe an Teilnehmenden an der Lehrveranstaltung u.a. im Hinblick auf ihren Studienfortschritt schwierig. Daher wird am Ende des jeweiligen Videos auf die Fragen hingewiesen, sie können aber jederzeit übersprungen werden. Ob die richtigen Antworten zu den Quiz-Fragen auf Online-Plattformen wie z.B. Studydrive oder in Chat-Gruppen geteilt werden, ist aus diesem Grund zweitrangig. Da die Bearbeitung der einzelnen Fragen freiwillig ist und keinerlei Einfluss auf die Benotung hat, werden aller Voraussicht nach vor allem die Studierenden diese beantworten, die ihre Fähigkeiten überprüfen möchten. Daher würden sie sich selbst schaden, wenn sie sich im Vorfeld die Lösungen anschauen würden. Entsprechend können die Fragen auch in mehreren aufeinanderfolgenden Semestern eingesetzt werden.

Semesterprojekt

Ein wichtiger Bestandteil des Lehrkonzeptes ist das Semesterprojekt, welches im Folgenden detailliert betrachten wird. Das Semesterprojekt dient als ein Out-of-Class Element des Lehrkonzepts, für welches sich die Studierenden miteinander in kleineren Gruppen ohne direkte Beteiligung eines Lehrenden mit den Lehrinhalten an einem Use Case beschäftigen. Um das Semesterprojekt von den anderen Out-of-Class Aktivitäten unseres Lehrkonzepts abzugrenzen, ist es wichtig zu betonen, dass das Semesterprojekt nicht zur Vermittlung neuer Inhalte, sondern als Möglichkeit zum Vertiefen und Anwenden des bereits Gelernten konzipiert ist.

Gruppenbildung

Für das Semesterprojekt werden die Studierenden in feste Gruppen eingeteilt. Diese bestehen aus bis zu fünf Personen, die bis zur Abgabe des Semesterprojekts zusammenarbeiten. Die Entscheidung für ein gruppenbasiertes Semesterprojekt folgt u.a. Erkenntnissen, welche untermauen, dass der Lernerfolg von Studierenden in kleineren Gruppen von drei bis fünf Personen höher ausfällt als für Studierende, die sich allein mit den Inhalten auseinandersetzen (Gillies, 2016). Unter der Annahme, dass sich ca. 90 Studierende für dieses spezielle Kursprogramm anmelden werden, würden 18 Gruppen gebildet. Diese 18 Gruppen werden gleichmäßig den vom Lehrstuhl vorher entworfenen Use Cases zugeteilt. Dabei sollten sich maximal fünf bis sechs Gruppen mit demselben Use Case beschäftigen. Entsprechend sind drei verschiedene Use Cases vorzubereiten. Die Gruppeneinteilung findet über ein Online-Portal statt, auf welchem die Use Cases durch die Studierenden ihren persönlichen Interessengebieten entsprechend bewerten werden können. Basierend auf dem Interessen der Studierenden an den einzelnen Use Cases werden im Anschluss die Gruppen zugeteilt.

