cii Student Papers 2024

Karlsruhe Institute of Technology
aifb
CRITICAL INFORMATION INFRASTRUCTURES RESEARCH GROUP

cii Student Papers - 2024

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/1000173991

Editorial

Critical information infrastructures (cii) are sociotechnical systems comprising essential software components and information systems with a pivotal impact on individuals, organizations, governments, economies, and society. For more than six years, our research group has been investigating various research- and practice-driven challenges at the Karlsruhe Institute of Technology (KIT) while looking at the design, development, and evaluation of reliable and secure information systems. The main driver for our research is theorizing on and designing the applications and methods required to create and innovate sociotechnical systems with promising value propositions. With this, we are multifaceted in the use contexts, including the Internet and healthcare industries, as well as industry-specific applications of secure and trustworthy artificial intelligence (AI) models. As we focus on human behavior affecting cii and vice versa, our research enables us to rigorously generate strong theoretical insights while producing research outputs relevant to practical audiences.

Each year, our research group supervises nearly 200 course works and theses of bachelor's and master's students during their studies at KIT. To us, research is an essential and inseparable part of university education. The research-based teaching and learning paradigm allows us to incorporate our research topics directly into students' education and learning experiences. It benefits students through stronger engagement, increased learning performance, and increased skills necessary for life-long learning and the effective dissemination of generated insights and knowledge (Blomster et al., 2014; Christe et al., 2015; Healey, 2005; Nuchwana, 2012; Rueß et al., 2016). In doing so, we are highly motivated to ensure that we can provide excellent teaching to students, whereby we apply inquiry-based learning methods and actively introduce our research topics to them in various seminars and lectures. As we think that sound research and working with a team go hand in hand, students primarily work in groups during our courses and deal with problems and issues related to sociotechnical challenges in the realm of cii. The course topics generally correspond to what we are currently researching to ensure high relevance. In addition, we allow students to propose their own research topics or conduct their studies in collaboration with small, medium, or large companies.

Following our cutting-edge information systems research, topics change from semester to semester. Research topics included but are not limited to disruptive health information systems (Thiebes et al., 2023), the secure design of cloud, fog, and edge services (Blume et al., 2023; Brecker, Lins, Trenz, et al., 2023), task-congruence in gamified healthcare information systems (Schmidt-Kraepelin et al., 2024), the evaluation of AI explanations for industry experts (Toussaint, Warsinsky, et al., 2024), designing and implementing requirements for distributed ledgers (Leinweber et al., 2023), adoption and trust concerns regarding the use of AI in autonomous vehicles (Renner et al., 2023), the effects of trust in organizational security practices and protective structures on employees' security-related precaution-taking (Greulich et al., 2024), the emergence and consequences of consumer skepticism toward web seals (Lins et al., 2024), and theory development for transparency of information privacy practices (Dehling & Sunyaev, 2023, 2024). Our team supports students throughout the research process, helping them identify and organize problems, apply appropriate research methods consistently, develop and communicate approaches to solutions, and write research papers.

Involving students in daily work and bringing research to students provides many benefits to the students, our research group, the research community, and practice in general. Students engage with present-day practice problems that research is trying to solve. Moreover, they can apply the theoretical principles and knowledge acquired in previous lectures while working on their seminar papers, deepening their understanding. By offering research-based learning courses, students can gain first-hand experience in self-reliant research and scientific writing and benefit from their now enhanced skillset for writing upcoming seminars, bachelor's, and master's theses. Consequently, we believe our students' works are of high value. Nonetheless, in the past, only few students continued their research after attending a seminar or a lecture, and their works often disappeared into drawers despite the disruptive and valuable insights students have come up with. As a research group, we always appreciate the work of great students and started publishing the best works in a miscellany dedicated to making them available to a broader audience. Our previous collections (see Sunyaev et al., 2021, 2022, 2023) have encouraged students to continue their research in various forms, including volunteer work in their free time, as part of their work as a student assistant in our

research group, and even pursuing a PhD. Students, furthermore, regularly interact with organizations during our courses, which can lead to collaborations and even pave the way for upcoming jobs. Moreover, it encourages great students to incorporate their insights into our research and, sometimes, together with these students, we advance and publish exceptional results in conference proceedings or journals (e.g., Bodynek et al., 2023; Furmanek et al., 2024; Hasse et al., 2024; Hofmann et al., 2024; Hu et al., 2023; Klein et al., 2022).

In the spirit of making students' works available, we continue the idea of publishing the best student works from our courses and are delighted to present this collection for the fourth time in a row. In this work, we bring together the best student works from the previous summer term of 2023 and winter term of 2023/2024. Contributions in this anthology come from four different courses that provide students with a broad range of topics related to cii:

Emerging Trends in Internet Technologies:

The seminar *Emerging Trends in Internet Technologies* aims to provide students with insights into current topics in the field of information systems while focusing mainly on fundamental and innovative Internet technologies. Kicking off with a short introduction and corresponding topics, students are offered a selection of topics around the lectures and present research of our group, including distributed ledger technology (Beyene et al., 2022; Leinweber et al., 2023), cloud, fog, and edge computing (Brecker, Lins, Trenz, et al., 2023), AI (Brecker, Lins, & Sunyaev, 2023), security (Adam et al., 2024; Greulich et al., 2024), and privacy (Dehling & Sunyaev, 2023; Renner et al., 2022). For example, our research group clarified the conceptualization of AI accountability to tackle the lack of conceptual clarity in research and practice (Nguyen et al., 2024).

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. 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), distributed ledger technology (Beyene et al., 2022; Hu et al., 2024), AI (Leiser et al., 2023; Thiebes et al., 2021), digital transformation in the healthcare sector (Guse et al., 2022), and gamification in healthcare (Schmidt-Kraepelin et al., 2023). An example of our interdisciplinary work in this field is a recent systematic mapping study on explainable AI for omics data. The study investigates current machine learning approaches for biomedical data and applied explainable AI methods by systematically analyzing extant literature. In doing so, the study not only provides a research discipline-spanning overview but also identifies open shortcomings of explainable AI for omics data and suggests several future research directions (Toussaint, Leiser, et al., 2024).

Digital Health:

The course *Digital Health* introduces master's students to 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., Guse et al., 2022; Pandl et al., 2021; Rädsch et al., 2021; Thiebes, Schlesner, et al., 2020; Warsinsky et al., 2021). After an introductory session on the challenge of digital transformation in healthcare, the following sessions focus on an in-depth exploration of selected topics 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 in the field of digital health is mental health, which is becoming increasingly important for both individuals and society (Hu et al., 2023). Given individual differences and the sensitivity of specific mental health issues, we focus on integrating different game design elements into mental health information systems, paying attention to users' specific preferences while also exploring the potential roles and effects of individual game design elements, such as avatars, in digital mental health interventions. By studying relevant literature and existing applications, our research aims to provide up-to-date insights into the meaningful use of gamified information systems in mental health (Hu et al., 2023).

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's students learn to handle the complexities of designing, developing, operating, and evaluating cii. At the beginning of the course, cii are introduced on a general level. 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 discussing the characteristics and challenges of cii (Dehling et al., 2019).

Selected Issues in Critical Information Infrastructures / Trustworthy Emerging Technologies:

In the summer term of 2023, the course *Selected Issues in cii* focused on examining emerging technologies and their impact on society and the economy. An emerging technology is a radically novel and relatively fast-growing technology characterized by a certain degree of coherence persisting over time and with the potential to exert a considerable impact on the socio-economic domain(s), which is observed in terms of the composition of actors, institutions, and patterns of interactions among those, along with the associated knowledge production processes (Rotolo et al., 2015). However, its most prominent impact lies in the future, so its emergence phase is still uncertain and ambiguous. The cii research group aims to examine emerging technologies to be on the cutting edge of research and invites students to join the endeavor, offering them an insightful and engaging learning experience. During the course, students collaboratively analyzed technologies like the Metaverse, human digital twins, and sustainable data centers while being supported by researchers from our research group. Students had the opportunity to contribute their ideas and conceptions. As a novel learning technique, the students had to develop a learning portfolio throughout the course, consisting of a topic concept, an interim video presentation, peer feedback, and a seminar paper. Given the considerable success of the course, the research group offered this course again for the winter term 2023/2024 under the new course name *Trustworthy Emerging Technologies*.

We selected the student works representing excellent and intriguing studies from these courses. The student works in this book cover a wide range of research problems, including summaries of the current state of explainable AI evaluation methods in healthcare, the integration of domain knowledge into medical image classification tasks, and the value creation and orchestration in digital platform-based ecosystems, as well as interview studies on human digital twins, smart home users' behavior in response to privacy implications, and the generative AI landscape.

- Butt, Nocus, Rose, and Tiemann conducted a literature review employing thematic analysis to provide a comprehensive overview of the current state of explainable AI evaluation methods in healthcare. They identified eight key themes and laid the foundation for developing context-specific evaluation frameworks for explainable AI in healthcare.
- Xanthakis and Dang explored how integrating domain knowledge into medical image classification tasks within informed machine-learning approaches can improve effectiveness and accuracy. Systematically reviewing 80 publications, they unraveled domain knowledge characteristics and investigated the interplay of data and knowledge in data preprocessing. They concluded that medical image analysis systems can be advanced by focusing on knowledge and data preprocessing.
- Braun, Häusle, Romanenko, and Schorling performed a systematic literature review to investigate digital platform-based ecosystems regarding their orchestration, technical aspects, and the value creation within such ecosystems. Their study emphasizes the importance of a shared vision among ecosystem participants, a clear distribution of roles, and modular resources in creating a thriving ecosystem.
- Fantino, Fischer, Günder, and Zekri conduct a systematic literature review to investigate the social and ethical implications of the metaverse on the younger generation. They identify the advantages and risks associated with the metaverse and provide a basis for developing coping measures to

mitigate them and leverage the advantages identified for a safe, inclusive, and transparent future digital world.

- Xanthakis, Muff, Timpe, and Weeber conducted expert interviews to investigate human digital twins' current and future state in healthcare. Employing inductive and deductive coding strategies, they revealed opportunities, benefits, barriers, and dangers of human digital twins in healthcare in technical, legal, and ethical dimensions.
- Fantino, Faßbender, Günder, and Steinecke conducted an interview study and employed thematic analysis to assess smart home user behavior in response to privacy implications. Their findings highlight that privacy expectations may differ depending on the user's social environment and context and should be considered when developing smart homes and respective regulations.
- König, Faber, Xie, and Loder interviewed industry experts to analyze the generative AI market landscape. Their study concludes that the generative AI market is competitive and dynamic, presenting high entry barriers for large language model foundation providers and a high risk of dependency on those for the large language model layer providers. They emphasize the importance of regulatory caution to achieve innovative and fair competition.

Concluding this brief overview, we are grateful that these students have taken the time to revise and improve their work to ensure the high quality of this miscellany. In addition to the students who wrote the articles, this book was only possible with the help of the dedicated researchers in our research group who mentored the students throughout the courses. We want 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 excellence in teaching and intend to publish the best papers in a compendium each year to bring students closer to scientific work.

Sincerely,

Ali Sunyaev, Guangyu Du, Maximilian Renner, Philipp A. Toussaint, Scott Thiebes, Sebastian Lins, Yannick Erb

Miscellany Team 2024

Supervising Research Associates

References

Adam, M., Lins, S., Sunyaev, A., & Benlian, A. (2024). The Contingent Effects of IS Certifications on a Website's Trustworthiness. *Journal of the Association for Information Systems*, *25*(3), 594–617. **https://doi.org/10.17705/1jais.00836**

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**

Blomster, J., Venn, S., & Virtanen, V. (2014). Towards Developing a Common Conception of Research-Based Teaching and Learning in an Academic Community. *Higher Education Studies*, *4*(4), 62–75. **https://doi.org/10.5539/hes.v4n4p62**

Blume, M., Lins, S., & Sunyaev, A. (2023). Uncovering Effective Roles and Tasks for Fog Systems. In G. A. Papadopoulos, F. Rademacher, & J. Soldani (Eds.), *Service-Oriented and Cloud Computing*. *ESOCC 2023*. *Lecture Notes in Computer Science* (Vol. 14183, pp. 119–135). Springer. **https://doi.org/10.1007/978-3-031-46235-1_8**

Bodynek, M., Leiser, F., Thiebes, S., & Sunyaev, A. (2023). Applying Random Forests in Federated Learning: A Synthesis of Aggregation Techniques. *Wirtschaftsinformatik 2023 Proceedings*, Article 46. 18th International Conference on Wirtschaftsinformatik (WI2023), Paderborn, Germany. **https://aisel.aisnet.org/wi2023/46**

Brecker, K., Lins, S., & Sunyaev, A. (2023). Why it Remains Challenging to Assess Artificial Intelligence. *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS)*, 5242–5251. **https://hdl.handle.net/10125/103275**

Brecker, K., Lins, S., Trenz, M., & Sunyaev, A. (2023). Artificial Intelligence as a Service: Trade-Offs Impacting Service Design and Selection. *Proceedings of the 44th International Conference on Information Systems (ICIS)*.

Christe, D., Shah, A., Bhatt, J., Powell, L., & Kontsos, A. (2015). Raising interest in STEM education: A research-based learning framework. *2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services*, 167–169. **https://doi.org/10.1109/ETTLIS.2015.7048192**

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

Dehling, T., & Sunyaev, A. (2023). A Design Theory for Transparency of Information Privacy Practices. *Information Systems Research*, *Articles in Advance*, 1–22. **https://doi.org/10.1287/isre.2019.0239**

Dehling, T., & Sunyaev, A. (2024). *A design theory for transparency of information privacy practices : [Appendix]*. **https://doi.org/10.5445/IR/1000170914**

Furmanek, L., Lins, S., Blume, M., & Sunyaev, A. (2024). Developing a Hybrid Deployment Model for Highly Available Manufacturing Execution Systems. *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 2095–2100. **https://doi.org/10.1109/MIPRO60963.2024.10569530**

Greulich, M., Lins, S., Pienta, D., Thatcher, J. B., & Sunyaev, A. (2024). Exploring Contrasting Effects of Trust in Organizational Security Practices and Protective Structures on Employees' Security-Related Precaution Taking. *Information Systems Research*, *Articles in Advance*, 1–23. **https://doi.org/10.1287/isre.2021.0528**

Guse, R., Thiebes, S., Hennel, P., Rosenkranz, C., & Sunyaev, A. (2022). How Do Employees Perceive Digital Transformation and its Effects? A Theory of the Smart Machine Perspective. *ICIS 2022 Proceedings*. International Conference on Information Systems (ICIS) 2022, Copenhagen, Denmark. **https://aisel.aisnet.org/icis2022/digit_nxt_gen/digit_nxt_gen/6**

Hasse, F., Leiser, F., & Sunyaev, A. (2024). Informed machine learning for cardiomegaly detection in chest X-rays: a comparative study. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. **https://doi.org/10.1109/ISBI56570.2024.10635719**

Healey, M. (2005). Linking research and teaching: exploring disciplinary spaces and the role of inquiry-based learning. In R. Barnett (Ed.), *Reshaping the University: New Relationships between Research, Scholarship and Teaching* (pp. 67–78). McGraw Hill / Open University Press.

Hofmann, P., Brand, A., Späthe, E., Lins, S., & Sunyaev, A. (2024, September 1). AI-based Tools in Higher Education – A Comparative Analysis of University Guidelines. *Proceedings of Mensch Und Computer 2024 (MuC '24)*. Mensch und Computer 2024 (MuC '24), Karlsruhe, Germany. **https://doi.org/10.1145/3670653.3677513**

Hu, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2024). Mapping Distributed Ledger Technology Characteristics to Use Cases in Healthcare: A Structured Literature Review. *ACM Transactions on Computing for Healthcare*, *5*(3), Article 15. **https://doi.org/10.1145/3653076**

Hu, S., Usta, A., Schmidt-Kraepelin, M., Warsinsky, S., Thiebes, S., & Sunyaev, A. (2023). *Be Mindful of User Preferences: An Explorative Study on Game Design Elements in Mindfulness Applications*. Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, USA.

Klein, T., Fenn, T., Katzenbach, A., Teigeler, H., Lins, S., & Sunyaev, A. (2022). A Threat Model for Vehicular Fog Computing. *IEEE Access*, *10*, 133256–133278. **https://doi.org/10.1109/access.2022.3231189**

Leinweber, M., Kannengießer, N., Hartenstein, H., & Sunyaev, A. (2023). Leveraging Distributed Ledger Technology for Decentralized Mobility-as-a-Service Ticket Systems. In H. Proff (Ed.), *Towards the New Normal in Mobility* (pp. 547–567). Springer Fachmedien Wiesbaden. **https://doi.org/10.1007/978-3-658-39438-7_32**

Leiser, F., Rank, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2023). Medical informed machine learning: A scoping review and future research directions. *Artificial Intelligence in Medicine*, *145*, Article 102676. **https://www.sciencedirect.com/science/article/pii/S0933365723001902**

Lins, S., Greulich, M., Löbbers, J., Benlian, A., & Sunyaev, A. (2024). Why So Skeptical? Investigating the Emergence and Consequences of Consumer Skepticism toward Web Seals. *Information & Management*, *61*(2), Article 103920. **https://doi.org/10.1016/j.im.2024.103920**

Nguyen, L. H., Lins, S., Renner, M., & Sunyaev, A. (2024, June 13). Unraveling the Nuances of AI Accountability: A Synthesis of Dimensions Across Disciplines. *ECIS 2024 Proceedings*, Article 15. ECIS 2024, Paphos, Cyprus. **https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/15**

Nuchwana, L. (2012). How to Link Teaching and Research to Enhance Students' Learning Outcomes: Thai University Experience. *Procedia - Social and Behavioral Sciences*, *69*, 213–219. **https://doi.org/10.1016/j.sbspro.2012.11.401**

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), Article 9414. **https://doi.org/10.1038/s41598-021-88768-6**

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)*, 1294–1303. **https://hdl.handle.net/10125/70769**

Renner, M., Lins, S., Söllner, M., Jarvenpaa, S., & Sunyaev, A. (2023). Artificial Intelligence-Driven Convergence and its Moderating Effect on Multi-Source Trust Transfer. *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS)*, 5208–5217. **https://hdl.handle.net/10125/103271**

Renner, M., Lins, S., Söllner, M., Thiebes, S., & Sunyaev, A. (2022). Understanding the Necessary Conditions of Multi-Source Trust Transfer in Artificial Intelligence. *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*, 5901–5910. **http://hdl.handle.net/10125/80057**

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827–1843. **https://doi.org/10.1016/j.respol.2015.06.006**

Rueß, J., Gess, C., & Deicke, W. (2016). Forschendes Lernen und forschungsbezogene Lehre - empirisch gestützte Systematisierung des Forschungsbezugs hochschulischer Lehre. *Zeitschrift Für Hochschulentwicklung*, *11*(2), 23–44. **https://doi.org/10.3217/ZFHE-11-02/02**

Schmidt-Kraepelin, M., Ben Ayed, M., Warsinsky, S., Hu, S., Thiebes, S., & Sunyaev, A. (2024). Leaderboards in Gamified Information Systems for Health Behavior Change: The Role of Positioning, Psychological Needs, and Gamification User Types. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*, 3444–3453. **https://hdl.handle.net/10125/106800**

Schmidt-Kraepelin, M., Thiebes, S., Warsinsky, S. L., Petter, S., & Sunyaev, A. (2023, April 19). Narrative Transportation in Gamified Information Systems: The Role of Narrative-Task Congruence. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 215. 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany. **https://doi.org/10.1145/3544549.3585595**

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

Sunyaev, A., Renner, M., Toussaint, P. A., Thiebes, S., & Lins, S. (Eds.). (2022). *cii Student Papers - 2022*. Karlsruher Institut für Technologie (KIT). **https://doi.org/10.5445/IR/1000150078**

Sunyaev, A., Renner, M., Toussaint, P. A., Thiebes, S., & Lins, S. (Eds.). (2023). *cii Student Papers - 2023*. Karlsruher Institut für Technologie (KIT). **https://doi.org/10.5445/IR/1000162178**

Thiebes, S., Gao, F., Briggs, R. O., Schmidt-Kraepelin, M., & Sunyaev, A. (2023). Design Concerns for Multiorganizational, Multistakeholder Collaboration: A Study in the Healthcare Industry. *Journal of Management Information Systems*, *40*(1), 239–270. **https://doi.org/10.1080/07421222.2023.2172771**

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy Artificial Intelligence. *Electronic Markets*, *31*(2), 447–464. **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. *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), Article e14890. **https://doi.org/10.2196/14890**

Toussaint, P. A., Leiser, F., Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2024). Explainable artificial intelligence for omics data: a systematic mapping study. *Briefings in Bioinformatics*, *25*(1), Article bbad453. **https://doi.org/10.1093/bib/bbad453**

Toussaint, P. A., Warsinsky, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2024). Designing Gamification Concepts for Expert Explainable Artificial Intelligence Evaluation Tasks: A Problem Space Exploration. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*, 1338–1347. **https://hdl.handle.net/10125/106542**

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 (HICSS)*, 1305–1314. **http://hdl.handle.net/10125/70771**

Table of Contents

Explainable Artificial Intelligence Evaluation for Healthcare: A Literature Review

Emerging Trends in Digital Health, Summer Term 2023

Waleed Butt
Bachelor Student
Karlsruhe Institute of Technology
waleed.butt@student.kit.edu

Yannick Nocus
Bachelor Student
Karlsruhe Institute of Technology
yannick.nocus@student.kit.edu

Hannah Rose
Bachelor Student
Karlsruhe Institute of Technology
hannah.rose@student.kit.edu

Erja Tiemann
Bachelor Student
Karlsruhe Institute of Technology
erja.tiemann@student.kit.edu

Abstract

Background: *Explainable Artificial Intelligence (XAI) is increasingly important in healthcare, where transparency and trust are crucial. Traditional AI's "black box" nature often interferes with acceptance among medical professionals and patients, despite its potential to improve diagnostic and treatment processes. XAI approaches proposed in the literature have not yet been evaluated in practice by health professionals due to lack of time and conviction. Without XAI, doctors would have to resort to manual and traditional data processing methods, which are less precise and time-consuming. The consequences can be delayed and incorrect diagnoses, overloading of staff, and avoidable serious illnesses.*

Objective: *XAI methods depend on their context, especially their target group. At present, the research community lacks appropriate and mature evaluation approaches that convincingly assess XAI. This literature review aims to provide a comprehensive overview of existing evaluation methods in healthcare and to identify key issues and challenges in this interdisciplinary field.*

Methods: *Using Braun and Clarke's thematic analysis, literature from Scopus was systematically reviewed, focusing on peer-reviewed journals. As relevant identified literature was coded, and themes were derived to understand the current state of XAI evaluation methods in healthcare.*

Results: *Using a collection of relevant literature and a selection of codes, eight key themes were identified: XAI, experiment circumstances, stakeholders, evaluation criteria, evaluation methods, evaluation metrics, ethical and legal implications, and challenges.*

Conclusion: *This work underscores the necessity for standardized approaches and definitions in XAI evaluation to enhance comparability and reliability. Future research should focus on interdisciplinary collaboration and developing comprehensive, context-specific evaluation frameworks to advance XAI implementation in healthcare.*

Keywords: Explainable Artificial Intelligence, XAI, Healthcare, evaluation methods, thematic analysis, literature review, trust, transparency

Introduction

Background and Scope

Artificial Intelligence (AI) has proven to be a powerful tool for a multitude of applications and thus gained a vast amount of popularity from the public and many industries (Erdeniz et al., 2022). As it already performs more efficiently and effectively at certain tasks like image analysis (Muddamsetty et al., 2021) and classification (Slijepcevic et al., 2022) than humans, AI shows promising potential for the appliance in medicine.