Erläuterung der Use Cases

Die Use Cases sind als eine vereinfachte und flexiblere Alternative zu einer üblichen Case Study gedacht. Denn unter dem Begriff Case Study versteht man im Allgemeinen eine realitätsnahe Problemdarstellung eines bestimmten Falles aus der Praxis, bei dem durch gezielte Fragestellungen bestimmte Aspekte besonders hervorgehoben werden (Lasch & Schulte, 2006). Im Gegensatz dazu ist die praktische Problemstellung der Use Cases des Semesterprojekts abstrakter gefasst. Gleichzeitig werden für die Anwendung des Lehrinhalts auf die Problemstellung lediglich Leitfragen gegeben. Konkrete Lösungsansätze für die gegebene Problemstellung und den Transfer der Inhalte sind von den Studierenden selbst zu erarbeiten. Auf diese Weise soll zum einen die Motivation der Studierenden gefördert werden, da sie ihre Kreativität stärker ausleben können. Zum anderen erschwert dies den Austausch von Lösungen unter den Studierenden, da die Problemstellung gröber gefasst ist. Es ist dafür Sorge zu tragen, dass innerhalb des Jahrgangs nicht zwei Gruppen an einer zu ähnlichen Lösung arbeiten. Gleiches gilt für den Vergleich mit abgegebenen Semesterprojekts aus vorangegangenen Jahren. Zusätzlich reduziert die eher grobe Ausgestaltung der Problemstellung den Arbeitsaufwand, der für den Lehrstuhl mit dem Entwurf der Use Cases einhergeht. Die Berücksichtigung aktueller Forschungsprojekte ist bei diesem Entwurf ausdrücklich erwünscht. Auf Basis des Use Cases sollen die Studierenden zunächst ein Konzept für eine Lösung der Problemstellung erarbeiten. Dieses kann die Entwicklung eines neuartigen Geschäftsmodells oder eine Nachahmung einer bereits bestehenden Unternehmung aus der realen Welt sein. Beispiele für solch einen Use Case und mögliche zugehörige Lösungsansätze für die diesem zugrundeliegdene Problemstellung werden auf Nachfrage von den Autoren dieser Arbeit bereitgestellt. Sobald ein erstes Konzept zur Bearbeitung des Use Cases durch die Kleingruppen entworfen worden ist, erfolgt der Hauptteil des Semesterprojekts. In diesem sind die vermittelten Inhalte aus der Vorlesung auf den jeweiligen Anwendungsfall zu übertragen. Wie genau die Gruppen die verschiedenen Aspekte aus den einzelnen Vorlesungen umsetzen möchten, bleibt ihnen überlassen. Damit der Arbeitsaufwand nicht das festgelegte Maximum an Stunden übersteigt, werden Wahlpflicht-Themengebiete angeboten. So können die Gruppen entscheiden, ob sie beispielsweise lieber die Inhalte zu „Internet of Things“ oder zu „Distributed Ledger Technology“ in ihrem Projekt berücksichtigen. Dadurch wird zudem vermieden, dass die Gruppen Technologien in ihren Konzepten berücksichtigen müssen, die für ihr Geschäftsmodell nicht relevant sind. Außerdem sollte der Lehrstuhl zu einzelnen Kapiteln einige Leitfragen vorbereiten, sodass für die Gruppen auch eine Diskussionsgrundlage und Orientierungshilfe geschaffen wird.

Ablauf des Semesterprojekts

Am Ende des Semesters sind die Arbeitsergebnisse in Form einer schriftlichen Arbeit sowie der Abschlusspräsentation zusammenzufassen. Die konkreten Anforderungen im Hinblick auf geforderte Seitenanzahl, inhaltliche Ansatzpunkte mit kleinen Beispielen und Abgabetermine sind möglichst frühzeitig zu kommunizieren. Dies kann wahlweise über das Modulhandbuch oder die Website des Lehrstuhls erfolgen. Zusätzlich ist der Live-Kickoff in der ersten Vorlesungswoche hierfür eingeplant. Hinter dieser detaillierten Bekanntmachung steckt die Intention, dass sich die Studierenden bevor sie sich in einer Gruppe zusammenfinden, darüber bewusstwerden, dass das Semesterprojekt eine hohe Leistungs- und Kommunikationsbereitschaft sowie Motivation und Eigenverantwortung erfordert. Nach der

Gruppeneinteilung ist das oben beschriebene Grobkonzept des Semesterprojekt-Vorhabens zu erstellen. Obwohl das Semesterprojekt vor allem auf eigenständiges Arbeiten setzt, findet eine Überprüfung des Grobkonzepts durch den betreuenden Tutor statt. Die Einreichung des Grobkonzepts sollte in der ersten Hälfte des Vorlesungszeitraums erfolgen. Dadurch soll sichergestellt werden, dass das Konzept der Gruppe den Erwartungen an das Semesterprojekt entspricht und es zu keinen inhaltlichen Überschneidungen zu anderen Gruppen kommt. Zusätzlich werden die Studierenden dadurch animiert, frühzeitig mit der Arbeit am Semesterprojekt zu beginnen. Nach der sich anschließenden Anwendung der Lehrinhalte auf die konkrete Problemstellung wird am Ende des Vorlesungszeitraums, ca. zwei Wochen vor der Hauptklausur, die schriftliche Ausarbeitung abgegeben. Zusätzlich stellen die Gruppen ihre Ergebnisse in einer 15-minütigen Präsentation den anderen Studierenden vor, mit denen sie während des Semesters im Bootcamp zusammengearbeitet haben. Der Umfang der schriftlichen Ausarbeitung, welche als ausschließliche Bewertungsgrundlage des Semesterprojekts dienen wird, beschränkt sich auf 12 Seiten. Bewertet wird vor allem der Transfer der Inhalte der Lehreinheiten auf den jeweiligen Use Case. Für die Ausarbeitung erhält die Gruppe eine gemeinsame Note, die 50% der Gesamtnote ausmacht.