While AI is already seeing use in healthcare (Maddox et al., 2019; Muddamsetty et al., 2021), extensive implementation is hindered by the lack of acceptance. Reasons for this deficit include the black-box character of conventional AIs, referring to the shortfall of interpretable insight into the decision-making process (Muddamsetty et al., 2021). Since healthcare is a high-stakes field in which errors can be fatal, medical professionals have shown skepticism toward the use of AI for diagnosis and treatment (Muddamsetty et al., 2021). Nevertheless, AI is believed to potentially improve healthcare workflow efficiency, if transparency and interpretability are achieved (Yu & Wu, 2023). Furthermore, as of 2018, the European Union's General Data Protection Regulation (GDPR) law requires all algorithms used for patient care to be transparent by explaining the decision-making process (Salahuddin et al., 2021).

A promising solution for the previously mentioned difficulties could be explainable artificial intelligence (XAI). XAI technologies provide interpretable information (Li et al., 2020; Naiseh et al., 2023) about the decision of AI (Muddamsetty et al., 2021) and contribute to transparency and trust in its implementation (Korica et al., 2021). Research on XAI is still in its infancy, and the quality of an explanation depends on the context and the target group, respectively (Salahuddin et al., 2021). Therefore, XAI methods intended for the medical appliance must be evaluated with the involvement of medical experts and with the use of the right evaluation methods (Kaur et al., 2021). However, there is no uniform evaluation approach and the appliance of XAI is also subject to legal and ethical questions, which pose additional obstacles (Jung et al., 2023).

Objectives

Given the interdisciplinary nature of this topic, a holistic overview of existing XAI evaluation approaches mentioned in the literature could support the identification and development of suitable evaluation methods for XAI used in healthcare. In this work, a thematic analysis is applied to give a systematic overview of the literature on XAI evaluation approaches in healthcare.

To give a systemic overview, three consecutive subgoals (S) are perused:

S1 - Identification of the relevant literature

S2 - Analyzation and coding of the identified literature

S3 - Derivation of the overarching themes from the codes

Structure

This review is structured into six sections. The second section describes the fundamental definitions and assumptions used in our research. In the third section, we will outline the systematic steps of our analysis and render our findings in the fourth section. This will contain a collection of relevant literature, as well as the underlying codes and a selection of the identified themes. We will discuss our results in the fifth section including the implications and limitations, as well as possible future research. Lastly, we conclude our work in the sixth section.

Core Concepts

This section provides an in-depth explanation of the theoretical foundations and key concepts that are crucial for understanding the research question and our results. The initial part of this background chapter will focus on defining and explaining the key terms that will appear frequently in this paper.

Artificial Intelligence in Healthcare

The healthcare domain provides multiple areas of application for AI. It has already been shown to be a helpful tool for the analysis of immense amounts of clinical data (Jung et al., 2023). Also, algorithms have prominently been used as a support system for decision-making processes (Erdeniz et al., 2022; Slijepcevic et al., 2022) for example by integrating real-world evidence and thereby balancing human expert experience with evidence (Wysocki et al., 2023).

Despite the huge service AI provides, the nature of medicine also implies that there are risks. The primary concern of AI-supported decision-making processes is the safety of the patients (Jung et al., 2023). The health and lives of humans can be severely impaired by decisions, and therefore AI must be trusted to make or support these decisions (Siddiqui & Doyle, 2022). This makes its development and uses come with great responsibility (Kaur et al., 2022), and it's not clear who bears it. As an attempt to prevent harm, the EU has passed multiple guidelines, including the previously mentioned GDPR. These guidelines are meant to govern AI systems on requirements like explainability, trustworthiness, and safety (Kaur et al., 2022).

However, many AI models struggle with providing interpretable insights into the decision-making process, hence their claimed black-box nature. Nevertheless, unraveling this nature has become a legal and ethical requirement, as well as a necessity to foster clinical trust (Salahuddin et al., 2021).

Explainable Artificial Intelligence

As a solution to the problems and limitations associated with AI, the topic of XAI gained significant attention. The fundamental idea behind this concept is to produce performant and explainable models, which allow humans to understand, manage and appropriately trust AI (Barredo Arrieta et al., 2020). An increase in scientific publications referring to or containing XAI started to appear in the database Scopus in 2017, with growing numbers in the following years (comp. Figure 1) (Barredo Arrieta et al., 2020). This can serve as an indicator of the rising interest in the topic.

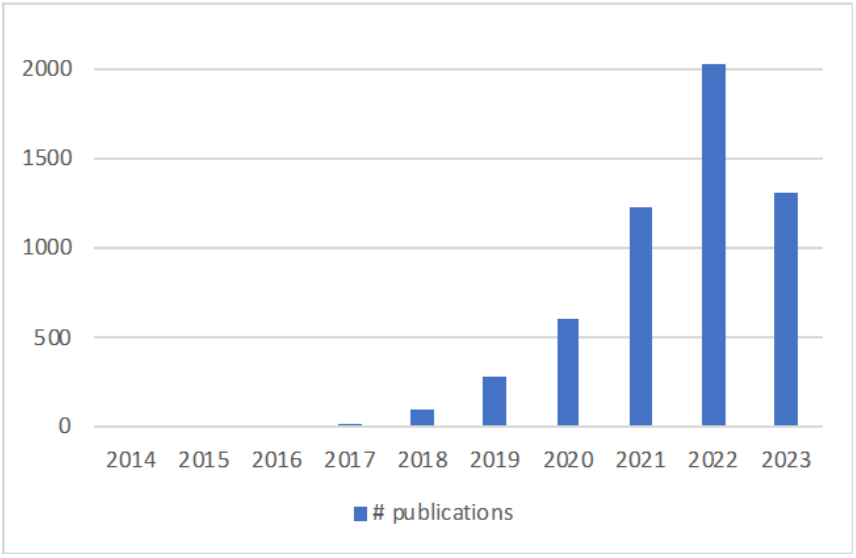

Figure 1: Number of Publications Regarding XAI

Pinpointing XAI to one definition has revealed itself to be challenging. This problem started with the non-uniform definition and use of the words explainability and interpretability. As we try to report on evaluation methods for explainable AI, we concluded to differentiate between explainable AI and interpretable AI. We define XAI as systems that, given a target group, provide information and reason to make its functioning clear and understandable (Barredo Arrieta et al., 2020). The main goal in the field of study is to create a process of interpretable models and methods that render more explainable models while maintaining performance levels (Carvalho et al., 2019). Furthermore, we restrict this definition to post-hoc models, referring to systems that intend to explain the learned patterns of predictive models already trained with data (Yu & Wu, 2023). The counterpart, model-based explainable AI, is more commonly associated with interpretability and is not the focus of this review.

Explainable Artificial Intelligence in Healthcare

In medicine, there is a growing need for AI approaches that not only work well, but are also trustworthy, transparent, interpretable, and explainable to a human expert (Holzinger et al., 2019). These requirements could be met by XAI. However, the results must be also understandable for laypersons. After all, if neither the patient nor the physician trusts the AI-based clinical diagnosis, its use becomes unlikely (Lötsch et al., 2021). Patients, healthcare professionals, and researchers cannot be expected to have the same level of knowledge and have different requirements for explainability and XAI. This further emphasizes the role of the audience and the subjective nature of explainability.

Explainable Artificial Intelligence Evaluation Methods

To properly assess the quality and usability of XAI models, it is necessary to have appropriate evaluation methods. These methods should focus on evaluating the explainability that XAI provides. However, there is no single definition of explainability but depends on the context (Carvalho et al., 2019). While no standard has yet been established, certain characteristics have been identified to majorly influence the quality of XAI methods. These criteria contain among others trust assessment (Jung et al., 2023) and causability (Müller et al., 2022). As these are important fundaments for the evaluation of XAI, they need to be further defined. While high trust in an XAI method can be interpreted as a positive aspect, it can also negatively influence its assessment. Excessive trust can lead to bias (Jung et al., 2023), which could, in turn, promote overreliance. This must be considered in trust assessment. Causability contains three components, namely effectiveness, efficiency, and satisfaction of the user (Müller et al., 2022). User satisfaction measures how well the output provided by XAI corresponds with the user's expectations and directly influences the perceived pragmatic utility (Wysocki et al., 2023).

Metrics and Criteria

Metrics and criteria are both important for the evaluation of XAI methods and the quality of the explanation they provide. However, like explainability and interpretability, these words are also frequently used as synonyms. We decided to differentiate between criteria and metrics, as it allows us to provide a finer insight into the state of the analyzed literature. We define criteria as characteristics of explanations. Those characteristics are closely linked to user demands. As an example, traceability (Müller et al., 2022) is one of the criteria we found. While this is an important factor that needs to be tested when evaluating an XAI method, it is also vague. Criteria, therefore, are more general goals and ambitions. Metrics, on the other hand, are more specific evaluation approaches, which aim to qualify and or quantify the output of XAI models and are to be measured by the evaluation methods. An example of a metric is the user agreement rate, as it can be measured accurately (Naiseh et al., 2023).

Materials and Methods

In this study, we utilized the Braun and Clarke thematic analysis method (Braun & Clarke, 2006) to gain insights into our research topic as it supported us uncover the underlying themes, patterns, and trends in the literature related to the evaluation of XAI in healthcare settings. It allowed us to identify recurring issues, challenges, and successes in the assessment of XAI systems, contributing to a comprehensive understanding of the topic. This method consists of six phases: (1) familiarization with the data, (2) generation of initial codes, (3) theme search, (4) theme review, (5) theme refinement, and (6) reporting.

Identifying Relevant Studies

In the first phase of Braun and Clarke's approach (Braun & Clarke, 2006), we identified the databases most suitable for our project. We investigated four databases namely EBSCOhost, IEEE Xplore, Scopus, and Science Direct. Out of these, we chose Scopus as our only database to look through, since it is an interdisciplinary database that covers the technological and medical field. In order not to miss relevant papers we used two different search strings (see Table 1 below) to get a broader coverage of the field. Although the term XAI is new, the concept of explainability is not. Therefore, we didn't exclude any studies by the year they were published. We only included studies published in peer-reviewed journals and conference proceedings. Only papers written either in English or German were included.

	Search String 1	**Search String 2**
Search String	TITLE-ABS-KEY ((((explainable OR interpretable) AND (ai OR "artificial intelligence" OR "machine learning" OR "deep learning")) OR xai) AND (health* OR medic* OR clinic*)) AND TITLE(evalu*)	TITLE-ABS-KEY (((explainable OR interpretable) AND (ai OR "artificial intelligence" OR "machine learning" OR "deep learning")) OR xai AND (evalu* OR performance)) and TITLE(health* OR medic* OR clinic*)
Fields	Title; abstract; keywords	Title; abstract; keywords
Databases	Scopus	Scopus
Publication Types	Journal articles; conference papers	Journal articles; conference papers
Date Range	Peer-reviewed publications: January 1999 to May 2023	Peer-reviewed publications: January 1999 to May 2023

Table 1: Search String

Screening and Coding

Screening of the title, abstract, and full-text article corresponded to the second phase of Braun and Clarke's approach (Braun & Clarke, 2006) and was performed independently in pairs by two reviewers, and discrepancies were resolved by reaching a consensus between the four reviewers. After screening the relevant literature, the coding process refers to the process of extracting data chunks that appear interesting. The given code represents the chunk of data in the literature (Braun & Clarke, 2006). Initially, we decided that each member should code 10 different papers but after coding the same paper with varying results, we decided we should split into pairs again. We double-coded in 2 pairs and each pair analyzed 10 papers. Each reviewer collected their identified codes numbered in a table. Exemplary codes are 'expert feedback' (Muddamsetty et al., 2021) or 'variable importance evaluation' (Yu & Wu, 2023).

Derivation of Themes

After coding according to our last chapter, our team thoroughly reviewed all the generated codes and made necessary improvements or removals. In this chapter, we followed phases (3), (4), and (5) of the Braun and Clark approach (Braun & Clarke, 2006). We identified similarities among the codes and grouped them into clusters by creating a mind map. Each code was ranked based on its frequency of occurrence. Lastly, we explored potential themes that seemed feasible.

Results

This chapter presents the results we obtained through the chapters in 'Materials and Methods' and is part of the reporting phase (6) of the Braun and Clarke approach (Braun & Clarke, 2006). The following chapter 'Collection of Literature' is the result of identifying relevant studies. The chapter 'Codes' presents the codes we collected as a result of our screening and coding process. The last chapter 'Themes' will cover the themes we derived in the previous chapter.

Collection of Literature

The collection of Relevant literature is *S1*. The literature selection process is depicted in Figure 2. With the 2 search strings, we identified a total of 594 papers and after removing 105 duplicates 489 remained. For the title and abstract screening, we split our team into 2 pairs where each pair screened half (244 and 245) of the papers. For a paper to be eligible for further analysis it needed to fulfill one of the following criteria: The paper either evaluated the explanation effectiveness, defined a metric to compare the explanation of an XAI method, compared explanations of XAI methods, or executed user studies with the focus on explainability. If the pair couldn't come to a consensus, then all 4 members discussed the paper. After the title and abstract screening 444 papers in total were excluded either because the papers were deemed off-topic or they evaluated something different. Due to time constraints, we didn't conduct a backward search of the screened papers. After conducting a full-text screening of 30 articles we removed further 10 articles and coded the remaining 20 articles.

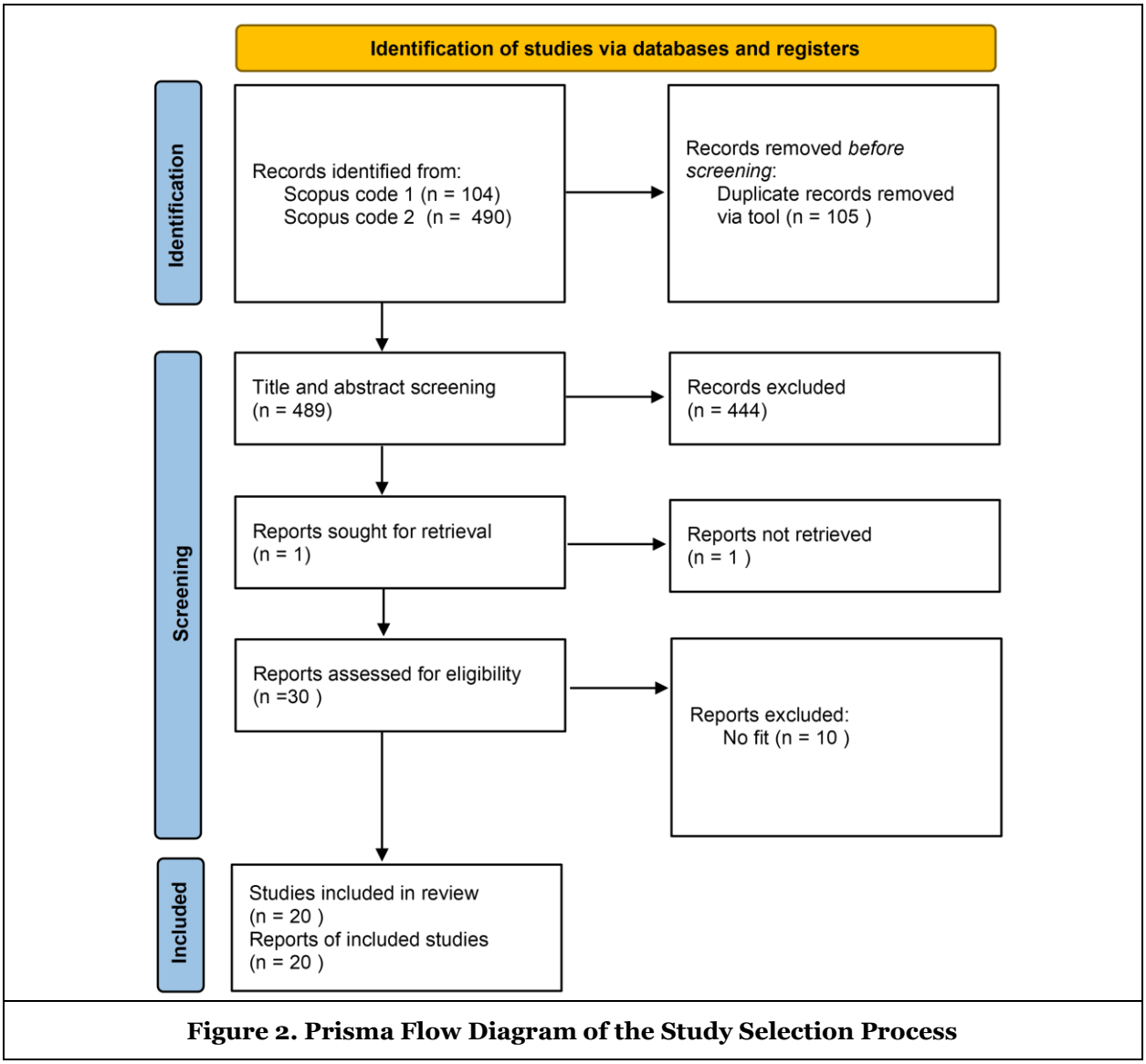

Figure 2. Prisma Flow Diagram of the Study Selection Process

Codes

During the process of double coding, we created a total amount of 515 codes. 'LIME' (short for 'Local Interpretable Model-Agnostic' (Ribeiro et al., 2016)) with an appearance of 10 has the highest count. Most of the codes are very specific and describe the used XAI method or applied mathematical and statistical models. Others are qualitative like 'trust', 'bias', or 'time'. An overview of the extracted codes from the analyzed literature can be found in Figure A-1 in the appendix. The chapter Codes achieves *S2*.

Themes

Through the mind map design process, we have identified eight themes. These themes consist of Explainable *Artificial Intelligence, Experiment Circumstances, Explainable Artificial Intelligence Stakeholders, Evaluation Criteria, Evaluation Methods, Evaluation Metrics, and Ethical, and Legal Implications*. A summary of all the themes can be found in Table 5. The themes provide a comprehensive understanding of the current research of XAI-Evaluation approaches in the healthcare system. This fulfills *S3*.