Vorteile des Semesterprojekts

Die beschriebene Umsetzung des Semesterprojekts fördert gezielt das kollaborative und kooperative Lernen, bei welchem die Studierenden sich innerhalb einer Gruppe austauschen (Lage et al., 2000). Darüber hinaus motiviert das Projekt die Studierenden zur kontinuierlichen Auseinandersetzung mit den Inhalten. Da durch das Semesterprojekt ein kooperatives Lernumfeld geschaffen werden soll, welches von tiefgreifenden Diskussionen innerhalb der Gruppe lebt, ist zu vermuten, dass sich der Lernprozess auch qualitativ steigert. Aus Studien geht hervor, dass kooperatives Lernen im Vergleich zum kompetitiven und selbständigen Lernen zu einem verbesserten Lernerfolg sowie höherer Sozialisation und Motivation führt und persönliche Weiterentwicklung der Studierenden stärker fördert (Gillies, 2016). Außerdem bietet die Gruppenarbeit die Möglichkeit, die Fähigkeiten der Teilnehmenden zur sozialen Interaktion zu trainieren. Jedem Gruppenmitglied sollte bewusst sein, dass die Note der schriftlichen Ausarbeitung einen erheblichen Einfluss auf die Gesamtnote hat. Diese gegenseitige Abhängigkeit wird zum einen die Motivation zur Teilhabe an der gruppeninternen Zusammenarbeit verstärken und zum anderen aufzeigen, wie man sich gegenüber anderen Gruppenmitgliedern verhalten sollte. So können die Studierenden neue Erfahrungen sammeln, wie man mit Problemen während der Zusammenarbeit mit unterschiedlichen Persönlichkeitstypen umgeht und lernen zudem, wie man das eigene Wissen vertieft, indem man die Ansichten anderer aufnimmt (Kuh, 2008). Auf eine möglichst ausgewogene Verteilung der Arbeitslast innerhalb der Gruppen ist zu achten. Sogenanntes Trittbrettfahren, bei dem sich einzelne Gruppenmitglieder im geringeren Umfang an der Bearbeitung des Semesterprojekts beteiligen, und vom zusätzlichen Engagement der anderen profitieren, gilt es zu vermeiden. Eine Möglichkeit hierzu könnten anonyme Feedbackbögen sein, mit denen die Gruppenmitglieder ihre Mitarbeit am Projekt gegenseitig beurteilen. Die Ergebnisse dieser Feedbackbögen würden entsprechend bei der individuellen Benotung berücksichtigt werden. Sollte sich auch hierdurch keine Besserung zeigen, ist es denkbar, dass Semesterprojekt von Gruppen- auf Einzelarbeit umzustellen.

Bootcamp

Neben den bereits beschriebenen Lehrvideos sowie dem Semesterprojekt stellt das sogenannte Bootcamp als In-Class-Element die dritte Komponente des Lehrkonzepts dar. In diesem sollen die theoretischen Konzepte aus den Lehrvideos aufgegriffen und die Studierenden auf deren Anwendung im Rahmen des Semesterprojekts vorbereitet werden. Dazu werden mehrere Bootcamp-Gruppen gebildet.

Zusammensetzung des Bootcamps

Die Bootcamp-Gruppen setzen sich aus jeweils drei Gruppen des Semesterprojekts zusammen und werden von einem Tutor betreut. Basierend auf dem alten Lehrkonzept wurden im Sommersemester 2021 parallel sechs verschiedene Tutorien zu „Angewandte Informatik- Internet Computing“ angeboten. Dabei wurde von vier Tutoren jeweils eine Tutoriums-Gruppe betreut, während ein Tutor zwei Gruppen übernommen hat. Äquivalent hierzu könnten bei gleichbleibender Anzahl an Tutoren sechs Bootcamp-Gruppen begleitet werden, welche sich wiederum aus insgesamt 18 Gruppen des Semesterprojekts zusammensetzen. Der vorgesehenen Gruppengröße von fünf Studierenden für das Semesterprojekt entsprechend könnten somit