Explainable Artificial Intelligence

We identified XAI as one of the major themes that serves as the foundation of our research issue. Under this theme, we gather information on the complexity of XAI. We listed all the models that were used and categorized the type of explanation that was utilized, and the method of presentation of the explanation.

Models. We found 19 different XAI-Models that were used in the evaluation experiments. Especially LIME (Hammoud et al., 2022), SHAP (Korica et al., 2021), and GradCam (Rao et al., 2022) or GradCam++ (Siddiqui & Doyle, 2022) respectively were used often. In Table 2 all used models are listed.

Explanation Class. Explanation classes categorize the kind of explanation given by an XAI model. Often these classes are divided into local and global explanations. Another frequently named class is counterfactual explanations, which according to Naiseh et al. (2023) is distinct from other XAI classes, while Dieber & Kirrane (2022) argue it is part of local explanations.

Local explainability is a subject-level explanation, which explains the model for a specific individual sample (Yu & Wu, 2023). There are multiple ways to achieve local explainability. In the study of Naiseh et al. (2023), local explanation is introduced to be achievable by quantifying the contribution value for each input data feature to the recommendation or generating local rules or decision trees of a recommendation.

On the other hand, global explanations aim to explain the overall logic of a black-box model used in AI-based recommendations. This includes presenting the weights of different data features as decision trees, rules, or ranking styles (Naiseh et al., 2023). Erdeniz et al. (2022) define global explanations as an output based on the training data and not personalized for each prediction specifically. An example of a global explanation method is the Pearson Correlation Coefficient.

Counterfactual explanations are introduced in Naiseh et al. (2023) as the answer to the user's questions "what-if" to observe the effect of a modified data feature on the recommendation.

Explanation Presentation Style. After analyzing the nature of explanations, we looked at how an explanation is shown to the user. We discovered maps to be a frequently used tool to visually present information to the user. Information mostly concerns image analysis. Additionally, interfaces and language descriptions were key findings. For instance, heat maps or saliency maps can be used to highlight relevant regions regarding a prediction (Salahuddin et al., 2021). Highlighted regions visualize the relative weights (or importance) of features in the data before feeding them into deep learning mechanisms for automated diagnosis (Wang et al., 2023). El Shawi et al. (2019) concluded that their tabular user interface helped participants detect bias more effectively than the aggregated user interface. Research has shown that providing clinicians with language descriptions in a human-readable format can enhance their understanding and interpretation, as highlighted by Slijepcevic et al. (2022).

Explanation Presentation Style. After analyzing the nature of explanations, we looked at how an explanation is shown to the user. We discovered maps to be a frequently used tool to visually present information to the user. Information mostly concerns image analysis. Additionally, interfaces and language

descriptions were key findings. For instance, heat maps or saliency maps can be used to highlight relevant regions regarding a prediction (Salahuddin et al., 2021). Highlighted regions visualize the relative weights (or importance) of features in the data before feeding them into deep learning mechanisms for automated diagnosis (Wang et al., 2023). El Shawi et al. (2019) concluded that their tabular user interface helped participants detect bias more effectively than the aggregated user interface. Research has shown that providing clinicians with language descriptions in a human-readable format can enhance their understanding and interpretation, as highlighted by Slijepcevic et al. (2022).

XAI-Models	Exemplary references
LIME	(El Shawi et al., 2019; ElShawi et al., 2021; Erdeniz et al., 2022; Hakkoum et al., 2022; Jung et al., 2023; Kaur et al., 2022; Korica et al., 2021)
Smooth Grad	(Rao et al., 2022; Siddiqui & Doyle, 2022)
SHAP	(El Shawi et al., 2019; ElShawi et al., 2021; Korica et al., 2021; Li et al., 2020; Slijepcevic et al., 2022)
GradCam	(Muddamsetty et al., 2021; Rao et al., 2022; Siddiqui & Doyle, 2022)
Deep Taylor	(Singh et al., 2021)
GradInput	(Rao et al., 2022)
GraphNET	(Rao et al., 2022)
GraphSAGE	(Rao et al., 2022)
CAM	(Rao et al., 2022)
IG	(Rao et al., 2022)
MAPLE	(ElShawi et al., 2021)
SIDU	(Muddamsetty et al., 2021)
GradCAM++	(Korica et al., 2021; Siddiqui & Doyle, 2022)
Anchors	(El Shawi et al., 2019; ElShawi et al., 2021)
Hypothetical AI	(Naiseh et al., 2023)
Layer-wise Relevance Propagation (LRP)	(Slijepcevic et al., 2022)
LORE	(El Shawi et al., 2019)
Manifold	(Li et al., 2020)
PGExplainer	(Rao et al., 2022)

Table 2: Overview of all Found XAI Models

Experimental Circumstances

Because the healthcare industry is vast and encompasses numerous specializations and variations, we compiled information from various medical fields and data sources cited in the analyzed papers.

Medical Field. Our code overview includes all the medical fields we have identified (comp. Table 3). Medical fields include therapies and diagnosis. The medical field varies from diagnosing retinal conditions (Singh et al., 2021) to predicting in-hospital mortality (Li et al., 2020).

Data. Mostly data was used in regard to training and testing an XAI model. In the study by (Wysocki et al., 2023), the XAI model was intentionally exposed to error-sensitive data to effectively gauge the overreliance of the expert user.

Medical Field	Exemplary references
CDSS	(Erdeniz et al., 2022; Wysocki et al., 2023)
Cancer	(Hakkoum et al., 2022)
Chemotherapy	(Naiseh et al., 2023)
Covid	(Wysocki et al., 2023)
Clinical Gait Analysis	(Slijepcevic et al., 2022)
Mortality	(El Shawi et al., 2019)
Retinal Diseases	(Singh et al., 2021)
Companion Diagnostics	(Müller et al., 2022)
Subarachnoid Hemorrhage	(Yu & Wu, 2023)

Table 3: Table of Medical Fields That Were Mentioned in Analyzed Papers

Explainable Artificial Intelligence Stakeholder

The Stakeholders involved in XAI extend beyond just healthcare professionals (HCP) and developers. The versatility is demonstrated in the main findings, which are standpoint, use, and groups.

Standpoint. The standpoint regarding XAI models focuses on the end-user. Our research revealed that a user's prior knowledge and trust significantly affect their evaluation. In the study of (Wysocki et al., 2023), it was concluded that healthcare professionals with less expertise found the XAI model more helpful even without an explanation. However, a lack of technical knowledge can hinder the ability to interpret both global and local explanations (Naiseh et al., 2023).

Trust. Trust can be divided into affect-based trust, based on humans' emotional responses to AI systems, and cognition-based trust, based on humans' intellectual perceptions of AI reasoning (Naiseh et al., 2023). The lack of technical knowledge or unfamiliarity may decrease user's trust because of misunderstanding. The misunderstanding can be a result of the design of XAI methods, which are made and used by data scientists (Naiseh et al., 2023).

Use. It has been observed that several factors can influence the agreement of HCP with XAI models. These factors include information overload (Naiseh et al., 2023; Wysocki et al., 2023), as well as justification or support of their previous decision (Wysocki et al., 2023). Additionally, there is a risk of automatous bias characterized by an overreliance on the model (Naiseh et al., 2023; Wysocki et al., 2023). Information overload could hinder the use of XAI because of the related cognitive effort (Wysocki et al., 2023). Another aspect why information overload hinders the use of XAI is the time consumption and difficulty to fit XAI into the everyday workflow (Wysocki et al., 2023). Automatous bias is closely linked to the user's trust in the XAI model. In experiments where the XAI model would wrongly discharge the patient, an HCP should be able to apply their clinical judgment and recognize that the patient would require admission (Wysocki et al., 2023). An HCP should be able to override the model's decision.

In (Erdeniz et al., 2022) the term 'Informed decision' is introduced. It means that HCPs should decide whether or not to rely on an XAI model based on their experience and clinical judgment and use explainability as an assistant.

Groups. The identified groups show which professions were noticed in the evaluation process. In the medical sector, experienced and unexperienced healthcare professionals (Slijepcevic et al., 2022; Wysocki et al., 2023), as well as doctors (Kaur et al., 2022; Salahuddin et al., 2021) and clinicians (El Shawi et al., 2019; Slijepcevic et al., 2022), were part of studies. On the other hand, researchers (Li et al., 2020), statisticians (Yu & Wu, 2023), and developers (Schoonderwoerd et al., 2021) played a role as well.

Evaluation Criteria

We observed evaluation criteria to be linked with the user of the XAI model and because of that the evaluation criteria are closely linked with the XAI stakeholders' demands and use.

We found user satisfaction (Jung et al., 2023), representativeness (Naiseh et al., 2023), user trust and reliance (UTR) (Erdeniz et al., 2022), and causability (Müller et al., 2022). Causability is already of great importance in the biomedical domain. There it is used to enable medical experts to reproduce and understand why an algorithm generated a certain result. UTR scores how much a professional agrees with the explanation of an XAI model on a scale from 1 for strongly disagree to 5 for strongly agree (Erdeniz et al., 2022). Representativeness is an information demand of participants. Representativeness describes how many cases could be covered by a given explanation of XAI classes (Naiseh et al., 2023).

Evaluation Methods

The focus of this paper is the evaluation methods utilized for XAI in the healthcare industry. This leads us to the following theme, which is further divided into three subthemes: application-, functionality-, and human-grounded evaluation methods. We split our papers into these categories as shown in Table 4. These topics will be elaborated on in the following sections.

Application-Grounded Evaluation. Application-grounded evaluation refers to quantifying the explanations generated by experts and how they can assist other humans in completing certain tasks (Muddamsetty et al., 2021). This means humans and real tasks are used to do experiments and the viewer will get an immediate interpretation (Wysocki et al., 2023). To ensure the quality of the assessment, experts in the relevant field are tasked with completing specific assignments within the context of the given application. One particular example is using eye-tracking in a specific context. A study collected eye-tracking data of retinal fundus images for medical applications (Muddamsetty et al., 2021). They used recordings to generate and evaluate the explanation provided by the XAI algorithms. Heatmaps of human expert eye fixations were generated and compared to those created by the XAI (Muddamsetty et al., 2021). A second experiment used Diabetic Retinopathy disease levels eye-tracking data recordings to evaluate XAI methods. Three experts' opinions were individually compared to the XAI methods. A different aspect is having the presence of a doctor within an application-grounded evaluation since it will help detect if biases are introduced and are leading to over- or underdiagnosis (Salahuddin et al., 2021).

Functionality-Grounded Evaluation. Functionality-grounded evaluation can be an option, particularly when time and cost are constraints, as it allows for evaluation metrics that do not require human interaction. When assessing the quality of explanations, it can be beneficial to use proxy tasks instead of relying on human evaluation (Salahuddin et al., 2021). One way to evaluate functionality is through the receiver operating characteristics measure (ROC), which is commonly used to assess the similarity between two saliency maps (Salahuddin et al., 2021). The area under the curve (AUC) indicates that random positive instances are ranked higher than negative ones. To visualize these results, an XAI explanation heatmap is used as a binary classifier, helping to keep an overview of true and false positive rates (Muddamsetty et al., 2021). The study conducting the AUC also explained that the evaluation of XAI methods in the medical field should have human experts supervising, to gain trust and transparency (Muddamsetty et al., 2021). One way to determine the importance of variables is through a statistical approach called variable importance evaluation. In a recent study, a new and efficient framework for this evaluation was proposed to interpret nonlinear prediction models for dichotomous outcomes. The researchers used the odds ratio, a widely accepted concept, to quantify variable importance. This method is straightforward and can be easily understood by both statisticians and biomedical scientists (Yu & Wu, 2023). SpRAy, short for Spectral Relevance Analysis, is a unique statistical method that analyzes a model's prediction strategies based on relevance patterns from XAI. Through a clustering technique, it identifies a structure within the given patterns (Slijepcevic et al., 2022).

Human-Grounded Evaluation. Human-grounded evaluation refers to the assessment of simplified tasks performed by non-expert humans. For instance, various explanations would be provided to users and they would have to determine which one is superior (Muddamsetty et al., 2021). In healthcare, it can be difficult to separate human- and application-grounded since healthcare workers and doctors are experts, but they may not have extensive knowledge about AI. However, human-grounded evaluation is an important form of user study. For example, one study asked participants to follow or reject recommendations they believed were generated by an AI-based model, when in reality they were not (Naiseh et al., 2023). Other studies used surveys to assess task completion time, performance accuracy, usefulness, and typicalness of explanations (Jung et al., 2023).

Paper categorization	Total # counted
Application-grounded evaluation	7
Functionality-grounded evaluation	6
Human-grounded evaluation	2
none of the above	5

Table 4: Papers Categorized

Evaluation Metrics

Evaluation metrics evaluate the functional characteristics of an XAI explanation. We found two types of evaluation metrics*:* quantitative and qualitative.

Quantitative Metrics. User studies often use behavioral indicators as a quantitative metric. One such indicator is agreement, which is a binary variable that shows when a participant agrees with an AI recommendation (Naiseh et al., 2023). The opposite is a switch, which indicates when a participant decides to switch from the recommended AI option. However, agreement or switching does not necessarily mean it was the right choice. To address this, a binary variable called Human-AI performance was introduced. This variable shows whether the collaborative Human-AI task resulted in a successful decision, meaning that participants agreed with the correct recommendations and rejected the false ones (Naiseh et al., 2023).

Qualitative Metrics. Stability is a qualitative metric that assesses the consistency of explanations (Siddiqui & Doyle, 2022) for instances that belong to the same class. In other words, if two instances belong to the same class, their explanations should be comparable (El Shawi et al., 2019). Similarity is another metric that measures how similar or dissimilar the explanations for instances are. Both stability and similarity were the most frequently used metrics (El Shawi et al., 2019; ElShawi et al., 2021; Siddiqui & Doyle, 2022). The more similar the instances are, the closer their explanations should be. Specific examples would be the Jaccard Similarity and the Hamming Similarity (Siddiqui & Doyle, 2022). On the other hand, separability measures the dissimilarities between instances. If two instances are dissimilar, their explanations should be different too (El Shawi et al., 2019). Kullback-Leibler Divergence compares the probability density functions of two instances to show their dissimilarity (Muddamsetty et al., 2021). When considering a different metric, it is crucial to be aware of biases and detect the metric. Bias detection is the ability to identify biases in training data explanations (El Shawi et al., 2019). Trust calibration is a complex subject when it comes to AI. Medical professionals need to have a healthy judgment regarding the current state of AI capabilities and decide whether to follow or disregard the recommendations provided by AI (Naiseh et al., 2023).

Ethical and Legal Implications

The next theme discusses a variety of ethical and legal implications regarding the usage of XAI in healthcare. Both have been gaining importance since XAI became more prominent in healthcare.

Legal Implications. Concerning legal matters, there are a few things to consider. Firstly, various regulatory bodies, including the Food and Drugs Administration and the European Parliament, have introduced frameworks for responsible AI (Ahmad et al., 2021). Additionally to the standard usability criteria, other components have gained an increased focus to meet the regulations required for medical AI (El Shawi et al., 2019). For example, the output must be retractable, interpretable, and comprehensible under the necessary regulations (Müller et al., 2022). Despite the introduction of new laws (e.g. GDPR), medical professionals are still expressing concerns about privacy laws, especially as privatized AI applications become increasingly common. A recent survey asked physicians and other medical professionals about their fears regarding AI, and the majority see it as a significant threat (Siddiqui & Doyle, 2022).

Ethical Implications. There are three primary ethical concerns related to AI. The first is known as ethics washing, where the appearance of ethical guidelines is used to cover up a lack of investment in meaningful AI systems and infrastructure (Ahmad et al., 2021). Another issue is discrimination, which can occur if the training data for AI algorithms are not of sufficient quality and inclusivity. In such cases, the AI may learn

to discriminate as a result of gaps in the data, since machine learning systems are only as good as the data they are given (Müller et al., 2022). Finally, inequalities across racial and ethnic groups may reveal systematic problems that need to be corrected by humans, such as clinicians, discharge planners, and vulnerable populations who may be negatively impacted (Ahmad et al., 2021).

Challenges

Our last theme are challenges faced when evaluating XAI in healthcare. In the following text, we will explore some of these challenges in detail.

Explanation and interpretability. During our research, we encountered a challenge distinguishing between the concepts of explainability and interpretability. Although these terms are often used interchangeably, it is important to note that explanation is a result of interpretation. Interpretation involves creating a statement that can be comprehended by a human expert, while the explanation is a compilation of elements from the interpretable domain (Wysocki et al., 2023).

Lack of Controlled Environment. Another challenge that we identified was the COVID-19 pandemic. As a result, some studies had to be conducted online, which was not ideal since face-to-face interactions are preferred. This meant that the researchers couldn't control the devices of individual healthcare professionals, which may have impacted the consistency and quality of the studies. In other studies, HCP could have been provided with more time to work with XAI models to build trust in them. However, this would require a significant time commitment (Wysocki et al., 2023).

Perception of Patient and Clinician. To create XAI models, developers must possess a thorough comprehension of the medical decision-making process, as well as the explanations necessary within it. They should be aware that specific information is crucial for making the correct choices. However, it is important to note whether an explanation is intended for other clinicians or the patient's understanding since it can greatly impact its effectiveness (Schoonderwoerd et al., 2021). Thus, overall explainability is a collaborative process between the explainer and the explainee, involving multiple steps of interaction between humans and AI (Naiseh et al., 2023). Nevertheless, it remains a challenge to explain to experts, that are also understandable by non-experts (Jung et al., 2023).

Discussion

This paper aims to give an overview of existing evaluation approaches for XAI in healthcare using thematic analysis after Braun and Clarke (Braun & Clarke, 2006). The findings of this literature review offer a comprehensive overview of the various methods and techniques used to assess the explainability of XAI models in healthcare.

Principle Findings

During our investigation we found a great diversity of XAI evaluation approaches in healthcare, covering a wide range of aspects. One difficulty during our research was the different use of the term explainability and interpretability. Although they are not the same (Korica et al., 2021) they are often conflated with each other. Our definition of explainability is not equal to interpretability. We will readdress this matter in implication for practice and research.

We distinguished between evaluation metrics and evaluation criteria. Evaluation metrics focus on the functional characteristics while evaluation criteria reflect users' demands. Most of the reviewed studies utilized either or both quantitative and qualitative metrics to evaluate the quality of explanations. Qualitative metrics such as stability and similarity (El Shawi et al., 2019; Siddiqui & Doyle, 2022) were frequently employed to measure the consistency and objectivity of provided explanations. Evaluation criteria like user satisfaction (Jung et al., 2023) and UTR (Erdeniz et al., 2022) were utilized to assess the acceptability and usefulness of explanations from the perspective of users. Our research found that not all papers distinguish between what is classified as a metric and what is classified as an evaluation criterion.

During our research, we implemented the well-established categorization of evaluation methods into application-, functionality-, and human-grounded evaluation. For human-grounded evaluations, non-experts were included in a user study, along with survey assessments and scoring systems. Comparing

explanations of XAI methods with different metrics and criteria was also a commonly used evaluation method.

The most commonly evaluated XAI model was LIME, as well as other local XAI models like SHAP. The reason why a specific model was evaluated was sometimes mentioned. The analysis of XAI stakeholders shows that XAI was mostly evaluated with the use by HCP. No evaluation with a layperson was found. A big part is the discrepancy between developers or other scientists like statisticians, and professionals of the healthcare system. Doctors and clinicians often weren't satisfied by the kind of explanation. The demands of what an explanation should explain differ due to professions. Doctors and clinicians prefer textual explanation while developers and statisticians can interoperate numbers better.

Theme	Main findings
XAI	– XAI-Models divided into local and global explanations – Explanation presentation style is important
Experiment Circumstances	– Different fields of application in healthcare
XAI-Stakeholder	– Groups can be healthcare professionals (doctors, clinicians), statisticians, and researchers – Stakeholders agreeing and making an informed decision over AI explanations
Evaluation Criteria	– Evaluation criteria closely linked to XAI-model users – Examples: feedback, UTR, causability, etc.
Evaluation Methods	– Divided in application-, functionality-, and human-grounded evaluation
Evaluation Metrics	– Quantitative metrics: behavioral indicators such as agreeing or switching from the recommended AI option – Qualitative metrics like stability, similarity, or separability
Ethical, social, and legal implications	– Legal concerns lead to new laws – Ethical concerns are discrimination and systematic problems
Challenges	– Lack of controlled environment – Different perceptions of patients and clinicians

Table 5: Overview of all Identified Themes and Main Findings

Implications for Practice and Research

We hope that our work provides a holistic overview of the topics discussed in the literature and supports the establishment of a fundament for future research. We notice a big interest in the topic of XAI, as well as more specifically its implementation in healthcare. With the rising prominence of this topic, we believe that the handling of AI integration in the healthcare workflow creates a need for a standardized approach as well as clear definitions. In addition, more interdisciplinary is required.