bis zu 90 Studierende von den sechs Tutoren begleitet werden. Abhängig von der erwarteten Anzahl an Teilnehmenden an der Lehrveranstaltung kann es erforderlich sein, zusätzliche Tutoren einzustellen. Unklar ist derzeit noch der Arbeitsaufwand, der durch Rückfragen und individuelle Besprechungen mit den Gruppen des Semesterprojekts entsteht. Sollte sich dieser, wie erwartet, in Grenzen halten, erscheint es zielführender, die Tutoren zusätzliche Gruppen begleiten zu lassen. Dies wird insbesondere deshalb als zielführend erachtet, da der Aufwand zur Vorbereitung der Bootcamp-Treffen im Vergleich zu den bisherigen Tutorien deutlich abnimmt. Abgesehen vom gesonderten Termin für die Abschlusspräsentationen des Semesterprojekts finden drei Treffen der Bootcamp-Gruppen statt, die gleichmäßig über das Semester verteilt sind. Jeder der drei Termine deckt inhaltlich drei Kapitel der Vorlesung ab. Insgesamt werden somit neun der elf Vorlesungskapitel im Rahmen des Bootcamps intensiver besprochen. Der Ablauf ist auch Tabelle 1 zu entnehmen. Zu den Kapiteln 1 und 11 erfolgt keine ausführliche Besprechung der Inhalte, da diese als Einführung bzw. als Ausblick gedacht sind. Entsprechend werden die Kapitel 2 bis 4 (erstes Treffen), 5 bis 7 (zweites Treffen) und 8 bis 10 (drittes Treffen) gemeinsam betrachtet. Zur Teilnahme an den drei Terminen sind alle Studierenden verpflichtet. Dies stellt eine Voraussetzung für die Prüfungsleistung dar. Abgehalten werden können die Treffen wahlweise über Videokonferenzen oder in Seminarräumen. Entsprechende Präferenzen der Studierenden werden bei der Anmeldung zum Semesterprojekt und der damit verbundenen Gruppenbildung berücksichtigt.

Ablauf der Treffen

Die Treffen an sich sind zeitlich untergliedert in weitere Teilabschnitte, die sich wiederholen. Dabei wird von jeder der drei Gruppen des Semesterprojekts, die gemeinsam eine Bootcamp-Gruppe bilden, zu einem der drei Kapitel eine Kurzzusammenfassung vorbereitet und den anderen Studierenden vorgestellt. Mit den Kurzzusammenfassung werden gleich mehrere Ziele verfolgt. Zum einen sind die Studierenden während der Vorbereitung dazu gezwungen, sich so mit dem Vorlesungsstoff auseinander zu setzen, dass sie wichtige Inhalte von eher unwichtigerem unterscheiden können. Dadurch müssen sie sich mit dem gesamten Inhalt des Kapitels und den zugrundeliegenden Konzepten aktiv auseinandersetzen, um die Inhalte richtig einordnen zu können. Auf diese Weise werden Impulse zur weiteren Festigung des Gelernten im Gehirn gegeben (Ullmann, 2016). Dasselbe trifft auf die Studierenden zu, die bei der Vorstellung der jeweiligen Kurzzusammenfassung zuhören. Da die Lerninhalte in dieser in anderer Form als zuvor in den Lehrvideos vorgestellt werden, werden auch bei ihnen neue Impulse gesetzt. An die fünf bis zehn-minütige Kurzzusammenfassung schließt sich ein ca. zehnminütiges Quiz an, welches im Vorfeld von der präsentierenden Gruppe in Ilias erstellt wurde. Dazu werden alle Studierenden einer Bootcamp-Gruppe einem gemeinsamen Ilias-Kurs zugewiesen und erhalten die Rechte, die zum Anlegen der Quiz-Fragen von Nöten sind. Das vorbereitete Quiz wird nun von den anderen Studierenden in Einzelarbeit beantwortet. Abschließend werden die Lösungen zu den Quiz-Fragen besprochen und häufig aufgetretene Fehler genauer diskutiert. Durch die Integration des Quizes in die Treffen werden erneut die Erkenntnisse zur positiven Wirkung von Tests auf den langfristigen Lernerfolg aufgegriffen (Roediger et al. 2011). Sobald die beschriebenen Schritte zu einem Kapitel durchlaufen wurden, werden diese für die nächsten beiden Kapitel, die von den anderen zwei Gruppen vorbereitet wurden, wiederholt. Der Fokus liegt aus den genannten Gründen während des gesamten Treffens darauf, dass die Studierenden sich die Inhalte selbst gegenseitig nochmals erklären und sich bei offenen Fragen unterstützen. Dadurch soll der Vorlesungsstoff weiter gefestigt sowie Verständnisschwierigkeiten, die nicht bereits in den Semesterprojekt-Gruppen geklärt werden konnten, besprochen werden. Daher ist der betreuende Tutor bei den Treffen primär zur Qualitätskontrolle anwesend und soll lediglich einschreiten, wenn inhaltliche Fragen nicht von den Studierenden allein geklärt werden können.