The Need for a Standardized Approach and Definitions

The methods utilized to evaluate the effectiveness of the output from XAI methods were heterogeneous. We encountered a wide variety of approaches which ranged from user studies with non-professionals to defining metrics and using them to evaluate how good the explanations performed. There is little common ground and therefore it is difficult to compare the results from various studies. If best practices and norms were constituted, XAI Methods and their performance could better be compared. Best practices shouldn't be all general, since the use still depends on the context. A possible solution could be to strongly categorize XAI methods and their intention.

Furthermore, there is also a need for uniform definitions to make the results more comparable. As highlighted in our work, the conflation of explainability and interpretability, or the distinction between

metrics and evaluation criteria is lacking and poses problems for the comparison of XAI evaluation methods and research.

More Interdisciplinarity

While an ideal XAI model would provide explanations satisfying all the user demands and criteria, the realization of such a model is hard since demands are dependent on the user. They are not only influenced by the medical domain of the HCPs but also the personal and subjective preferences. Additionally, there's a discrepancy between developers and HCP. It may be more feasible to provide the next generations of HCPs with an additional amount of education in statistics and informatics so that they are more comfortable with the output generated by XAI. This can lead to more trust by HCP and consequently more informed decision-making with XAI.

Limitations and Future Research

In our research, we encountered a vast amount of potential literature. Due to time constraints, we were able to analyze only 20 out of the 30 papers we identified as feasible, thereby limiting our scope. Additionally, our focus on the SCOPUS database, chosen for its coverage of diverse domains such as technology and medicine, may have led us to overlook significant publications from other databases. These limitations may have impacted the selection of themes.

Furthermore, the selection and coding of literature following Braun and Clarke (Braun & Clarke, 2006) inherently involves subjective judgment. Our individual preferences and opinions inevitably influenced the results, but through conducting most steps in pairs we tried to stay as objective as possible.

Given these limitations, future research should address the broader spectrum of available literature and include additional databases with similar focus points as Scopus (e.g. IEEE Xplore) to ensure a more comprehensive review. Expanding the scope to encompass more databases and papers would enhance the reproducibility and robustness of the findings. Additionally, the inclusion of preprints could offer insights into the most up-to-date findings, as the research on XAI is rapidly evolving.

Additionally, a more fine-grained analysis of different medical fields and their specific requirements and use cases for XAI could significantly advance our understanding of the topic. Conducting focused reviews in the heterogeneous subfields of medicine and AI could yield more targeted insights. Such detailed reviews would contribute to developing a scientific consensus and guide further research in this interdisciplinary area.

Conclusion

In this work, we aimed to systematically analyze relevant literature for the topic of evaluation of XAI in healthcare. We identified eight overarching themes, namely XAI, XAI-Stakeholder, Experiment Circumstances, Ethical and Legal Implications, Challenges, Evaluation Metric, Evaluation Criteria, and Evaluation Method. The interest in this topic seems to be high, but due to its infancy, there is little common ground and standards. Our review could support future research as a starting point or as an orientation.

References

Ahmad, M. A., Overman, S., Allen, C., Kumar, V., Teredesai, A., & Eckert, C. (2021). Software as a Medical Device: Regulating AI in Healthcare via Responsible AI. *Proceedings of the 27th ACM SIGKDD Conference on Knowlsedge Discovery & Data Mining*, 4023–4024. **https://doi.org/10.1145/3447548.3470823**

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/10.1016/j.inffus.2019.12.012**

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. **https://doi.org/10.1191/1478088706qp063oa**

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), 832. **https://doi.org/10.3390/electronics8080832**

Dieber, J., & Kirrane, S. (2022). A novel model usability evaluation framework (MUsE) for explainable artificial intelligence. *Information Fusion*, *81*, 143–153. **https://doi.org/10.1016/j.inffus.2021.11.017**

El Shawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2019). Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 275–280. **https://doi.org/10.1109/CBMS.2019.00065**

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. **https://doi.org/10.1111/coin.12410**

Erdeniz, S. P., Veeranki, S., Schrempf, M., Jauk, S., Tran, T. N. T., Felfernig, A., Kramer, D., & Leodolter, W. (2022). *Explaining Machine Learning Predictions of Decision Support Systems in Healthcare*. **https://doi.org/10.1515/cdbme-2022-1031**

Hakkoum, H., Abnane, I., & Idri, A. (2022). Evaluating Interpretability of Multilayer Perceptron and Support Vector Machines for Breast Cancer Classification. *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, 1–6. **https://doi.org/10.1109/AICCSA56895.2022.10017521**

Hammoud, I., Prasanna, P., Ramakrishnan, I., Singer, A., Henry, M., & Thode, H. (2022). EventScore: An Automated Real-time Early Warning Score for Clinical Events. *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, 192–200. **https://doi.org/10.1109/ICHI54592.2022.00038**

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4), e1312. **https://doi.org/10.1002/widm.1312**

Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, *9*(5), e16110. **https://doi.org/10.1016/j.heliyon.2023.e16110**

Kaur, D., Uslu, S., & Durresi, A. (2022). Trustworthy AI Explanations as an Interface in Medical Diagnostic Systems. In L. Barolli, H. Miwa, & T. Enokido (Eds.), *Advances in Network-Based Information Systems* (Vol. 526, pp. 119–130). Springer International Publishing. **https://doi.org/10.1007/978-3-031-14314-4_12**

Kaur, D., Uslu, S., Durresi, A., Badve, S., & Dundar, M. (2021). Trustworthy Explainability Acceptance: A New Metric to Measure the Trustworthiness of Interpretable AI Medical Diagnostic Systems. In L. Barolli, K. Yim, & T. Enokido (Eds.), *Complex, Intelligent and Software Intensive Systems* (Vol. 278, pp. 35–46). Springer International Publishing. **https://doi.org/10.1007/978-3-030-79725-6_4**

Korica, P., Gayar, N. E., & Pang, W. (2021). Explainable Artificial Intelligence in Healthcare: Opportunities, Gaps and Challenges and a Novel Way to Look at the Problem Space. In H. Yin, D. Camacho, P. Tino, R. Allmendinger, A. J. Tallón-Ballesteros, K. Tang, S.-B. Cho, P. Novais, & S. Nascimento (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2021* (Vol. 13113, pp. 333–342). Springer International Publishing. **https://doi.org/10.1007/978-3-030-91608-4_33**

Li, Y., Fujiwara, T., Choi, Y. K., Kim, K. K., & Ma, K.-L. (2020). A visual analytics system for multi-model comparison on clinical data predictions. *Visual Informatics*, *4*(2), 122–131. **https://doi.org/10.1016/j.visinf.2020.04.005**

Lötsch, J., Kringel, D., & Ultsch, A. (2021). Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedInformatics*, *2*(1), 1–17. **https://doi.org/10.3390/biomedinformatics2010001**

Maddox, T. M., Rumsfeld, J. S., & Payne, P. R. O. (2019). Questions for Artificial Intelligence in Health Care. *JAMA*, *321*(1), 31. **https://doi.org/10.1001/jama.2018.18932**

Muddamsetty, S. M., Jahromi, M. N. S., & Moeslund, T. B. (2021). Expert Level Evaluations for Explainable AI (XAI) Methods in the Medical Domain. In A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, & R. Vezzani (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges* (Vol. 12663, pp. 35–46). Springer International Publishing. **https://doi.org/10.1007/978-3-030-68796-0_3**

Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., & Zatloukal, K. (2022). Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. *New Biotechnology*, *70*, 67–72. **https://doi.org/10.1016/j.nbt.2022.05.002**

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, *169*, 102941. **https://doi.org/10.1016/j.ijhcs.2022.102941**

Rao, J., Zheng, S., Lu, Y., & Yang, Y. (2022). Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns*, *3*(12), 100628. **https://doi.org/10.1016/j.patter.2022.100628**

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* (Version 3). arXiv. **https://doi.org/10.48550/ARXIV.1602.04938**

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2021). *Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods* (Version 1). arXiv. **https://doi.org/10.48550/ARXIV.2111.02398**

Schoonderwoerd, T. A. J., Jorritsma, W., Neerincx, M. A., & Van Den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, *154*, 102684. **https://doi.org/10.1016/j.ijhcs.2021.102684**

Siddiqui, K., & Doyle, T. E. (2022). Trust Metrics for Medical Deep Learning Using Explainable-AI Ensemble for Time Series Classification. *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 370–377. **https://doi.org/10.1109/CCECE49351.2022.9918458**

Singh, A., Jothi Balaji, J., Rasheed, M. A., Jayakumar, V., Raman, R., & Lakshminarayanan, V. (2021). Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis. *Clinical Ophthalmology*, *Volume 15*, 2573–2581. **https://doi.org/10.2147/OPTH.S312236**

Slijepcevic, D., Horst, F., Lapuschkin, S., Horsak, B., Raberger, A.-M., Kranzl, A., Samek, W., Breiteneder, C., Schöllhorn, W. I., & Zeppelzauer, M. (2022). Explaining Machine Learning Models for Clinical Gait Analysis. *ACM Transactions on Computing for Healthcare*, *3*(2), 1–27. **https://doi.org/10.1145/3474121**

Wang, Y.-C., Chen, T.-C. T., & Chiu, M.-C. (2023). An improved explainable artificial intelligence tool in healthcare for hospital recommendation. *Healthcare Analytics*, *3*, 100147. **https://doi.org/10.1016/j.health.2023.100147**

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*, 103839. **https://doi.org/10.1016/j.artint.2022.103839**

Yu, D., & Wu, H. (2023). Variable importance evaluation with personalized odds ratio for machine learning model interpretability with applications to electronic health records-based mortality prediction. *Statistics in Medicine*, *42*(6), 761–780. **https://doi.org/10.1002/sim.9642**

Appendix

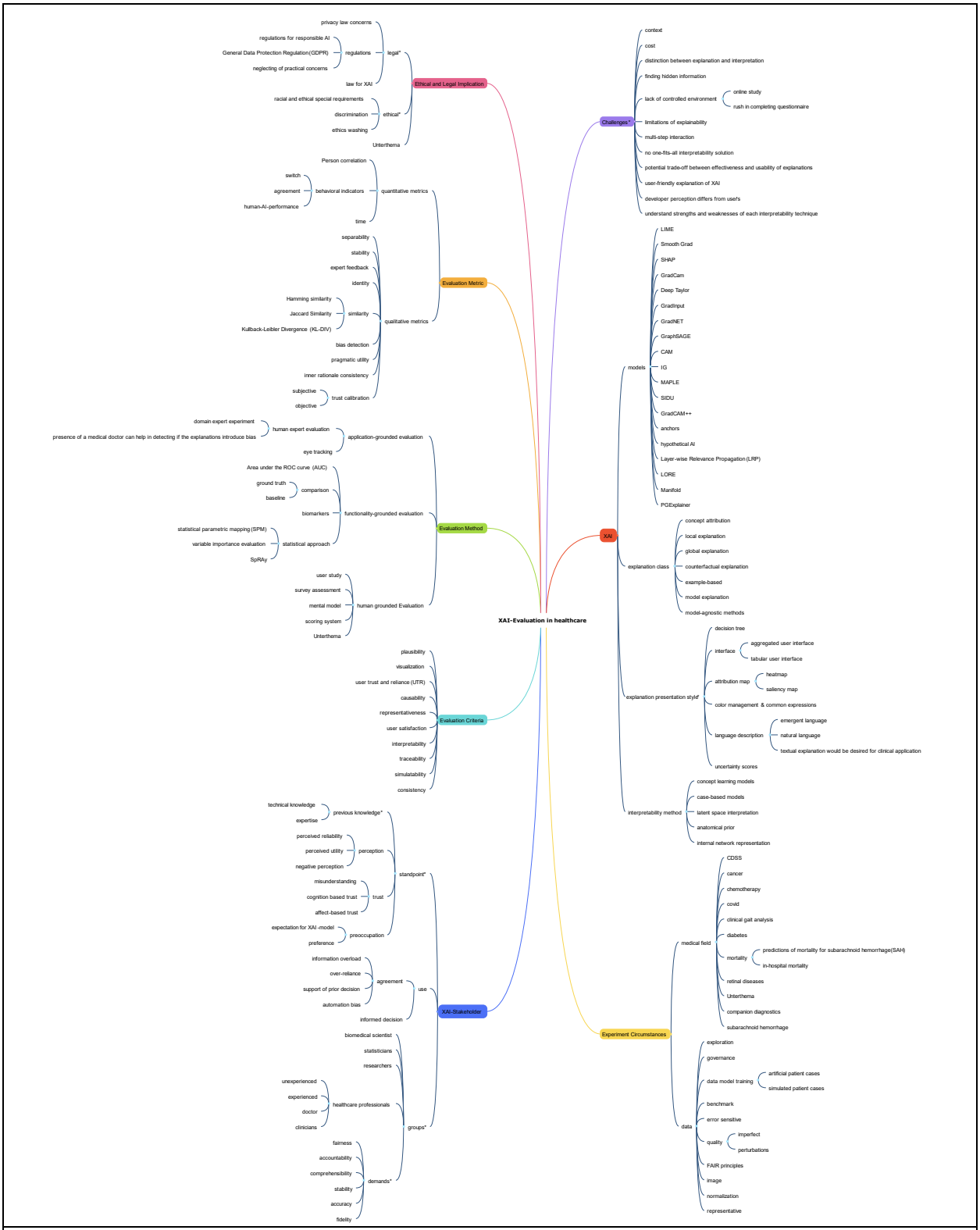

Figure A-1: Mind Map with the Extracted Codes from the Analyzed Literature. Codes Marked with * Were Subsequently Added

Preprocessing Approaches for Informed Machine Learning in Medical Imaging

Emerging Trends in Digital Health, Summer Term 2023

Vasileios Xanthakis
Master Student
Karlsruhe Institute of Technology
vasileios.xanthakis@student.kit.edu

Kieu Anh Dang
Master Student
Karlsruhe Institute of Technology
kieu.dang@student.kit.edu

Abstract

Background: *In recent years, the integration of machine learning algorithms and neural networks in medical image classification has become increasingly significant. These technologies offer promising capabilities for assisting healthcare professionals in interpreting various types of medical images, such as X-rays, MRIs, CT scans, and pathology slides. However, the effectiveness of these algorithms heavily relies on the availability of sufficient and high-quality training data, which often presents a bottleneck due to limited and inaccessible patient datasets.*

Objective: *This paper aims to explore how informed machine learning approaches can improve the accuracy and effectiveness of medical image classification models by integrating domain knowledge into the imaging algorithm. Specifically, the research investigates preprocessing steps utilized in this process, focusing on both the input data and the domain knowledge.*

Methods: *A systematic review of 80 publications was conducted to identify and analyze the preprocessing approaches employed in integrating domain knowledge into medical image classification tasks. The study categorized the preprocessing steps into three main categories: "Data Preprocessing," "Knowledge Preprocessing," and "Knowledge-Enhanced Data Preprocessing." Additionally, the characteristics of the domain knowledge were examined to understand the relationships between preprocessing steps and the structure of available knowledge.*

Results: *The analysis revealed that the adaptation of existing image preprocessing pipelines and machine learning image classification tasks required minimal changes. Instead, the emphasis was on preprocessing the domain knowledge and designing a compatible architecture. The findings highlighted the importance of effectively incorporating specialized knowledge and expert input to enhance model accuracy and effectiveness.*

Conclusion: *Informed machine learning approaches, which integrate domain knowledge into medical image classification algorithms, offer promising opportunities to improve diagnostic accuracy and efficiency in healthcare. By focusing on preprocessing steps related to both data and domain knowledge, researchers can design more effective and compatible models, ultimately advancing the capabilities of medical image analysis systems.*

Keywords: medical imaging, expert knowledge, knowledge informed AI, explainability, preprocessing

Introduction

Medical imaging plays a vital role in the examination and early identification of diseases. Nonetheless, a significant hurdle lies in the laborious screening and analysis process performed by physicians. This predicament leads to inefficiency, extended working hours for doctors, and a restricted number of medical images that can be thoroughly assessed (Stoitsis et al., 2006). As a result, it becomes imperative to explore the potential application of artificial intelligence (AI) to alleviate this burden and enhance efficiency. Specifically, Machine Learning (ML) techniques can be employed for a range of tasks, including object classification, segmentation, and localization (Yoon et al., 2019). However, machine learning encounters a noteworthy challenge due to the scarcity of extensive datasets, which complicates achieving optimal performance (Roth et al., 2020).

Due to the limited availability of extensive datasets, alternative approaches are being sought to enhance the training process of machine learning algorithms. One potential avenue involves incorporating domain knowledge to augment existing datasets with supplementary material (Lee & He, 2019). However, domain knowledge originates from diverse sources and possesses various forms of representation (von Rueden et al., 2023). The preprocessing requirements involved in this integration process remain unclear. Consequently, the present challenge lies in comprehending the procedures based on the characteristics of domain knowledge, thereby enabling the appropriate integration of domain knowledge into imaging-based machine learning algorithms within each distinct scenario. To provide more insight on this topic, the following research questions arise:

Research question 1: What are the preprocessing requirements for the integration of domain knowledge depending on the structure of the knowledge?

Research question 2: What underlying connections and trends between domain knowledge structure and preprocessing approaches can be identified?

By systematically reviewing a collection of 80 papers and creating a comprehensive overview of preprocessing procedures for the integration of domain knowledge, this study uncovers valuable insights.

For the integration of domain knowledge, the preprocessing steps can be distinguished between Data Preprocessing, Knowledge Processing, and Knowledge-enhanced Data Preprocessing. However, the results also show that the preprocessing steps required for incorporating domain knowledge exhibit minimal divergence from standard Data Preprocessing. Consequently, the approach to preprocessing data and incorporating domain knowledge should be approached holistically, as there exists a blurred distinction between the two processes. Lastly, the surveyed literature emphasizes the importance of architectural design. A well-structured and purpose-driven architectural design plays a pivotal role in ensuring the seamless integration of domain knowledge.

This paper is organized as follows: First, the background and essential terminologies in the context of medical imaging and informed machine learning are introduced. Next, the methodology used in this study is outlined. An overview of the results is presented, and the proposed categorization is explained based on examples from the examined literature. In the discussion sections, the principal findings of this work are presented, and their implications are considered, limitations of the followed approach are pointed out, and prospects for future research are presented.

Background

In this background chapter, the foundation for this thesis on the integration of domain knowledge in medical imaging AI systems is laid. The chapter introduces and defines key concepts essential to understanding the scope, categorization, results, and findings of our work. The primary concepts explored here are medical imaging and the application of machine learning in this field, domain knowledge, including the characteristics and types of Expert Knowledge, and Data Preprocessing, with particular emphasis laid on the preprocessing of imaging data in Convolutional Neural Networks (CNNs). The interaction between these concepts is highlighted, setting the stage for a more in-depth exploration in the main part of the paper.

Medical Imaging

Medical images are visual representations of various aspects of the human body or medical conditions captured using imaging techniques. These images play a vital role in diagnosing, treating, and monitoring illnesses and injuries. They provide valuable insights into the internal structures, functions, and abnormalities within the body (Panayides et al., 2020). The term "medical imaging" refers to the broad, multidisciplinary field concerned with the acquisition, processing, visualization, and interpretation of structural and functional images of living organisms and is often employed in clinical or research applications (Meijering, 2020). Typically, these images arise from the interaction between electromagnetic waves of varied wavelengths and biological tissues. Only ultrasound employs mechanical soundwaves. The prevailing imaging methodologies comprise X-ray computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), single-photon emission computed tomography (SPECT), and ultrasound (US) imaging. To utilize these imaging technologies effectively, it is essential to acquire high-quality images as well as accurate interpretation (Panayides et al., 2020).

This connection between image quality and interpretation is integral within the realm of "Imaging Informatics," a comprehensive domain encompassing the entire spectrum of activities related to medical image creation, distribution, storage, retrieval, processing [...], interpretation, reporting, and communication (Panayides et al., 2020). Focusing specifically on medical analysis, this study focuses on three vital tasks: (1) diagnosing diseases, (2) detecting lesions, organs, and abnormalities, and (3) segmenting lesions and organs.

Despite significant enhancements in sensitivity, efficiency, and image quality of medical imaging instruments over the past decades, the task of image interpretation has traditionally relied on human expertise. However, even seasoned professionals are susceptible to subjectivity, variability, and fatigue, which can affect diagnostic accuracy (Kwee & Kwee, 2021). Recognizing these challenges, there is a growing interest in exploring avenues to enhance workplace efficiency, automate repetitive tasks, and facilitate collaborative analysis with multiple experts. To implement these remedies, computational methods like artificial intelligence can be utilized (Yoon et al., 2019).