Benotung des Bootcamps

Eine Benotung der erstellten Kurzzusammenfassungen und Quiz-Aufgaben erfolgt nicht. Lediglich die Anwesenheit bei allen drei Terminen wird vorausgesetzt. Es wird zum Zeitpunkt des Entwurfs dieses Lehrkonzepts davon ausgegangen, dass der soziale Druck sowie die kontinuierliche Auseinandersetzung mit den Vorlesungsinhalten im Rahmen des Semesterprojekts zu einer ausreichenden Motivation führen. Zudem wird die Reihenfolge variiert, nach der den Teilgruppen einer Bootcamp-Gruppe die Kapitel zugewiesen werden, damit sich der zeitliche Druck während der Vorbereitung fair auf die einzelnen Gruppen verteilt. So bereitet die erste Gruppe für das erste Treffen Kapitel Zwei vor. Somit hat sie den

größten Abstand zwischen der Woche, in der das Anschauen des Kapitels vorgesehen ist, und der des Treffens. Beim zweiten Treffen ist die Gruppe dagegen für das sechste Kapitel zuständig und hat statt drei zwei Wochen Zeit zur Vorbereitung von Kurzzusammenfassung und Quiz. Beim letzten Treffen verantwortet sie die Vorbereitung des zehnten Kapitels und hat entsprechend eine Woche Abstand zum Treffen. Sollte trotz der beschriebenen Maßnahmen die Qualität der Kurzzusammenfassungen und der Quiz-Fragen nicht zufriedenstellend sein, können diese alternativ auch als Teil der Prüfungsleistung mitberücksichtigt werden. Da jedoch nicht auszuschließen ist, dass Kurzzusammenfassung und Quiz zwischen verschiedenen Gruppen oder Studienjahrgängen ausgetauscht werden, ist die Berücksichtigung im Rahmen der Prüfungsleistung aus unserer Sicht nicht erstrebenswert. Außerdem droht sich in diesem Fall, der Charakter des Bootcamps zu ändern. Denn bei diesem steht das gemeinsame Lernen und gegenseitige Unterstützen im Vordergrund, um die benötigte Grundlage zum eigenständigen Arbeiten im Rahmen des Semesterprojekts zu schaffen (Mayer, 2004). Denkbar ist des Weiteren, dass die von den Gruppen erstellten Quiz-Fragen nach den Treffen auch den anderen Bootcamp-Gruppen in einem Fragenpool zur Verfügung gestellt werden und die vom Lehrstuhl zu den Lehrvideos gestellten Fragen ergänzen. Dies wäre eine weitere Möglichkeit zur individuellen Vorbereitung auf die Abschlussklausur, würde aber auch die Anzahl an Studierenden erhöhen, die Zugang zu Quiz-Fragen zu einzelnen Teilkapiteln haben. Letzteres ist explizit zu begrüßen, um den Studierenden regelmäßige Tests des eigenen Wissens zu ermöglichen, die das Lernen verbessern (Roediger et al., 2011). Dadurch steigt allerdings die Wahrscheinlichkeit dafür, dass Fragen an zukünftige Jahrgänge weitergegeben werden, und senkt weiter den Anreiz, diese als Teil der Prüfungsleistung zu betrachten.

Grenzen und Alternativen des Lehrkonzeptes

Dieses Lehrkonzept hat jedoch neben seinen Stärken, wie beispielsweise der Förderung des kontinuierlichen Lernens, der Vertiefung des Lehrinhalts durch das Semesterprojekt und der gesteigerten Flexibilität, welche vor allem in Zeiten der Pandemie einen erhöhten Stellenwert aufweist, auch gewisse Grenzen. Dementsprechend ist eine genaue vorherige Analyse der Ausgangssituation essenziell, um die Potenziale des Konzeptes ausschöpfen zu können. Diese Analyse sollte im Vorfeld durch den Lehrstuhl durchgeführt werden, um die Struktur (Inhalt, Anzahl der Studierenden, zu verwendende pädagogische Mittel wie beispielsweise Übungsaufgaben) der Teilleistung zu verstehen und die Anwendbarkeit des Lehrkonzeptes zu prüfen. Das Hauptanwendungsgebiet findet dieses Konzept in sehr inhaltsvollen Fächern, welche eine geringere Transfertiefe besitzen.

Eigenschaften der Lehrveranstaltung

Eine Anwendung des Lehrkonzeptes wäre bei sehr stark besuchten Veranstaltungen mit einer Gesamtanzahl von mehr als 200 Studierender nicht mehr handhabbar. Zum einen müssten in diesem Fall sehr viele Tutoren dem Lehrstuhl zur Verfügung stehen. Zum anderen würde dies auch einen erhöhten Aufwand für den Dozenten bzw. dessen Lehrbeauftragten bedeuten, da aufgrund der gesteigerten Anzahl an Kleingruppen eine Vielzahl an Use Cases generiert werden müssten.