Machine Learning for Medical Imaging

Current research shows that artificial intelligence (AI) is exhibiting its capacity to enhance efficiency within the medical domain, exemplified by its capabilities to tackle tasks like patient health monitoring, clinical trial participation, or medication management (Väänänen et al., 2021). In this context, machine learning techniques prove to be instrumental. Through the employment of machine learning algorithms, the analysis, classification, and interpretation of medical images are facilitated. Consequently, this enables radiologists to receive support in detecting anomalies, recognizing intricate patterns, and ultimately arriving at precise diagnoses. This synergy between AI and machine learning augments the diagnostic capabilities of medical practitioners (Nomani et al., 2022).

For machine learning algorithms, however, the lack of available datasets poses a challenge. This challenge arises due to the unique characteristics of medical data and the intricacies of acquiring, annotating, and utilizing such data for machine-learning purposes. As medical images are often generated in clinical settings with strict ethical and privacy considerations, access to these images is restricted. Consequently, the availability of large and diverse datasets for training robust machine learning models is limited. The process of annotation must also deal with certain issues. Moreover, as annotating medical images requires specialized expertise, often involving the interpretation of complex anatomical structures, pathologies, and subtle abnormalities, it can be time-consuming and costly to obtain accurate annotations. Lastly, as certain medical conditions are rare, accumulating sufficient data for training models specifically tailored to these conditions also poses a hurdle. This scarcity hampers the development of effective solutions for diagnosing or treating rare diseases (Xie et al., 2020).

Domain Knowledge

Owing to the scarcity of available data, the integration of informed machine learning assumes significant importance. Informed machine learning, in this context, presents a learning process that derives insights from a cohesive knowledge repository, encompassing inputs from both data driven sources and prior

knowledge. This expertise, stemming from an independent source, is conveyed through structured representations, and integrated into the machine-learning pipeline. Informed machine learning can help mitigate the challenge of lack of data by leveraging domain knowledge, expert insights, and existing information to enhance the training and performance of machine learning models (von Rueden et al., 2023).

Definition and Characteristics of Domain Knowledge

An aspect that plays a vital role in the incorporation of informed machine learning in the clinical praxis is domain knowledge. Domain knowledge in medical imaging refers to specialized expertise and understanding of the principles, practices, and intricacies involved in the acquisition, interpretation, and analysis of medical images. It encompasses a deep understanding of both the medical and imaging aspects of the field, allowing professionals to effectively utilize and interpret various imaging modalities for diagnostic and clinical purposes (Xie et al., 2020). Medical imaging domain knowledge can be categorized into five distinct areas: the training patterns, the overall diagnostic patterns employed when viewing images, the specific regions they typically emphasize, the features (such as characteristics, structures, and shapes) they particularly prioritize, and additional information used for diagnostic purposes (Xie et al., 2020).

Integrating domain knowledge into a machine learning algorithm requires a translation process, wherein the domain expertise is transformed into a format that is both comprehensible and actionable for the system. Hence, for the process of domain knowledge integration preprocessing data, domain knowledge, as well as the examination of the machine learning model, are necessary. It is therefore important to take a closer look at the sources, structures, and steps of applications to gain a more profound insight. To have a mutual understanding of these concepts, this paper uses the structure and definitions of the study from von Rueden et al. (2023).

Source

The source of categorical knowledge pertains to the origin of prior knowledge to be incorporated into machine learning. This source can stem from a recognized domain of knowledge as well as from insights gathered from a specific group of individuals with relevant expertise. Different types of sources of domain knowledge are presented in Table 1.

Source	Definition
Scientific Knowledge	We subsume the subjects of science, technology, engineering, and mathematics under Scientific Knowledge.
World Knowledge	By World Knowledge we refer to facts from everyday life that are known to almost everyone and can thus also be called general knowledge. It can be more or less formal.
Expert Knowledge	We consider Expert Knowledge to be knowledge that is held by a particular group of experts. Within the expert's community it can also be called common knowledge. Such knowledge is rather informal and needs to be formalized, e.g., with human-machine interfaces.

Table 1. Definition of Different Sources of Domain Knowledge as Presented in von Rueden et al. (2023)

Structure (Representation)

The structure of knowledge depicts how knowledge is formally represented. It is important to note, that different representations can be mathematically transformed into each other. Nevertheless, reviewed articles are mostly classified into a specific type of knowledge structure. The main forms in which domain knowledge is typically represented are Algebraic Equations, Differential Equations, Simulation Results, Spatial Invariances, Logic Rules, Knowledge Graphs, Probabilistic Relations, and Human Feedback. The different structures of knowledge are shown in Table 2.

Structure	Definition
Algebraic Equations	Algebraic Equations represent knowledge as equality or inequality relations between mathematical expressions consisting of variables or constants.
Differential Equations	Differential Equations are a subset of Algebraic Equations, which describe relations between functions and their spatial or temporal derivatives.
Simulation Results	Simulation Results describe the numerical outcome of a computer simulation, which is an approximate imitation of the behavior of a real-world process.
Spatial Invariances	Spatial Invariances describe properties that do not change under mathematical transformations such as translations and rotations.
Logic Rules	Logic provides a way of formalizing knowledge about facts and dependencies and allows for translating ordinary language statements (e.g., IF A THEN B) into formal Logic Rules (A ñ B).
Knowledge Graphs	A graph is a pair pV, Eq, where V are its vertices and E denotes edges.
Probabilistic Relations	Prior knowledge could be assumptions on the conditional independence or the correlation structure of random variables or even a full description of the joint probability distributions.
Human Feedback	Human Feedback refers to technologies that transform knowledge via direct interfaces between users and machines.

Table 2. Definition of Different Structures of Domain Knowledge as Presented in von Rueden et al. (2023)

Step of Application

Steps of application describe the part in the machine learning pipeline in which knowledge is integrated. The majority of integration approaches can be classified under the four main components of machine learning systems, these being Training Data, Hypothesis Set, Learning Algorithm, and Final Hypothesis. An overview of different integration options can be seen in Table 3.

Source	Definition
Training Data	A standard way of incorporating knowledge into machine learning is to embody it in the underlying Training Data.
Hypothesis Set	Integrating knowledge into the Hypothesis Set is common, say, through the definition of a neural network's architecture and hyper-parameters.
Learning Algorithm	Learning Algorithms typically involve a loss function that can be modified according to additional knowledge, e.g., by designing an appropriate regularizer.
Final Hypothesis	The output of a learning pipeline, i.e., the Final Hypothesis, can be benchmarked or validated against existing knowledge.

Table 3. Definition of Different Steps of Application of Domain Knowledge as Presented in von Rueden et al. (2023)

Data Preprocessing

Data Preprocessing is essential for any AI application as it is aimed at refining raw data to enable effective and precise analysis by machine learning algorithms. In doing so, it serves to mitigate the risk of pursuing inappropriate analysis methodologies for the given dataset. Furthermore, this preparatory stage imparts a deeper understanding of the data's inherent characteristics, thereby facilitating more meaningful analytical insights. Consequently, this process enhances the capacity to glean more profound knowledge from the dataset at hand (J. Huang et al., 2021).

In most practical applications, a multifaceted approach to Data Preprocessing is imperative. As Data Preprocessing encompasses a series of preparatory steps that precede the initiation of the actual data analysis phase, identifying the specific category of Data Preprocessing required assumes a pivotal role in this context. In comparison to data that has not undergone standardization, Data Preprocessing contributes to the improvement of classification outcomes significantly. Therefore, Data Preprocessing constitutes an indispensable step within the framework of any machine learning pipeline, underlining its importance in the broader spectrum of data-driven analysis (Famili et al., 1997).

Data Preprocessing in CNNs

In image-based problems, CNNs have been highlighted frequently. CNNs are a type of deep learning architecture commonly used for tasks involving image and spatial data, such as image classification, object detection, and segmentation. To preprocess data in CNNs preparatory steps and techniques are applied to input data before it is fed into the CNN model for training or inference (Tang et al., 2018).

For CNNs, the objective of Data Preprocessing is to ensure that the input data is well-structured, normalized, and augmented, enabling the CNN model to learn meaningful patterns and features effectively. By optimizing the data input, preprocessing contributes to the overall performance and generalization capabilities of the network. In the case of CNNs, it has been shown to improve the robustness of classification tasks as well as the recognition rate and efficiency of implemented algorithms (Xu et al., 2019).

Method

To examine the research questions, a systematic literature review was conducted, and a comprehensive overview was developed. The systematic literature review followed a structured process proposed by (Webster & Watson, 2002), involving a thorough selection, analysis, and evaluation of relevant studies. Afterward, an overview development approach was applied inspired by (Nickerson et al., 2013), which included classifying the findings into categories. Lastly, the identified categories of preprocessing steps were mapped to the structure of knowledge. The entire process is depicted in Figure 1.

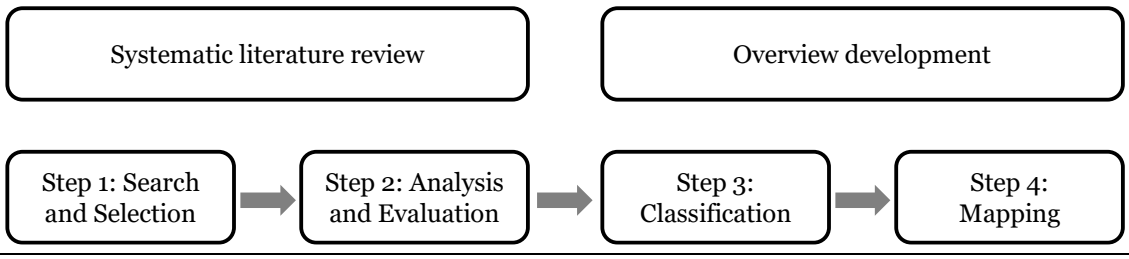

Figure 1. Analysis Steps

Search and Selection of Studies

The selection of studies for this research was based on a previous paper titled "Medical Informed Machine Learning: A scoping review and future research directions," wherein 177 relevant papers were analyzed. The search process was conducted across various digital databases, namely ACM Digital Library, AIS eLibrary, IEEE Xplore, ProQuest, PubMed, and Scopus, using a specific search string: string: TI-KE-AB[("machine learning" OR "artificial intelligence" OR "deep learning") AND (health* OR medic*) AND ("domain knowledge" OR expert-based OR theory-guided OR theory-driven OR physics-informed OR physics-guided).

In the context of the presented research, the selected papers were subject to comprehensive analysis concerning the utilization of domain knowledge, the implementation of machine learning models, and the identification of future research directions. It was stated that, in particular, the source and structure of domain knowledge need further examination. Therefore, the chosen articles from this study were further utilized to explore the technical and preprocessing requirements for the integration of domain knowledge.

Out of the 177 papers considered, a significant majority focused on medical images. Due to resource limitations, the decision was made to prioritize the most relevant area, which led to the selection of medical images as the primary subject of investigation in this study. As a result, a total of 80 papers specifically related to medical images were thoroughly examined during this research.

Analysis and Evaluation

For the analysis, the selected papers underwent a dual-authored review protocol. Initially, the first ten papers were jointly read and analyzed by both authors to establish a common comprehension and concordance concerning the coding procedure. Subsequently, the remained papers were evenly distributed between the two authors, with each autonomously encoding 30 papers. After all papers were analyzed, the results were consolidated, and individual cases were discussed. This methodological approach fosters a comprehensive and dependable evaluation of the considered papers.

To examine the technical and preprocessing requirements of domain knowledge integration, the preprocessing steps were first identified for domain knowledge and data integration individually. After extracting all specific preprocessing steps, general preprocessing steps were concluded. During the coding process, however, a third aspect of preprocessing was identified. This is explained in the section "Knowledge Enhanced Data Preprocessing". The example provided in Table 4 demonstrates the framework for capturing and classifying the preprocessing steps instituted within each paper.

Title	**Prepros. step data (general)**	**Prepros. step data (specific)**	**Prepros. step knowledge (general)**	**Prepros. step knowledge (specific)**
A computational framework for cancer response assessment based on oncological PET-CT scans	Masking	Automatic Tumor Segmentation Mask	ROI Segmentation	Segmentation masks created by physicians

Table 4. Classification Table

Classification

To provide a comprehensive overview of this topic, the development of a taxonomy was advisable. However, as only two authors took part in this research, the execution of a taxonomy development process was not feasible. Consequently, the adoption of an overview development approach proved to be more appropriate. Therefore, the framework for the overview development, based on the work of (Nickerson et al., 2013), was employed with specific modifications made to accommodate the absence of comprehensive taxonomy development. Upon careful review and extensive discussion, a deductive-inductive approach was employed for the coding process.

During the coding process, specific objective, and subjective ending conditions, as well as definitions for various categories, were considered to ensure the rigor and quality of the coding results. These ending conditions, in conjunction with the definitions for each category, were employed to precisely delineate the categories, ensuring no essential aspects are overlooked. In its entirety, the overview development process adhered to five specific objective ending conditions. The first ending condition described the analysis of a representative sample of objects. In addition, at least one object must be classified under each characteristic within every category. Due to a lack of resources, however, objects that could be unequivocally assigned to categories were excluded from further scrutiny. The overview development reached its final stage when no new categories or characteristics were introduced, combined, or divided during the last iteration. Moreover, every category was expected to be distinct and non-repetitive.

Among the subjective ending conditions, the categorization had to exhibit conciseness, robustness, and explanatory power to fulfill the established criteria. The conciseness described the quantity and quality of categories without having an excessive number of different categories.

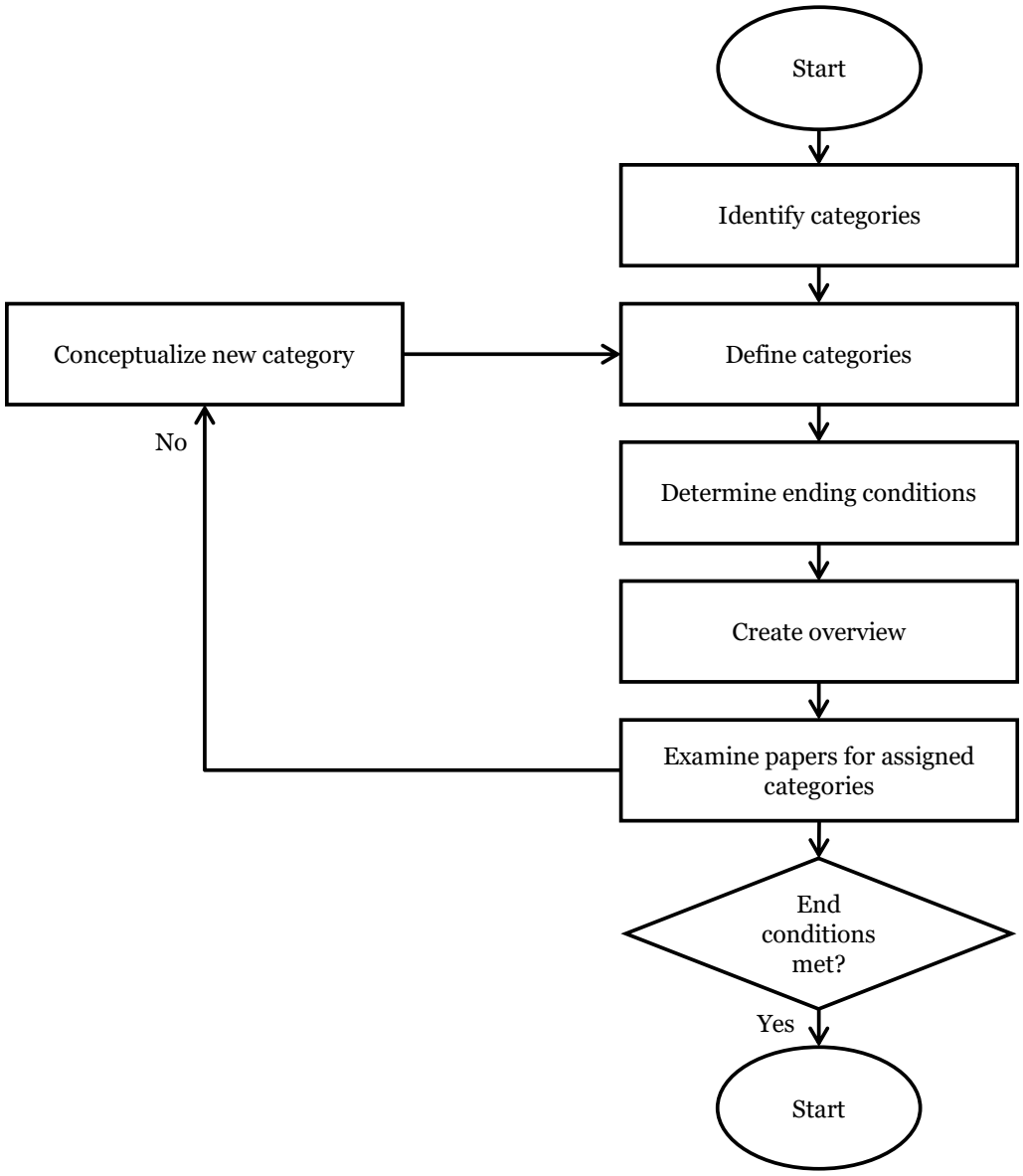

Figure 2. Classification Process

Furthermore, categorization had to be robust, which means that distinct categories needed to be distinguished sufficiently to be of interest. However, a cautious approach was taken to this requirement because the investigated topic has inherent complexities that might make it difficult to clearly separate categories. While effort was made to create clear distinctions between categories, it was vital to recognize that certain categories could have connections or overlaps. This approach was chosen because it made sense within the context of this study's subject matter. These connections between categories helped to reveal important and relevant insights.

Once the ending conditions were defined, the categories were identified defined, the overarching overview was constructed. The process ended when the predefined ending conditions were met. In cases where these conditions were not met, a novel category was conceptualized. This procedural flow is visually represented in Figure 2.

Mapping

After developing a classification of different categories of preprocessing steps, the subsequent task involved examining how different preprocessing methods relate to the underlying source and structure of knowledge. Due to the limited number of examined papers, the focus was laid solely on the structure of knowledge. The different structures of knowledge that were considered for the mapping process are based on the study of von Rueden et al. (2023).

To establish these connections, the identified preprocessing steps were systematically linked to the structures of domain knowledge as outlined in the background section. The organization and execution of this correlation were facilitated using the spreadsheet software Excel. After the mapping of knowledge structures onto the preprocessing methods, the outcomes underwent comprehensive evaluation and interpretation.

Results

During the examination of the specified literature, we identified three key prerequisites that need to be fulfilled in order to properly integrate domain knowledge in machine learning systems. These requirements (explained below) are preprocessing of the data, preprocessing of the domain knowledge, and adaption of the architecture design.

- Data Preprocessing: This refers to the preprocessing of the training data, for it to be compatible with the system. In the case of imaging applications, this includes both image preprocessing and metadata preprocessing.
- Knowledge Preprocessing: The available knowledge needs to be coded in a way that enables it to be interpretable by the computer and compatible with the machine learning algorithm.
- Architecture Design: The architecture of the machine learning system needs to be adapted to accommodate the domain knowledge being added to the system. This can either include additional steps or channels for the additional information or a complete redesign of the overall architecture and pipeline.

We have classified the first two prerequisites (Data Preprocessing, Knowledge Preprocessing) as the necessary preprocessing steps for the incorporation of domain knowledge in Imaging AI algorithms. The third aspect of domain knowledge integration (Architecture Design) needs to be addressed much earlier in the development process as it lays the foundation and sets the overall structure for the entire system during the conceptual stage of development (Baragry & Reed, 1999). For this reason, it lies outside the scope of this paper and was not examined during our research.

The following tables summarize the main preprocessing requirements that were identified in the available dataset according to the specific characteristics of the domain knowledge.

Although the literature was examined in regard to both the source and the structure of the integrated knowledge, we have chosen to only focus on the structure of the knowledge for the purposes of this categorization. The reasons for this approach lie with the limited number of works available and the variety of identified categories. Choosing to examine all possible combinations between structure and source would only allow for the classification of a handful of research projects in each category, which is not sufficient for formulating concrete and robust conclusions on the discussed approaches. Between the two characteristics, the structure of the knowledge was found to be much more relevant for the process of integration in the machine learning system, as we are examining the process of transforming data from the available structure to a system-compatible format. The categorization will therefore be based on the structure of the available domain knowledge, as it was shown to directly impact the necessary preprocessing steps.

It is also important to clarify, that the presented steps are not the sole preprocessing measures implemented in the available literature, but rather the ones that were most stressed and discussed in each scenario. Additionally, numerous examined works were found to employ a combination of preprocessing steps, while certain standardized preprocessing methods were omitted from some of the papers but are still assumed to have been part of the overall analysis process.