Neben der Veranstaltungsgröße spielt jedoch auch die Thematik der Veranstaltung eine ausschlaggebende Rolle. Stark theoretische Themenbereiche mit weniger Praxisbezug sind für dieses Lehrkonzept wegen der beschränkten Möglichkeit, Use Cases zu erstellen, nicht geeignet.

Ebenso trifft dies auf Vorlesungsinhalte zu, die komplexere Übungsaufgaben erfordern, um den Lehrstoff weiter zu vertiefen, da in diesem Konzept kein Element für die Bearbeitung dieser Aufgaben vorgesehen ist.

Austausch von Lösungen

Findet das vorliegende Lehrkonzept Anwendung, so sind weitere Herausforderungen in der Praxis zu erwarten. Wird dieses Konzept über mehrere Semester hinweg angewandt, so wird es immer schwieriger werden, Use Cases zu erstellen, die sich von den Themen der vorherigen Semester abgrenzen. Nützlich wäre es hier, falls möglich, aktuelle Entwicklungen einzubeziehen, um neue Aspekte einzubringen. Dies wird vermutlich aber nicht immer möglich sein, weshalb es zu sich ähnelnden Use Cases kommen kann. Da die Semesterprojekte zu einem Teil mit in die Bewertung einfließen, ist es naheliegend deren Lösung bzw. Lösungsansatz zu verbreiten oder im Internet zu veröffentlichen. Dadurch läuft das Lehrkonzept Gefahr, dass selbstständige Arbeiten zu vernachlässigen. Daher ist es von Bedeutung, die Use Cases in spezifischen

Aspekten so abzuwandeln, dass zumindest neue Ansätze mit eingebracht werden müssen. Demzufolge wäre auch die Nutzung verbreiteter Unterlagen aus Vorjahren kein Ausschlusskriterium mehr, da sich dadurch die Studierenden dennoch mit dem Stoff auseinandersetzen und ergänzende Punkte erarbeiten müssen. Die bekannten Unterlagen würden hier nur Denkanstöße liefern und keine Lösung für den gesamten Use Case darstellen. Dies lässt sich analog auf Zusammenfassungen beziehen, da diese ebenfalls nicht die gesamte Fragestellung des Use Cases umfassen würden.

Individuelle Bedürfnisse der Studierenden

Der zuvor als positiv deklarierter Aspekt des selbständigen Lernens und Arbeitens kann für Studierende, die über eine schlechte Selbstorganisation verfügen, gleichzeitig eine Hürde darstellen. Die Studierenden haben viele Freiheiten und müssen sich selbständig organisieren, da das Lehrkonzept wenig verpflichtende Termine vorsieht. Studierende, denen diese Form des selbstständigen Arbeitens grundsätzlich Schwierigkeiten bereitet, weisen hier unter Umständen weniger Motivation auf, sich einzufinden. Andererseits kann die Teilnahme an der Lehrveranstaltung auch als Anreiz wirken, um ihre Selbstorganisation zu verbessern.

Da das Lehrkonzept auf das Arbeiten in Gruppen ausgelegt ist, kann dies abschreckend für sehr introvertierte Studierende sein, die wenig Kontakte an der Universität haben. In deren Augen wären andere Teilleistungen unter Umständen attraktiver. Auf der anderen Seite kann dies jedoch auch als Chance gesehen werden, gerade im Zusammenhang mit der Corona-Pandemie, in der die wenigsten Kurse in Präsenz stattfinden, neue Kontakte zu knüpfen und Anschluss zu finden.

Wie bei jeder gruppenbasierenden Arbeit besteht auch hier die Gefahr des Trittbrettfahres. Wenn mehrere Personen zusammenarbeiten, können in gewissen Fällen Einzelpersonen dies für sich nutzen, um weniger Aufwand in die Arbeit zu stecken als andere Gruppenmitglieder und dennoch an deren erbrachter Arbeit zu partizipieren. Um diese Gefahr abzuschwächen, sind die Präsentationen der Zusammenfassungen und der Use Cases der Gruppe hilfreich, da hierzu sich alle Gruppenmitglieder mit dem Thema auseinandersetzen müssen.

Eine weitere Herausforderung könnte sein, dass das Konzept im Modulhandbuch zunächst abschreckend auf Studierende wirkt, da durch die Zusammenfassungen und Use Cases der Eindruck eines unverhältnismäßig hohen Zeitaufwandes erweckt werden könnte. Hierzu lässt sich jedoch anmerken, dass durch die Use Cases bereits 50% der Prüfungsleistung erbracht werden und somit der Lernaufwand in der Prüfungsphase reduziert wird. Die im Bootcamp erarbeiteten Zusammenfassungen senken den Aufwand während der Prüfungsphase zusätzlich. Auf diese Reduktion des Arbeitsaufwands sollte spätestens während der Kickoff-Veranstaltung bzw. im Idealfall vor Beginn der Lehrveranstaltung hingewiesen werden.