Data Preprocessing

In the following, the main observed preprocessing steps will be introduced, and their relevance in the scope of this thesis will be explained. An overview of the preprocessing methods is provided in Table 5.

Preprocessing Method	Knowledge Structure	Papers
Image Preprocessing Pipeline	Algebraic Equations	(Chae et al., 2022)
		(Bian et al., 2022)
		(Yaman et al., 2021)
	Probabilistic Relations	(Andrushia et al., 2021)
		(Yu et al., 2021)
		(Qoku & Buettner, 2022)
		(Tardy & Mateus, 2021)
	Spatial Invariances	(Pezzotti et al., 2020)
		(Cheng et al., 2021)
		(Nguyen et al., 2021)
		(Qiu, 2020)
Data Augmentation	Logic Rules	(Z. Zhang et al., 2020)
		(Hemelings et al., 2020)
		(Lizhi et al., 2021)
	Spatial Invariances	(J. Huang et al., 2021)
		(Zhou et al., 2017)
		(Xie et al., 2021)
	Probabilistic Relation	(Gou et al., 2021)
		(Tian et al., 2021)

Table 5. Categorization of Data Preprocessing Steps According to the Structure of the Domain Knowledge

Image Preprocessing Pipeline

The most prominent preprocessing procedure is the implementation of an Image Preprocessing Pipeline for the preparation of training and test samples. This pipeline encompasses standard image preprocessing techniques commonly employed in Convolutional Neural Network (CNN) systems. These include data cleansing and integration, selection and transformation, data mining, masking, as well as evaluation and presentation (Maharana et al., 2022). Such preprocessing steps are typically implemented by the computer through predefined preprocessing algorithms without requiring explicit input from an expert (Yu et al., 2021). Although diverse variations were identified, no discernible pattern emerged to indicate the specific steps chosen or the decision-making process behind them. Consequently, each decision is taken on an individual basis.

It is important to note that this pipeline is not limited to a particular knowledge structure or even restricted solely to the domain of imaging, as similar methods have been observed in other contexts (Holzinger, 2017). Although image preprocessing was consistently present across all papers, particular emphasis was laid on this step in the case of instances involving Algebraic Equations (Chae et al., 2022), Probabilistic Relations (Andrushia et al., 2021), and Spatial Invariances (J. Huang et al., 2021), as evidenced in respective papers. However, no definitive correlation between this method and distinct knowledge structures has been established. One plausible interpretation is that these specific types of knowledge structures are more amenable to integration due to their lack of stringent requirements for image preprocessing. The weight of

the preprocessing steps is therefore laid on the less specialized and more general preprocessing procedures that aim to increase the compatibility of image data with the machine learning algorithm.

Data Augmentation

Data Augmentation is a preprocessing technique commonly utilized to optimize the extraction of features from data by enlarging the size of the dataset through generated data (Maharana et al., 2022). The primary objective of Data Augmentation is to optimize the recognition rate by obtaining a feature set with minimal elements while also generating similar feature sets for diverse instances of the same symbol. Through this approach, image augmentation techniques aim to prevent overfitting, balance class distributions, enforce invariance, and improve generalization to real-world variability. In the case of imaging neural networks, this is most often achieved through the implementation of classical image transformations, such as rotating, cropping, zooming, histogram-based methods, or more advanced techniques, such as the use of Generative Adversarial Networks (Miko-lajczyk & Grochowski, 2018). It is no surprise that Data Augmentation is often employed in the field of medical imaging in conjunction with the integration of domain knowledge, as both methods aim to counteract the scarcity and high cost of medical data (Xie et al., 2020).

Regarding the relevant literature, it is interesting to note that although examined works recognize the shortcomings and limitations of Data Augmentation, considering it insufficient if used in isolation (Yu et al., 2021), several researchers incorporate it in the analysis pipeline as part of the Data Preprocessing step. The wide implementation of Data Augmentation in different contexts and preprocessing pipelines is also enabled through the high transferability and flexibility of this method (Borghesi et al., 2020). In particular, Data Augmentation finds extensive application when the integrated knowledge follows the structure of Logic Rules (Yu et al., 2021), Spatial Invariances (J. Huang et al., 2021), or Probabilistic Relations (Yu et al., 2021). When these categories form a big part of the model's internal logic, the ability of Data Augmentation to reduce the risk of overfitting and increase the model's capacity to generalize for real worlds scenarios becomes even more valuable. This rings particularly true in the case of knowledge in the form of Spatial Invariances, which -per definition describe properties that remain unaltered and should therefore be recognizable even under the executed transformations (von Rueden et al., 2023). This aspect, in conjunction with the general sparsity of medical data, motivates the implementation of Data Augmentation preprocessing steps for the proper integration of domain knowledge.

Knowledge Preprocessing

The second aspect of preprocessing for integrating domain knowledge in machine learning involves transforming the knowledge itself. Beyond ensuring compatibility with the data and algorithm, converting the knowledge into a computer-usable and interpretable format is crucial. The steps and preprocessing approaches identified are listed in Table 6.

Logical Encoding

Logical Encoding refers to the process of transforming domain knowledge into a format that is interpretable by the computer and compatible with the used machine learning algorithm. Although most of the preprocessing steps presented in this paper could be classified under "encoding", as they translate existing knowledge in system-compatible input, the term here refers specifically to adapting already formalized knowledge for the specific scenario. Compared to other preprocessing pipelines, the degree of adaptation is minimal since formalization does not need to be executed beforehand and therefore warrants a separate categorization.

The method of Logical Encoding is mainly implemented when the knowledge is represented in the form of Logic Rules or Spatial Invariances. The review of relevant papers indicates that while this preprocessing step is widely utilized, the exact methods of transformation or adoption of rules are not extensively reported. In many instances, methodology reports primarily describe the transformation of rules in the mathematical or probabilistic models or in if-then statements (Yang et al., 2004; Z. Zhang et al., 2020). In other words, the weight lies in formalizing the knowledge and bringing it to a favorable structure, while the final encoding step is assumed to be executed manually. It is also worth mentioning that several papers emphasize the importance of Logical Encoding and highlight its significance in the context of Architecture Design (Z. Zhang et al., 2020).

Preprocessing Method	Knowledge Structure	Papers
Logical Encoding	Logic Rules	(Yang et al., 2004)
	Spatial Invariances	(Dravid et al., 2022)
		(Lizhi et al., 2021)
		(Fang et al., 2021)
		(Tardy & Mateus, 2021)
Supervised Knowledge Constrains	Spatial Invariances	(Baragry & Reed, 1999)
		(J. Huang et al., 2021)
Loss Function (Scoring Criterion)	Probabilistic Relations	(Yu et al., 2021)
		(Tardy & Mateus, 2021)
Knowledge Transfer	Logic Rules	(Li et al., 2019)

Table 6. Categorization of Knowledge Preprocessing Steps According to the Structure of the Domain Knowledge

Supervised Knowledge Constraints

The objective of knowledge constraints is to confine the extensive search space to mitigate errors and facilitate expedited and efficient training, ultimately converging towards an optimal or satisfactory solution (Cai & Qiu, 2018). In the case of medical imaging AI, Expert Knowledge is used to generate such constraints and limit the hypothesis space for the program during training, improving the efficiency and quality of the results (Cai & Qiu, 2018). Apart from enabling the model to produce more accurate results, regularization, and constraint embedding techniques aim to obtain neural networks whose output respects predefined desirable properties or for which certain relationships between input and output hold (Borghesi et al., 2020). An additional benefit of this approach is that it allows for the model to be trained on datasets that contain instances violating the constraints without the danger of learning biases or false relationships (Borghesi et al., 2020).

Forming knowledge constraints can be perceived as a sub-category of Logical Encoding, as it entails the representation of domain knowledge through a list of variable relationships or is integrated within a posterior regularization framework (Baragry & Reed, 1999; Frank et al., 2022). Knowledge Constraints were most often observed when the available knowledge was represented through Spatial Invariances between objects or specific patterns. The knowledge representations this approach was most often observed with are similar to the ones of Logical Encoding, highlighting the similarities between the two approaches. By doing so, the model can focus on learning relevant features and avoid unnecessary computations.

Loss Function (Scoring Criterion)

Among the examined papers, the incorporation or adaptation of the chosen Loss Function has been identified as another widely implemented step for the integration of domain knowledge in the algorithm. The Loss Function is used as an evaluation criterion in most machine learning systems to define the quality of the algorithm's predictions during training or inference (Hennig & Kutlukaya, 2007). Apart from the general categorization based on the type of machine learning system used (e.g., classification or regression), further adaptation of the Loss Function has been extensively researched, as it is essential for the chosen Loss Function to correspond to the goals of the development or research team (Wang et al., 2022).

In the examined use cases, the emphasis is placed on integrating additional constraints or setting new priorities for the optimization of the Loss Function. This involves modifying the loss term or introducing supplementary loss criteria (X. Huang et al., 2021). Numerous instances have reported the utilization of existing metrics at various stages of the architecture or the creation of novel metrics (X. Huang et al., 2021). In the context of domain knowledge integration, experts establish arbitrary priorities based on available knowledge, which they manually enforce by adapting the Loss Function. This adaptation can take the form of modifying the primary term or introducing additional variables into the function.

An interesting expansion upon the adaptation of the Loss Function is the introduction of a completely new Scoring Criterion. Through this approach, the experts exert their influence on the machine learning algorithm by dictating exactly how the training outcomes are to be evaluated.

Due to the direct influence the scoring criterion has on the effectiveness of the system, it must be derived from clinically established knowledge provided by human experts to ensure relevance and correctness (Luo et al., 2021). Consequently, experts possess the capability to modify or devise the metric (or a combination of metrics) used to assess and enhance the quality of the program.

Supervised Knowledge Transfer

Transfer learning is the process of "transferring" knowledge acquired from one source task or domain to a different target task, either in the same or in a new domain (Rios & Kavuluru, 2019). The process of reutilizing previously acquired knowledge from one training domain is also referred to as deep adaptation of domain knowledge through Supervised Knowledge Transfer. Supervised Knowledge Transfer is a prominent preprocessing step, primarily in the context of Knowledge Graphs (Sekuboyina et al., 2021).

In medical imaging, Supervised Knowledge Transfer was most often implemented when knowledge was available from different domains and could partially be applied in the case of medical imaging algorithms. It relied heavily on the input of experts both to identify fitting fields from which knowledge could be leveraged, as well as during the knowledge adaptation process (Ye et al., 2020). The reliance on expert input for the preprocessing of the source knowledge is tied to the need for accuracy and correctness during the preparation and the application of existing equations or constraints (Ye et al., 2020; Y. Zhang et al., 2020).

Knowledge Enhanced Data Preprocessing

In specific cases, Data Preprocessing and Knowledge Preprocessing coincide. This occurs when experts directly engage with the training tata in order to achieve a more effective preprocessing and boost the training effectiveness and subsequent performance of the model. Table 7 provides a categorization of the steps included in Knowledge Enhanced Data Preprocessing.

Region of Interest (ROI) Segmentation

ROI Segmentation is a crucial preprocessing step aimed at extracting the pertinent region of interest from an image, enabling the learning of useful features while conserving computational resources by excluding the background. This segmentation is achieved by creating a numeric mask, a numeric image of the same size as the input, where pixels in the ROI are set to 1 and all other pixels to 0 (Chai et al., 2018).

The segmentation step is commonly executed either by a model (Olender et al., 2020; Leung et al., 2020) (Chai et al., 2018) or manually performed by domain experts (Rajaraman et al., 2021). Alternatively, a combination of model execution and expert validation can be employed (Zheng et al., 2019).

ROI Segmentation is an essential step during the training of imaging models and is, therefore, not constricted to a particular knowledge structure. This method was implemented in the cases of knowledge in the form of Algebraic Equations (Olender et al., 2020), Logic Rules (Costin & Rotariu, 2011), and Spatial Invariances (Chai et al., 2018). In these scenarios, knowledge was often employed to boost the segmentation accuracy of the model and improve the performance of automated ROI Segmentation. In the case of Human Feedback integration into the system, using the experts' input during the segmentation step was shown to be an effective and efficient approach (Zheng et al., 2019). ROI Segmentation is, therefore, not a direct and exclusive prerequisite for integrating knowledge from the aforementioned sources but rather a general requirement for effective training of CNN networks that can be greatly enhanced through leveraging domain knowledge.

Labeling

The majority of the examined imaging machine-learning applications primarily focused on the classification of generated images and the detection of visual patterns. As with all classification systems, imaging AI systems, therefore, fall under the category of supervised learning, meaning that they rely on a set of pre-labeled images to learn the features they must use for the classification of new input (de Siqueira

et al., 2018). In medical imaging, the importance of accurate and precise Labeling is even higher, and expert input is often required during this process (Rajaraman et al., 2021).

Although almost all of the examined papers relied either on a Labeling step or on a set of prelabeled images, this method was stressed the most when the domain knowledge was represented in the form of Human Feedback (Luo et al., 2021). This correlation does not necessarily signal that labeled data is necessary to incorporate Human Feedback in a machine-learning system but could rather indicate that image Labeling or annotation is an effective method to integrate Expert Knowledge into the data analysis pipeline. Although this process may induce additional costs, these expenses are frequently mitigated through other domain knowledge sources or the augmentation of pre-existing labeled images (Xie et al., 2020). Manual Labeling, therefore, plays a pivotal role in enhancing the model's performance (Ye et al., 2020), thereby positioning Labeling within the framework of the Human Feedback knowledge structure.

Preprocessing Method	**Knowledge Structure**	**Papers**
ROI Segmentation	Algebraic Equations	(Olender et al., 2020)
		(Bian et al., 2022)
	Human Feedback	(J. Huang et al., 2021)
	Spatial Invariances	(Chai et al., 2018)
		(Leung et al., 2020)
		(Rajaraman et al., 2021)
	Logic Rules	(Costin & Rotariu, 2011)
		(Hemelings et al., 2020)
Labeling	Human Feedback	(Luo et al., 2021)
		(Zheng et al., 2019)
Feature Extraction	Human Feedback	(Baâzaoui et al., 2017)
		(Zhao et al., 2022)
	Knowledge Graphs	(Yu et al., 2021)
		(Velikova et al., 2013)
	Spatial Invariances	(Lian et al., 2021)
		(Fang et al., 2021)
		(Dravid et al., 2022)

Table 7. Categorization of Knowledge Enhanced Data Preprocessing Steps According to the Structure of Domain Knowledge

Feature Extraction

Feature Extraction is a fundamental process in machine-learning models aimed at extracting the most relevant information from raw data. This involves retrieving essential features from objects, which are subsequently used to construct feature vectors. These feature vectors serve as input for classifiers, enabling the recognition of target output units (Kumar & Bhatia, 2014). By employing Feature Extraction, a reduced set of elements can maximize the recognition rate while facilitating the generation of similar feature sets for various instances of the same symbol.

In the examined literature, Feature Extraction was often conducted by computers or existing artificial intelligence models, sometimes in conjunction with handcrafted features (Guo et al., 2019, Frank et al., 2022). Enhancing and enriching the features with additional characteristics was found to significantly improve the effectiveness of the models. In such cases, experts were called upon to identify or extract relevant features. For this reason, although automated Feature Extraction was observed, human input was found to be an important aspect in preprocessing the input data (Guo et al., 2019).

Feature Extraction was found to be particularly relevant in the case of a graph representation of knowledge (Knowledge Graphs), as the chosen features play a vital role in strengthening the connections between graph nodes during training (Wen et al., 2020). In the examined scenarios, the addition of handcrafted and expert-identified features was therefore found to enable models to outperform approaches that are restricted to input images and fully automated Feature Extraction.

Discussion

Principal Findings

Reflecting on the executed review and categorization of examined approaches for the integration of domain knowledge in biomedical imaging AI algorithms has yielded vital insights into the overall process and the current state of implemented methodologies. In the following, the principal findings regarding Data and Knowledge Preprocessing, and Architecture Design will be presented.

Limited Costs Associated with Data Preprocessing

One of the main objectives of this executed review was to evaluate additional and specialized preprocessing steps required to integrate domain knowledge in imaging machine learning algorithms. In the majority of the reviewed papers, the identified preprocessing steps for the input data were not found to deviate significantly from those of the approaches implemented in general AI imaging algorithms and CNNs. Furthermore, many of the techniques presented in the reviewed papers are already part of typical and standardized Image Preprocessing Pipelines, commonly employed in image recognition, classification, or segmentation applications to enhance computer performance.

Although a number of papers have delved into the adaption or combination of preprocessing measures to properly prepare the input data, the vast majority of the examined literature instead stresses the preprocessing required for the knowledge itself in order to make it interpretable by the computer and compatible with the system architecture. Even when the standardized steps are replaced with more advanced versions (as seen in the section “Knowledge Enhanced Data Preprocessing”), the added degree of complexity mainly revolves around injecting the expert input into the system instead of transforming the training and test data.

Interconnection of Data and Knowledge Preprocessing

The findings of this paper shed light on the interconnectedness between Data and Knowledge Preprocessing. Initially, it was expected that preprocessing steps for training data and domain knowledge could be clearly separated. However, the outcomes of this study reveal that these preprocessing steps often coincide, as domain experts frequently offer direct input during Image Preprocessing. This finding motivated the creation of a third category next to “Data Preprocessing” and “Knowledge Preprocessing”. The “Knowledge Enhanced Data Preprocessing” includes techniques for utilizing domain knowledge to improve the preprocessing of input data. In order to utilize expert input, the methods that fall under this category typically include steps to formalize the available and ensure its compatibility with the overall system. In other words, this new category combines the two formerly introduced preprocessing pipelines and eliminates the need for separate preprocessing of data or knowledge in the examined scenarios. Knowledge enhanced Data Preprocessing was most often achieved through expert labeling data, aiding in the process of ROI Segmentation, or optimizing the Feature Extraction process. Although these techniques can also be executed without expert input (as seen in typical AI systems), the relevant publications noted an increase in accuracy and model performance when the preprocessing pipeline was adapted to include domain knowledge.

Significance of Architecture Design

Although a compatible Architecture Design was found to be one of the three central prerequisites for proper integration of domain knowledge in imaging machine learning systems (along with Data Preprocessing and Knowledge Preprocessing), it was not examined separately, as it lies outside the scope of this thesis. Interestingly, during the literature review, Architecture Design was found to be a significant part of the integration process and was, in many cases, explored in more detail than Data or Knowledge Preprocessing.

Additionally, adjusting or completely redesigning the system architecture was not found to follow standardized approaches or templates and therefore requires a higher level of technical expertise. The added degree of complexity and individualization in system Architecture Design compared to the other preprocessing steps could potentially be a more significant technical bottleneck in future use cases.

Implications

Enhanced Data Preprocessing Efforts

The limited additional costs associated with Data Preprocessing to enable domain knowledge integration in imaging AI systems present interesting and valuable implications for the future of informed machine learning. The minimal degree of necessary adaptation shows that Data Preprocessing is not a bottleneck in the wider use of knowledge-enhanced machine learning systems and will, therefore, not pose an obstacle to their implementation. This becomes even more evident when the tangible benefits of Expert Knowledge integration, which were explored in previous research, are considered. In other words, this review concludes that the costs associated with Data Preprocessing to ensure compatibility with domain knowledge in no way overshadow the positive impact doing so can have on the accuracy and efficiency of the system. As the results have pointed out, the focus should instead lay on the preprocessing of the knowledge itself and the adaptation of the overall system architecture. Examining this aspect, in particular, will allow for a more precise estimation of the degree of complexity and the costs associated with Expert Knowledge-enhanced systems compared to traditional machine-learning approaches.

Holistic Consideration of Preprocessing

The categorization in this paper demonstrated that a clear distinction between Data and Knowledge Preprocessing is, in many cases, either impossible or impractical, as the expert (or the knowledge source) directly interacts with the data during preprocessing to implement an informed machine learning approach. Consequently, instead of using the originally proposed categorization of preprocessing steps for data and knowledge, it is advantageous to incorporate a category that encompasses expert-assisted Data Preprocessing to accommodate such instances comprehensively. Implementing a categorization system that includes knowledge-enhanced Data Preprocessing should enable a deeper understanding of potential bottlenecks and requirements for informed machine learning. The role of the domain expert also becomes clearer when such scenarios are considered. As he is often not responsible solely for providing the knowledge but also directly interacts with the algorithm during design or training, additional tasks may need to be integrated into the process, and further competencies become relevant in corresponding use cases.

Significance of Architecture Design

To effectively integrate domain knowledge at specific stages of the machine learning process, results drawn from the reviewed literature strongly advocate directing increased attention toward the design of the architecture itself. A well-structured and purposeful Architecture Design facilitates the seamless integration of domain-specific insights, ultimately leading to improved model performance and interpretability. This consideration should also allow for a more accurate and realistic estimation of the manual and technical costs associated with implementing informed machine learning in the area of biomedical imaging.