Fazit

Abschließend folgt ein kurzes Resümee des Lehrkonzepts, bevor auf die Implikationen und Limitationen dieser Arbeit eingegangen wird.

Zusammenfassend besteht das vorgestellte Lehrkonzept aus drei wesentlichen Bestandteilen. Der Wissensvermittlung mit Hilfe von Uploadvideos, der Wissensfestigung durch Bootcamps als auch der Wissensanwendung durch die Teilnahme an dem Semesterprojekt. Bei letzterem wird einer Gruppe aus ca. fünf Studierenden ein Use Case zugeteilt, der den theoretischen Lehrinhalten der Teilleistung einen Praxisbezug verleihen soll. Die Bearbeitung des Use Cases erfolgt in Form einer Projektarbeit, welche benotet wird und zusammen mit der erreichten Note der schriftlichen Abschlussklausur die Endnote für die Lehrveranstaltung bildet. Fortlaufend über das Semester hinweg werden ca. 15-minütige Lehrvideos hochgeladen und stehen den Studierenden nach Upload jederzeit zur Verfügung. Hierdurch wird den Studierenden ein hohes Maß an Flexibilität eingeräumt. Die Lehrvideos sollen den Lehrinhalt didaktisch hochwertig vermitteln, da diese als primäre Wissensquelle der Studierenden fungieren. Ein an jedes Lehrvideo anschließendes Quiz sorgt neben einer Vertiefung der Inhalte, für eine Selbstevaluierung der Studierenden während des Semesters. Zudem erklären sich die Studierenden während drei Bootcamp-Treffen die Lehrinhalte gegenseitig vor, wodurch ihr Wissen bezüglich der Lehrveranstaltung gefestigt

werden soll. Des Weiteren werden die Studierenden auf diese Weise bereits während des Semesters intensiv mit den Lehrinhalten vertraut gemacht.

Flipped Classroom 2.0 bietet Dozierenden die Möglichkeit, mehr „Leben“ in ihre Lehre zu bringen. Studierende werden angeleitet, miteinander zu interagieren und in Teams ihr Wissen anzuwenden. Das könnte das Interesse am Lehrstuhl und die Motivation im Hinblick auf die Inhalte der Lehrveranstatlung erhöhen.

Durch die Modularität bietet Flipped Classroom 2.0 nicht nur ein in sich abgeschlossenes Lehrkonzept, sondern erweitert den Pool an generellen Möglichkeiten zur Gestaltung von Lehrveranstaltungen. So sind Lehrerende nicht gezwungen, das Lehrkonzept Flipped Classroom 2.0 als Ganzes umzusetzen, sondern können, einem Baukastensystem entsprechend, auch lediglich einzelne Bestandteile adaptieren. Z.B. könnte statt einem klassischen Tutorium das Bootcamp angeboten werden. Analog hierzu ist es den Dozierenden möglich, durch die langfristige Reduktion des Lehraufwands durch die Aufzeichnung der Vorlesungsvideos durchgängig ein Lehrangebot zur Verfügung zu stellen. Dadurch ist die Durchführung der Lehrveranstaltung während eines Forschungssemesters genauso möglich wie die dauerhafte Bereitstellung der Lehrinhalte für interessierte Studierende. Ebenfalls ist eine Veröffentlichung der Videos auf Plattformen, die auch für Personenen außerhalb der jeweiligen Universität zugänglich sind, möglich, um so auch externen den Zugang zu den Lehrinhalten zu ermöglichen.

Leider konnte das Lehrkonzept „Flipped Classroom 2.0“ noch nicht vollumfänglich in der Praxis getestet werden. Vor diesem Schritt muss der Aufwand für die Produktion von Uploadvideos in einer didaktisch hochwertigen Form abgeschätzt und überprüft werden, ob alle Voraussetzungen für eine erfolgreiche Durchfühung der Veranstaltung (Budget, Equipment, technisches Wissen) erfüllt sind. Um den Umgang mit der möglichen Weitergabe von Lösungen zu den Use Case besser beurteilen zu können, sollte nach einigen Semestern eine Evaluierung des Lehrkonzepts erfolgen, um dieses ggf. anzupassen. Dies gilt insbesondere für die Erbringung eines Teils der Studienleistung in Gruppen. Denn eine große Ungleichheit der Arbeitsaufteilung in den Gruppen des Semesterprojekts gefährdet nicht nur den langfristigen Lernerfolg der Studierenden, die sich wenig am Semesterprojekt beteiligen, sondern kann auch die Motivation der anderen Gruppenmitglieder zur Teilnahme an der Lehrveranstaltung beeinträchtigen.