Limitations

The literature review yielded interesting findings with implications that are relevant, both for future research, as well as for real-world application. However, certain significant limitations were identified, which should be taken into consideration when evaluating the generated results or considering approaches for future research. These limitations mainly revolve around the amount of available information the drawn conclusion was based on and will be explained below.

Quantity of Examined Papers

One of the core limitations of this study is the limited number of papers that were used for analysis. Although general trends could be identified among the examined literature, the quantity of provided

database does not allow for the development of a robust and detailed taxonomy of the preprocessing requirements. Due to this issue, interesting insights research could not be sufficiently validated in the scope of this thesis. For example, additional methods to visualize the images and data were found to be an important step when integrating Human Feedback into the machine learning system. This is a valuable and important conclusion that partially differentiates the preprocessing steps in informed machine learning systems from the standardized CNN preprocessing pipeline. However, the limited number of papers explicitly discussing this approach does not allow for it to be considered as an additional category. With the rising popularity of informed machine learning and the benefits of knowledge-enhanced approaches being made clear in existing research, this limitation is expected to be overcome in the future as the integration of domain knowledge in AI systems gains traction. However, it should still be taken into account when considering the definitiveness of the results generated through this literature review.

Lack of Reporting

The second aspect that challenges the accuracy of the results is the oftentimes lack of information on the exact preprocessing steps executed in each scenario. This issue was particularly evident in the case of the "Image Preprocessing Pipeline", a Data Preprocessing approach presented in the section "Data Preprocessing". This term encompasses numerous variations through the combination of different steps. The exact realization of this approach is, therefore, often dependent on the specific use case. Furthermore, the exact order and execution of steps are not always addressed, meaning that resulting categorization only offers an overview of the different methods without specifying steps to adapt each approach to the specific scenario. Lastly, the lack of reporting in certain publications makes it reasonable to presume that preprocessing steps have, in many cases, not been adequately reported or have been summarized in the term "Preprocessing Pipeline". This limitation has not prevented the creation of a general categorization of the preprocessing linked to the integration of domain knowledge in medical imaging AI systems or the examination of connections between the introduced preprocessing methods and the characteristics of the available knowledge. However, it does constrain the precision and level of detail that can be derived from the examined literature and lays the weight of adapting the general categorization to specific scenarios to the development team in future use cases.

Future Research

The categorization generated through the executed literature review paves the way for further research that leverages and enriches the presented findings. Future approaches could either address the limitations presented above or expand upon the principal findings. Certain promising approaches for further consideration are presented below.

Examination of Pathways and Creation of Complete Taxonomy Through Additional Papers

Directly tied to the first limitation of the implemented approach, an expansion upon the original work with the purpose of featuring additional papers could enable a more comprehensive categorization of the available literature and lead to the development of a taxonomy of preprocessing methods for the integration of domain knowledge in machine learning applications. Additional information would also enable the examination of the preprocessing requirements according to the "paths" linking different characteristics of domain knowledge, such as its source and step of application, instead of only focusing on the representation aspect.

Examination of Architecture Adaptation for Domain Knowledge

As introduced in the "Principal Findings" section, (re-)designing the architecture of a machine learning system to accommodate for the information provided by domain knowledge is often responsible for a considerable portion of the additional work associated with this task. The importance of Architecture Design could motivate a reevaluation of the examined literature with greater focus laid on the software architecture requirements and adaptation methods in future research. One possible outcome of such an approach would be the creation of a categorization of design options for knowledge-enhanced machine learning architecture. Alternatively, the complexity of this aspect could warrant a different review approach compared to the one introduced in the methodology section of this work.

A Reporting Framework

The second major limitation identified during the course of this research was the lack of sufficient or homogenous reporting when it comes to preprocessing information used in machine learning systems. Although some papers presented their methodology in sufficient detail, this aspect of the system was often overlooked or briefly introduced. The lack of reporting in such cases raises issues in works of literature review with the objectives of comparison or categorization of different approaches, such as this one. Motivated by this shortcoming of current research, introducing a reporting standard for the preparation and execution of machine learning systems and similar technical tasks would prove useful in future research. Implementing such guidelines would enable a wider and more direct comparison and clustering of executed approaches while shielding review research from misinterpretations and errors. Finally, standardized reporting would allow research teams to significantly accelerate, and scale literature review works by only considering the relevant aspects of this process instead of extracting such information manually from each publication.

Examination of Additional Medical Applications

Apart from examining additional papers in the field of medical imaging, it would also be of interest to expand on the presented work by integrating further medical fields into future research. Aspects of the examined approaches are expected to differ depending on the exact field and application. For example, the basic preprocessing pipeline will definitely be different when working with images compared to other measurements or data formats. However, similarities are expected to be present amongst fields, particularly when it comes to preprocessing of the available knowledge. Exploring and evaluating such assumptions could enable a deeper understanding of the nature of informed machine learning in the field of medicine and enable the development of more holistic and effective approaches. For this reason, expanding the scope of this work to incorporate medical fields that have not yet been addressed could prove beneficial by providing both a wider and a deeper understanding of domain knowledge integration in medical machine learning systems.

Conclusion

Utilizing machine learning algorithms for the analysis and evaluation of medical images has the potential to alleviate the burden on healthcare practitioners. To tackle the issue of limited data in machine learning, the integration of domain knowledge has demonstrated its advantages. This research undertakes a systematic literature review of 80 studies to explore the preprocessing requirements for the implementation of informed machine-learning approaches in medical imaging. Consequently, a comprehensive overview of the identified preprocessing steps is developed.

The results of this research were organized under Data Preprocessing and Knowledge Preprocessing. The former includes techniques aimed at ensuring compatibility of the input data with the knowledge and the machine learning algorithm and included methods such as implementing an Image Preprocessing Pipeline and Data Augmentation. The latter encompasses methods to standardize and formalize Domain Knowledge and transform it into a representation that is compatible with and interpretable by the system. Logical Encoding, Supervised Knowledge Constraints, Loss Function adaptation, and Knowledge Transfer were all classified under Knowledge Preprocessing. Through the evaluation of the results, a third category, knowledge-enhanced Data Preprocessing, was established. This new categorization was deemed necessary to include methods of preprocessing the data through direct input from the expert. Such methods include ROI Segmentation, Labeling, and Feature Extraction. Identified trends between the preprocessing approaches and the different structures of domain knowledge were identified and explained.

Reflecting on the results of the literature review yielded interesting findings regarding the current landscape of informed machine-learning approaches. The Data Preprocessing methods were found to not deviate significantly from typical Data Preprocessing pipelines for image classification algorithms. The focus should therefore be laid on preprocessing the domain knowledge and adapting Architecture Design, which was found to be an important prerequisite despite lying outside the scope of this thesis. Finally, the limitations of the executed approach were discussed, mainly revolving around the availability of literature and preprocessing reporting, and potential approaches for further research were proposed, either tackling the presented limitations or expanding upon the generated findings.

References

Andrushia, A. D., Sagayam, K. M., Dang, H., Pomplun, M., & Quach, L. (2021). Visual-saliency-based abnormality detection for mri brain images—alzheimer's disease analysis. *Applied Sciences*, *11* (19). **https://doi.org/10.3390/app11199199**

Ba âzaoui, A., Barhoumi, W., & Zagrouba, E. (2017). Towards semantic visual features for malignancy description within medical images. *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 397–402. **https://doi.org/10.1109/ICCP.2017.8117037**

Baragry, J., & Reed, K. (1999). Why is it so hard to define software architecture?. *Proceedings 1998 Asia Pacific Software Engineering Conference (Cat. No.98EX240)*, 28–36. **https://doi.org/10.1109/APSEC.1998.733577**

Bian, X., Pan, H., Zhang, K., Li, P., Li, J., & Chen, C. (2022). Skin lesion image classification method based on extension theory and deep learning. *Multimedia Tools and Applications*, *81*, 1–21. **https://doi.org/10.1007/s11042-022-12376-3**

Borghesi, A., Baldo, F., & Milano, M. (2020). Improving deep learning models via constraint-based domain knowledge: a brief survey. *arXiv preprint arXiv:2005.10691*. **https://doi.org/10.48550/arXiv.2005.10691**

Cai, J., & Qiu, X. (2018). Constrained partial fuzzy clustering for brain magnetic resonance image segmentation. *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, 115–118. **https://doi.org/10.1109/ITME.2018.00035**

Chae, J., Zhang, Y., Zimmermann, R., Kim, D., & Kim, J. (2022). An attention-based deep learning model with interpretable patch-weight sharing for diagnosing cervical dysplasia. In K. Arai (Ed.), *Intelligent systems and applications,* 634–642. Springer International Publishing. **https://doi.org/10.1007/978-3-030-82199-9_43**

Chai, Y., Liu, H., & Xu, J. (2018). Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowledge-Based Systems*, *161*, 147–156. **https://doi.org/https://doi.org/10.1016/j.knosys.2018.07.043**

Cheng, P., Lin, L., Huang, Y., Lyu, J., & Tang, X. (2021). Prior guided fundus image quality enhancement via contrastive learning, 521–525. **https://doi.org/10.1109/ISBI48211.2021.9434005**

Costin, H., & Rotariu, C. (2011). Medical image processing by using soft computing methods and information fusion. *Proceedings of the 13th WSEAS International Conference on Mathe- matical Methods, Computational Techniques and Intelligent Systems, and 10th WSEAS International Conference on Non-Linear Analysis, Non-Linear Systems and Chaos, and 7th WSEAS International Conference on Dynamical Systems and Control, and 11th WSEAS International Conference on Wavelet Analysis and Multirate Systems: Recent Researches in Computational Techniques, Non-Linear Systems and Control*, 182–191. **https://doi.org/ 10.5555/2039846.2039879**

de Siqueira, G. O., Canuto, S. D., Gon ̧calves, M. A., & Laender, A. H. F. (2018). A pragmatic approach to hierarchical categorization of research expertise in the presence of scarce information. *International Journal on Digital Libraries*, *21*, 61–73. **https://api.semanticscholar.org/CorpusID:69837310**

Dravid, A., Schiffers, F., Gong, B., & Katsaggelos, A. K. (2022). medxgan: Visual explanations for medical classifiers through a generative latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2936-2945. **https://doi.org/ 10.1109/CVPRW56347.2022.00331**

Famili, F., Shen, W.-M., Weber, R., & Simoudis, E. (1997). Data preprocessing and intelligent data analysis. *Intell. Data Anal.*, *1*, 3–23. **https://doi.org/ 10.3233/IDA-1997-1102**

Fang, J., Xu, Y., Zhao, Y., Yan, Y., Liu, J., & Liu, J. (2021). *Weighing features of lung and heart regions for thoracic disease classification*. **https://doi.org/10.1186/s12880-021-00627-y**

Frank, O., Schipper, N., Vaturi, M., Soldati, G., Smargiassi, A., Inchingolo, R., Torri, E., Perrone, T., Mento, F., Demi, L., Galun, M., Eldar, Y. C., & Bagon, S. (2022). Integrating domain knowledge into deep networks for lung ultrasound with applications to covid-19. *IEEE Transactions on Medical Imaging*, *41* (3), 571–581. **https://doi.org/10.1109/TMI.2021.3117246**

Gou, C., Shen, T., Zheng, W., Xue, H., Yu, H., Ji, Q., ... & Wang, F. Y. (2019). Parallel Medical Imaging for Intelligent Medical Image Analysis: Concepts, Methods, and Applications. *arXiv preprint arXiv:1903.04855*. **https://doi.org/10.48550/arXiv.1903.04855**

Guo, X., Yang, C., Lam, P. L., Woo, P. Y., & Yuan, Y. (2020). Domain knowledge based brain tumor segmentation and overall survival prediction. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II 5*, 285-295. Springer International Publishing. **https://doi.org/10.1007/978-3-030-46643-5_28**

Hemelings, R., Elen, B., Blaschko, M., Jacob, J., Stalmans, I., & De Boever, P. (2020). Patho- logical myopia classification with simultaneous lesion segmentation using deep learning. *Computer Methods and Programs in Biomedicine*, *199*, 105920. **https://doi.org/10.1016/j.cmpb.2020.105920**

Hennig, C., & Kutlukaya, M. (2007). Some thoughts about the design of loss functions. *REVSTAT-Statistical Journal*, *5* (1), 19–39. **https://doi.org/10.57805/revstat.v5i1.40**

Holzinger, A. (2017). Big data calls for machine learning. **https://doi.org/10.1016/B978-0-12-801238-3.10877-3**

Huang, J., Yan, H., Li, J., Stewart, H., & Setzer, F. (2021). Combining anatomical constraints and deep learning for 3-d cbct dental image multi-label segmentation, 2750–2755. **https://doi.org/10.1109/ICDE51399.2021.00319**

Huang, X., Yue, X., Xu, Z., & Chen, Y. (2021). Integrating general and specific priors into deep convolutional neural networks for bladder tumor segmentation. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. **https://doi.org/10.1109/IJCNN52387.2021.9533813**

Kumar, G., & Bhatia, P. K. (2014). A detailed review of feature extraction in image processing systems. *2014 Fourth International Conference on Advanced Computing Communication Technologies*, 5–12. **https://doi.org/10.1109/ACCT.2014.74**

Kwee, T. C., & Kwee, R. M. (2021). Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: Growth expectations and role of artificial intelligence. *Insights into Imaging*, *12*. **https://doi.org/10.1186/s13244-021-01031-4**

Lee, J., & He, Q. P. (2019). Understanding the effect of specialization on hospital performance through knowledge-guided machine learning. *Comput. Chem. Eng.*, *125*, 490–498. **https://doi.org/10.1016/j.compchemeng.2019.03.040**

Leung, K. H., Marashdeh, W., Wray, R., Ashrafinia, S., Pomper, M. G., Rahmim, A., & Jha,

A. K. (2020). A physics-guided modular deep-learning based automated framework for tumor segmentation in PET. *Physics in Medicine &amp Biology*, *65* (24), 245032. **https://doi.org/10.1088/1361-6560/ab8535**

Li, Z., Zhang, S., Zhang, J., Huang, K., Wang, Y., & Yu, Y. (2019). MVP-Net: multi-view FPN with position-aware attention for deep universal lesion detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22* (pp. 13-21). Springer International Publishing. **https://doi.org/10.1007/978-3-030-32226-7_2**

Lian, J., Liu, J., Zhang, S., Gao, K., Liu, X., Zhang, D., & Yu, Y. (2021). A structure-aware relation network for thoracic diseases detection and segmentation. **https://doi.org/10.1109/TMI.2021.3070847**

Lizhi, S., Liu, Z., Yan, Y., Liu, J., Ye, X., Xia, H., Zhu, X., Zhang, Y., Zhang, Z., Chen, H., He, W., Liu, C., Lu, M., Huang, Y., Sun, K., Zhou, X., Yang, G., Lu, J., & Tian, J. (2021). Patient-level prediction of multi-classification task at prostate mri based on end- to-end framework learning from diagnostic logic of radiologists. *IEEE Transactions on Biomedical Engineering*, 1–1. **https://doi.org/10.1109/TBME.2021.3082176**

Luo, J., Kitamura, G., Doganay, E., Arefan, D., & Wu, S. (2021). Medical knowledge-guided deep curriculum learning for elbow fracture diagnosis from x-ray images. In K. Drukker & M. A. Mazurowski (Eds.), *Medical imaging 2021: Computer-aided diagnosis*. SPIE. **https://doi.org/10.1117/12.2582184**

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques [International Conference on Intelligent Engineering Approach(ICIEA-2022)]. *Global Transitions Proceedings*, *3* (1), 91–99. **https://doi.org/https://doi.org/10.1016/j.gltp.2022.04.020**

Meijering, E. (2020). A bird's-eye view of deep learning in bioimage analysis. *Computational and Structural Biotechnology Journal*, *18*, 2312–2325. **https://doi.org/10.1016/j.csbj.2020.08.003**

Miko-lajczyk, A., & Grochowski, M. (2018). Data augmentation for improving deep learning in im- age classification problem. *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 117–122. **https://doi.org/10.1109/IIPHDW.2018.8388338**

Nguyen, D., Truong, M., Than, N., Prange, A., & Sonntag, D. (2021). *Self-supervised domain adaptation for diabetic retinopathy grading using vessel image reconstruction*.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*, 336–359. **https://doi.org/10.1057/ejis.2012.26**

Nomani, A., Ansari, Y., Nasirpour, M. H., Masoumian, A., Pour, E. S., & Valizadeh, A. (2022). Psownns-cnn: A computational radiology for breast cancer diagnosis improvement based on image processing using machine learning methods. *Computational Intelligence and Neuroscience*, *2022*. **https://doi.org/10.1155/2022/5667264**

Olender, M. L., Athanasiou, L. S., Michalis, L. K., Fotiadis, D. I., & Edelman, E. R. (2020). A domain enriched deep learning approach to classify atherosclerosis using intravascular ultrasound imaging. *IEEE Journal of Selected Topics in Signal Processing*, *14* (6), 1210– 1220. **https://doi.org/10.1109/JSTSP.2020.3002385**

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., & Pattichis, C. S. (2020). Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, *24* (7), 1837–1857. **https://doi.org/10.1109/JBHI.2020.2991043**

Pezzotti, N., Yousefi, S., Elmahdy, M. S., van Gemert, J., Schü lke, C., Doneva, M., Nielsen, T., Kastryulin, S., Lelieveldt, B. P. F., van Osch, M. J. P., de Weerdt, E., & Staring, M. (2020). An adaptive intelligence algorithm for undersampled knee mri reconstruction. **https://doi.org/10.1109/ACCESS.2020.3034287**

Qiu, T. (2020). Tongue identification for small samples based on meta learning, 295–299. **https://doi.org/10.1109/CIBDA50819.2020.00073**

Qoku, A., & Buettner, F. (2022). *Encoding domain knowledge in multi-view latent variable mod- els: A bayesian approach with structured sparsity*. **https://doi.org/10.48550/arXiv.2204.06242**

Rajaraman, S., Folio, L. R., Dimperio, J., Alderson, P. O., & Antani, S. K. (2021). Improved semantic segmentation of tuberculosis—consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations. *Diagnostics*, *11* (4). **https://doi.org/10.3390/diagnostics11040616**

Rios, A., & Kavuluru, R. (2019). Neural transfer learning for assigning diagnosis codes to EMRs. *Artificial intelligence in medicine*, *96*, 116–122. **https://doi.org/10.1016/j.artmed.2019.04.002**

Roth, H. R., Yang, D., Xu, Z., Wang, X., & Xu, D. (2020). Going to extremes: Weakly supervised medical image segmentation. *Mach. Learn. Knowl. Extr.*, *3*, 507–524. **https://doi.org/10.3390/make3020026**

Sekuboyina, A., Oñoro-Rubio, D., Kleesiek, J., & Malone, B. (2021). A relational-learning perspective to multi-label chest x-ray classification. **https://doi.org/10.1109/ISBI48211.2021.9433786**

Stoitsis, J. S., Valavanis, I. K., Mougiakakou, S. G., Golemati, S., Nikita, A., & Nikita, K. S. (2006). Computer aided diagnosis based on medical image processing and artificial in- telligence methods. *Nuclear Instruments & Methods in Physics Research Section A- accelerators Spectrometers Detectors and Associated Equipment*, *569*, 591–595. **https://doi.org/10.1016/j.nima.2006.08.**

Tang, Z., Chen, Z., Bao, Y., & Li, H. (2018). Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring. *Structural Control and Health Monitoring*, *26*. **https://doi.org/10.1002/stc.2296**

Tardy, M., & Mateus, D. (2021). Looking for abnormalities in mammograms with self- and weakly supervised reconstruction. *IEEE Transactions on Medical Imaging*, *PP*, 1–1. **https://doi.org/10.1109/TMI.2021.3050040**

Tian, Y., Liu, F., Pang, G., Chen, Y., Liu, Y., Verjans, J., Singh, R., & Carneiro, G. (2021). Self-supervised multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. **https://doi.org/10.1016/j.media.2023.102930**

Väänänen, A., Haataja, K., Vehviläinen-Julkunen, K., & Toivanen, P. J. (2021). Ai in healthcare: A narrative review. *F1000Research*, *10*, 6. **https://doi.org/10.12688/f1000research.26997.1**

Velikova, M., Lucas, P. J., Samulski, M., & Karssemeijer, N. (2013). On the interplay of machine learning and background knowledge in image interpretation by bayesian networks. *Artificial intelligence in medicine*, *57*. **https://doi.org/10.1016/j.artmed.2012.12.004**

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., & Schuecker, J. (2023). Informed machine learning – a taxonomy and survey of integrat- ing prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 614–633. **https://doi.org/10.1109/TKDE.2021.3079836**