Literaturverzeichnis

Abeysekera, L., & Dawson, P. (2015). Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research. *Higher Education Research & Development*, *34*(1), 1–14. https://doi.org/10.1080/07294360.2014.934336

Al-Samarraie, H., Shamsuddin, A., & Alzahrani, A. I. (2020). A flipped classroom model in higher education: a review of the evidence across disciplines. *Educational Technology Research and Development*, *68*(3), 1017–1051. https://doi.org/10.1007/s11423-019-09718-8

Blank, S. (2013). *Why the Lean Start-Up Changes Everything*. https://hbr.org/2013/05/why-the-lean-start-up-changes-everything

Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: a meta-analysis. *Educational Technology Research and Development*, *67*(4), 793–824. https://doi.org/10.1007/s11423-018-9633-7

Christensen, C. M., & Eyring, H. J. E. (2011). *Innovative University: Changing the DNA of Higher Education from the Inside Out*. John Wiley & Sons Inc.

Dehaene, S. (2020). *How We Learn: The New Science of Education and the Brain*. Allen Lane.

Gerbig-Calcagni, I. (2010). *Wie aufmerksam sind Studierende in Vorlesungen und wie viel können sie behalten?* Pädagogische Hochschule Weingarten.

Gillies, R. (2016). Cooperative Learning: Review of Research and Practice. *Australian Journal of Teacher Education*, *41*(3), 39–54. https://doi.org/10.14221/ajte.2016v41n3.3

Hoyer, T., & Mundt, F. (2014). e: t: p: M–ein Blended-Learning-Konzept für Großveranstaltungen. Lernräume gestalten–Bildungskontexte vielfältig denken. In K. Rummler (Hrsg.), *Lernräume gestalten – Bildungskontexte vielfältig denken. Medien in der Wissenschaft* (S. 249–259). Waxmann.

Jensen, J. L., Kummer, T. A., & Godoy, P. D. d. M. (2015). Improvements from a Flipped Classroom May Simply Be the Fruits of Active Learning. *CBE—Life Sciences Education*, *14*(1), ar5. https://doi.org/10.1187/cbe.14-08-0129

Kuh, G. D. (2008). *High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter*. Association of American Colleges and Universities.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the Classroom: A Gateway to Creating an Inclusive Learning Environment. *The Journal of Economic Education*, *31*(1), 30. https://doi.org/10.2307/1183338

Lasch, R., & Schulte, G. (2006). Die Fallstudie als didaktische Methode. In *Quantitative Logistik-Fallstudien* (S. 5–12). Gabler. https://doi.org/10.1007/978-3-8349-9111-9_2

Mayer, R. E. (2004). Should There Be a Three-Strikes Rule Against Pure Discovery Learning? *American Psychologist*, *59*(1), 14–19. https://doi.org/10.1037/0003-066X.59.1.14

Metzger, C., & Schulmeister, R. (2020). Zum Lernverhalten im Bachelorstudium. Zeitbudget-Analysen studentischer Workload im ZEITLast-Projekt. In *Studentischer Workload* (S. 233–251). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-28931-7_9

Prince, M. (2004). Does Active Learning Work? A Review of the Research. *Journal of Engineering Education*, *93*(3), 223–231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x

Ries, E. (2011). *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Currency.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Strelan, P., Osborn, A., & Palmer, E. (2020). Student satisfaction with courses and instructors in a flipped classroom: A meta-analysis. *Journal of Computer Assisted Learning*, *36*(3), 295–314. https://doi.org/10.1111/jcal.12421

Ullmann, E. (2016). *Lernen aus neurobiologischer Perspektive*. https://www.uni-wuerzburg.de/fileadmin/06000060/04_Fort-_und_Weiterbildungen_Lehrkraefte/Herbsttagungen/Herbsttagung_2016/20161006_WS_04_Neurobiologie.pdf

Werner, J., Ebel, C., Spannagel, C., & Bayer, S. (2018). Flipped Classroom – Zeit für deinen Unterricht. In *Flipped Classroom – Zeit für deinen Unterricht* (S. 13–18). Verlag Bertelsmann Stiftung.

Willige, J., Grützmacher, J., Sudheimer, S., & Naumann, H. (2018). *Studienqualitätsmonitor SQM 201*