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, *9*. **https://doi.org/10.1007/s40745-020-00253- 5**

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26* (2), xiii–xxiii. **https://doi.org/10.5555/2017160.2017162**

Wen, G., Ma, J., Hu, Y., Li, H., & Jiang, L. (2020). Grouping attributes zero-shot learning for tongue constitution recognition. *Artificial Intelligence in Medicine*, *109*, 101951. **https://doi.org/https://doi.org/10.1016/j.artmed.2020.101951**

Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., & Yu, S. (2020). A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical image analysis*, *69*, 101985. **https://doi.org/10.1016/j.media.2021.101985**

Xie, X., Niu, J., Liu, X., Li, Q., Wang, Y., & Tang, S. (2021). Dk-consistency: A domain knowl- edge guided consistency regularization method for semi-supervised breast cancer diagnosis, 3435–3442. **https://doi.org/10.1109/BIBM52615.2021.9669494**

Xu, Z., Liu, X., & Zhang, K. (2019). Mechanical properties prediction for hot rolled alloy steel using convolutional neural network. *IEEE Access*, *7*, 47068–47078. **https://doi.org/10.1109/ACCESS.2019.2909586**

Yaman, B., Shenoy, C., Deng, Z., Moeller, S., El-Rewaidy, H., Nezafat, R., & Ak¸cakaya, M. (2021). Self-supervised physics-guided deep learning reconstruction for high-resolution 3d lge cmr, 100–104. **https://doi.org/10.1109/ISBI48211.2021.9434054**

Yang, Y., Chen, S., Lin, H., & Ye, Y. (2004). A chromatic image understanding system for lung cancer cell identification based on fuzzy knowledge. In: Orchard, B., Yang, C., Ali, M. (eds), *Innovations in applied artificial intelligence*. 392–401. Springer Berlin Heidelberg. **https://doi.org/10.1007/978-3-540-24677-0_41**

Ye, Y., Mao, J., Liu, L., Zhang, S., Shen, L., & Sun, M. (2020). Automatic diagnosis of familial exudative vitreoretinopathy using a fusion neural network for wide-angle retinal images. *IEEE Access*, *8*, 162–173. **https://doi.org/10.1109/ACCESS.2019.2961418**

Yoon, Y.-m., Hwang, T., Choi, H., & Lee, H. (2019). Classification of radiographic lung pattern based on texture analysis and machine learning. *Journal of Veterinary Science*, *20*. **https:// doi.org/ 10.4142/jvs.2019.20.e44**

Yu, W., Zhou, H., Goldin, J., Wong, W., & Kim, G. H. (2021). End-to-end domain knowl- edge assisted automatic diagnosis of idiopathic pulmonary fibrosis (ipf) using computed tomography (ct). *Medical Physics*, *48*. **https://doi.org/10.1002/mp.14754**

Zhang, Y., Niu, S., Qiu, Z., Wei, Y., Zhao, P., Yao, J., Huang, J., Wu, Q., & Tan, M. (2020). Covid-da: Deep domain adaptation from typical pneumonia to covid-19. arXiv *preprint arXiv:2005.01577*. **https://doi.org/10.48550/arXiv.2005.01577**

Zhang, Z., Yu, S., Qin, W., Liang, X., Xie, Y., & Cao, G. (2020). Ct super resolution via zero shot learning. *arXiv preprint arXiv:2012.08943*. **https://doi.org/10.48550/arXiv.2012.08943**

Zhao, S.-X., Chen, Y., Yang, K.-F., Luo, Y., Ma, B.-Y., & Li, Y.-J. (2022). A local and global feature disentangled network: Toward classification of benign-malignant thyroid nodules from ultrasound image. *IEEE Transactions on Medical Imaging*, *41* (6), 1497–1509. **https://doi.org/10.1109/TMI.2022.3140797**

Zheng, H., Chen, Y., Yue, X., & Ma, C. (2019). Deep interactive segmentation of uncertain regions with shadowed sets. *ISICDM 2019: Proceedings of the Third International Sym- posium on Image Computing and Digital Medicine*, 244–248. **https://doi.org/10.1145/3364836.3364885**

Zhou, J., Li, Z., Zhi, W., Liang, B., Moses, D., & Dawes, L. (2017). Using convolutional neural networks and transfer learning for bone age classification, 1–6. **https://doi.org/10.1109/DICTA.2017.8227503**

Initial Orchestration of Digital Platform-Based Ecosystems

Emerging Trends in Internet Technologies, Summer Term 2023

Lukas Braun
Master Student
Karlsruhe Institute of Technology
lukas.braun3@student.kit.edu

Felix Häusle
Master Student
Karlsruhe Institute of Technology
felix.haeusle@student.kit.edu

Artur Romanenko
Master Student
Karlsruhe Institute of Technology
a.romanenko@web.de

Leo Schorling
Bachelor Student
Karlsruhe Institute of Technology
leofschorling@icloud.com

Abstract

Background: *Technological giants such as Alphabet, Apple, and Microsoft have fundamentally transformed the corporate landscape by developing digital platform-based ecosystems. The concept of creating not just a platform but an entire ecosystem has allowed these companies to incorporate numerous complementors, thereby generating significant added value for customers and other ecosystem participants.*

Objective: *Despite the success of this collaborative model and the increasing adoption of similar business models by other companies, there is still a lack of understanding regarding the orchestration approaches during the initial phase of a digital platform-based ecosystem. This research aims to explore the orchestration strategies that contribute to value co-creation in the early stages of these ecosystems.*

Methods: *A systematic literature review was conducted to investigate the orchestration, value creation, and technical aspects of digital platform-based ecosystems. The analysis was based on 17 thematic fields derived from coding relevant literature.*

Results: *The findings highlight that a common vision among participants, modular implementation of boundary resources, and clear role distribution are crucial for the success of the ecosystem. These elements are essential for effectively coordinating the diverse interests and contributions of the actors in the ecosystem.*

Conclusion: *The study provides a structured overview of different orchestration approaches in the initial phase of digital platform-based ecosystems. It underlines the importance of shared vision, modular resources, and role clarity in fostering successful ecosystems. Future research should continue to refine these insights and explore their applicability in various contexts.*

Keywords: orchestration approaches, digital platform-based ecosystems, value co-creation, initial phase

Einleitung

Problemstellung

Technologische Innovationen und weitere Veränderungstreiber rufen einen Wandel in der heutigen Geschäftswelt hervor und stellen Unternehmen vor die Herausforderung der „digitalen Revolution“ (Birkinshaw, 2018, S. 187). Dadurch entstehen völlig neue Produktkategorien, Wettbewerbsformen und Arbeitsweisen. Unternehmen sehen sich gezwungen grundlegende Entscheidungen in Bezug auf ihre Größe, ihren Umfang und ihre Organisation als Ganzes zu treffen. Dabei konnten „Tech-Giganten“ wie Amazon, Alphabet, Microsoft, Alibaba oder Apple durch ihre plattformbasierten Geschäftsmodelle, die allesamt auf umfassenderen Ökosystemansätzen aufbauen, und den damit verbundenen Vorteilen ihre Überlegenheit im Vergleich zu den traditionellen Geschäftsmodellen behaupten (Birkinshaw, 2018; Jacobides et al., 2024).

Digitale plattformbasierte Ökosysteme stellen eine neue Form der Zusammenarbeit dar und basieren auf einer fokalen digitalen Plattform, die durch Komplementaritäten von verschiedenen Agierenden ergänzt wird. Diese tragen somit zum Wertangebot der Plattform bei (Jacobides et al., 2018; Nerbel & Kreutzer, 2023). Die Agierenden und Teilnehmenden sind dabei nicht im klassischen Sinne vertraglich miteinander verbunden, sondern agieren wirtschaftlich unabhängig voneinander (Autio, 2022). Daher bedarf es einer zielgerichteten und sorgfältigen Orchestrierung der Plattformbetreibenden. Unter Orchestrierung versteht man im Grunde das Überzeugen der Agierenden der Plattform, sich in einer Weise zu verhalten, die mit der Ökosystemvision und dem Werteversprechen übereinstimmt (Autio, 2022). Im Gegensatz zu etablierten plattformbasierten Ökosystemen mit einem definierten Wertangebot und einer relativ stabilen Nutzendenbasis, treffen initiale und aufstrebende Plattformen auf unterschiedliche Schwierigkeiten innerhalb ihrer Entstehungsphase. So stellt die anfängliche Initialisierung eines solchen Ökosystems die Agierenden vor die Herausforderung ein kollektives Werteversprechen zu finden und ihre Aktivitäten auf ein gemeinsames Ziel auszurichten (Ofe & Sandberg, 2019). In der Praxis ist es den Agierenden und insbesondere den Plattformbetreibenden daher oftmals nicht klar, welche Aspekte oder Ansätze zur initialen Orchestrierung zu berücksichtigen sind. In der Forschung existieren bereits einige Frameworks oder Ansätze zur Gestaltung und Orchestrierung von plattformbasierten Ökosystemen (Autio, 2022; Nerbel & Kreutzer, 2023; Wulfert et al., 2022), jedoch fehlt bislang eine strukturierte Übersicht über verschiedene Orchestrierungsansätze während der Initialisierung. Eine strukturierte Übersicht ermöglicht es Zusammenhänge zwischen verschiedenen Aspekten und Ansätzen aufzudecken. So können redundante Aktivitäten und Entwurfsprinzipien vermieden werden und eine zielgerichtete Orchestrierung des Ökosystems erfolgen.

Zielsetzung

Das Ziel dieser Arbeit ist es, mit der Beantwortung der Forschungsfrage die Lücke innerhalb der Literatur zu der angesprochenen Problemstellung zu schließen.

Forschungsfrage:

Welche Orchestrierungsansätze bei der gemeinsamen Wertschöpfungsfindung existieren im Zuge der Initialisierung eines digitalen plattformbasierten Ökosystems?

Die Forschungsfrage lässt sich dabei aufteilen in mehrere Teilziele. Eines davon ist die Erkenntnis darüber, ob es einheitliche oder unterschiedliche Orchestrierungsansätze gibt. Falls unterschiedliche Ansätze existieren, sollen außerdem Merkmale ermittelt werden, worin sich diese Ansätze unterscheiden und warum unterschiedliche Ansätze existieren. Ein weiteres Ziel ist es, Aktivitäten zu identifizieren, welche einen Orchestrierungsansatz definieren.

Digitale Plattformbasierte Ökosysteme

Begriffsdefinition und Eigenschaften

Entliehen aus der Biologie bezeichnet ein Ökosystem im Kontext der Ökonomie eine Form der Kollaboration, bei der die einzelnen Agierenden nicht vertraglich miteinander in Verbindung stehen und

dennoch voneinander abhängig sind (Jacobides et al., 2018). Dabei geht die aktuelle Forschung von drei Arten von Ökosystemen aus: einem Business-Ökosystem, das sich insbesondere auf eine Firma und seine Umwelt bezieht, einem Innovationsökosystem, welches sich um eine neue Innovation oder ein neues Werteversprechen bildet, und einem plattformbasierten Ökosystem. Hier steht eine digitale Plattform im Mittelpunkt des Ökosystems, die von Plattformbetreibenden gestellt wird und es den Komplementierenden ermöglicht durch diese direkt oder indirekt mit der Kundschaft der Plattform zu interagieren. Handelt es sich dabei um eine über ein Netzwerk erreichbare Plattform innerhalb eines Ökosystems, so spricht man von einem digitalen plattformbasierten Ökosystem. Allerdings ist hier unter Netzwerk nicht zwingend das Internet zu verstehen, auch wenn die meisten digitalen, plattformbasierten Ökosysteme über dieses erreichbar sind. Eine weitere Möglichkeit bildet beispielsweise ein eigenes gesichertes lokales Netzwerk, das nur innerhalb eines oder mehrerer Firmennetzwerke verfügbar ist, um die Daten nach außen hin abzusichern. Eine solche digitale Plattform kann vor allem im Kontext einer vernetzten Produktion zum Einsatz kommen (Stonig et al., 2022).

Der Begriff und die Idee einer Plattform wurde dabei dem Bereich des Entwicklungsdesigns entnommen, in welchem unter einer Plattform eine modular aufgebaute Produktarchitektur verstanden wird (Jacobides et al., 2024; Ulrich, 1995). Damit wird bereits durch den Namen der Schwerpunkt auf die modulare Architektur eines plattformbasierten Ökosystems gelegt, welche es ermöglicht durch eine Verteilung der Komplexität auf alle Teilnehmende die Koordinierungs- und Verwaltungskosten zu reduzieren (Mukhopadhyay & Bouwman, 2019) und eine effiziente Koordination der unabhängigen Unternehmen sicherzustellen (Jacobides et al., 2018; Zeng et al., 2023).

Gemein ist allen drei Arten von Ökosystemen, dass sich der Wert sowohl für die Kundschaft als auch für die Anbietenden durch zusätzliche Teilnehmende am Ökosystem erhöht. Dadurch ergeben sich, bezogen auf das Ökosystem, sowohl Netzwerkeffekte durch eine höhere Zahl an Teilnehmenden als auch Lock-In-Effekte, die bei einem Wechsel des Ökosystems zu Kosten für den Agierenden führen würden. Diese können entweder durch den Aufwand, die neuen Grenzressourcen anzubinden (Wulfert et al., 2022), oder durch eine hohe Interdependenz zwischen den Ökosystemteilnehmenden entstehen (Schreieck et al., 2021).

Als Beispiel kann hier das Android- und iOS-Ökosystem der Hersteller Google und Apple genannt werden, die durch unterschiedlich genutzte Programmiersprachen und Richtlinien (sowohl in der Gestaltung der Nutzendenoberfläche als auch in der Verfügbarkeit von Hardware-Ressourcen) einen einfachen Wechsel für App-Entwickelnde von der einen zur anderen Plattform und damit auch zwischen den Ökosystemen erschweren. Hierfür muss die entsprechende App an die geltenden Regularien in Hinblick auf das Design und die Rechte- und Ressourcenverwaltung angepasst und in der jeweils verfügbaren Sprache neu kompiliert werden. Damit gehen erhebliche Entwicklungskosten einher, wodurch Google bzw. Apple als Plattformbetreibende gegenüber den Entwickelnden als Komplementierenden eine dauerhafte oder zumindest langanhaltende Bindung an ihr Ökosystem erreichen. Dementsprechend sollte der Wahl des Ökosystems und der zugrundeliegenden Plattform eine besondere Bedeutung zukommen.

Orchestrierung

Ökosysteme bilden sich vor allem dann, wenn ein erheblicher Koordinierungsbedarf besteht, der nicht auf Märkten bewältigt werden kann, aber auch nicht das Diktat und die Autoritätsstruktur zentraler Agierenden erfordert (Jacobides et al., 2018). Denn die Stärke von Ökosystemen und ihre Besonderheit besteht darin, dass sie eine Struktur bieten, innerhalb derer Komplementaritäten aller Art in der Produktion und/oder im Verbrauch eingedämmt und koordiniert werden können, ohne dass eine vertikale Integration erforderlich ist (Jacobides et al., 2018).

Steht ein Unternehmen vor einem solchen Koordinationsproblem, so kann es sich entweder einem bestehenden Ökosystem anschließen oder versuchen ein eigenes Ökosystem aufzubauen. Sich einem Ökosystem anzuschließen, bedeutet wiederum die Regeln des Ökosystems zu berücksichtigen und sich diesen unterzuordnen. Meist werden diese von einer orchestrierenden Partei, die im Falle eines plattformbasierten Ökosystems in aller Regel auch die plattformbetreibende Partei ist, festgelegt. Dadurch wird der orchestrierenden Partei die Möglichkeit gegeben, maßgeblich Einfluss auf die Gestaltung des Ökosystems zu nehmen. Gleichzeitig behalten die einzelnen Agierenden innerhalb des vereinbarten Rahmens einen hohen Grad an Autonomie (Jacobides et al., 2018). Daraus ergeben sich innerhalb eines Ökosystems verschiedene Rollen, die sich in den Aufgaben und Zielen der Agierenden widerspiegeln. Innerhalb der Literatur lassen sich verschiedene Ansätze finden, um diese Rollen zu klassifizieren

(Biedebach und Hanelt, 2020; Iansiti und Levien, 2004; Iyer et al., 2006). Dementsprechend lassen sich die konkreten Agierenden auch verschieden einordnen und je nach genutzter Klassifizierung auch eine unterschiedliche Anzahl von Rollen definieren. Um die dabei auftretenden gegensätzlichen Interessen der einzelnen Agierenden auszubalancieren, bedarf es oft einer gezielten Orchestrierung durch Orchestrierende (Ofe & Sandberg, 2019). Als Beispiel einer solchen Koordinierungsmaßnahme kann die Harmonisierung und Qualitätssicherung der Angebote der Komplementierenden durch die fokale Firma genannt werden (Nerbel & Kreutzer, 2023).

Ein Ökosystem lässt sich in drei Lebensphasen einteilen: Die Entstehungsphase (auch initiale Phase genannt), die Expansionsphase und die Reifephase (Han et al., 2022). In jeder dieser Phasen unterscheiden sich die Charakteristiken der Ökosysteme und damit auch die notwendigen Ansätze und Strategien zur Orchestrierung. Diese Arbeit konzentriert sich auf die Entstehungsphase. In dieser üben die fokalen Firmen häufig eine Form des „sanften" Einflusses (z.B. Framing und Dialog) aus, um die Agierenden innerhalb des Ökosystems zueinander zuführen und gemeinsam eine Vision zu etablieren (Han et al., 2022; Liu & Rong, 2015; Snihur et al., 2018). Gleichzeitig stellt diese Phase die Orchestrierende Partei vor besondere Herausforderungen, da neu entstehende Ökosysteme anfällig für Marktunsicherheiten und Druck von kongruierenden Ökosystemen sind und aufgrund unterschiedlicher Motive, Interessen und Rahmenbedingungen die gemeinsamen Visionen der Agierenden mehrdeutig sein können. Damit kommt einer effizienten Orchestrierung innerhalb der initialen Phase eine besondere Bedeutung zuteil.

Methodik

Forschungsdesign

Die Orchestrierung von plattformbasierten Ökosystemen gewinnt seit dem letzten Jahrzehnt zunehmend an Bedeutung. Dies spiegelt sich in der Anzahl der publizierten Arbeiten in diesem Forschungsbereich wider und unterstreicht die Relevanz einer effizienten Orchestrierung in Hinblick auf die digitale Transformation von Unternehmen (Autio, 2022; Nerbel & Kreutzer, 2023). Um die wesentlichen Aspekte der initialen Orchestrierungsansätze von digitalen plattformbasierten Ökosystemen zu identifizieren und zu untersuchen, wird ein qualitativer Forschungsansatz gewählt. Dieser umfasst die „thematische Analyse" in Anlehnung an Saunders et al. (2019). Der wesentliche Zweck der thematischen Analyse ist die Identifikation von Themenfeldern und Mustern in einem gewählten Datensatz (Saunders et al., 2019). Der flexible und zugleich systematische Ansatz der thematischen Analyse ermöglicht daher eine zielgerichtete Kodierung und Formalisierung von Schlüsselkonzepten in Bezug auf initiale Orchestrierungsansätze von digitalen plattformbasierten Ökosystemen. Aufgrund der Novität des Forschungsgegenstandes wurde ein induktiver Ansatz gewählt. Bei einem induktiven Ansatz lassen sich die wesentlichen Konzepte und Themenfelder aus den Daten ableiten, ohne sich strikt an bereits existierende Rahmenwerke aus der Theorie festzulegen (Saunders et al., 2019). Die gewählte Forschungsmethodik kann daher als geeignet angesehen werden, um die bestehende Problemstellung zu beantworten.

Die thematische Analyse nach Saunders et al. (2019) umfasst im Wesentlichen folgende Schritte: Vertraut werden mit den Daten, Themenfeldersuche und Erkennen von Zusammenhängen und schließlich die Verfeinerung von Themenfeldern, sowie deren Prüfung. Die beschriebenen Schritte müssen jedoch nicht als kausale Abfolge angesehen werden, sondern werden in der Regel in einem iterativen Verfahren durchlaufen. Im ersten Schritt dienten die Arbeiten von Autio (2022) und Jacobides et al. (2024) als Grundlage, um eine Übersicht über das Themengebiet der plattformbasierten Ökosystemen zu erhalten und mit den Daten vertraut zu werden. Darauf aufbauend wurde eine systematische Literaturanalyse durchgeführt, welche zur Generierung der zu untersuchenden Datenbasis dient. Die systematische Literaturanalyse wurde dabei in Anlehnung an Wolfswinkel et al. (2013) durchgeführt und wird im folgenden Kapitel im Detail näher beschrieben. Das Vorgehen zur Kodierung der Daten, sowie die Identifikation von Themenfeldern und Schlüsselkonzepten wird ebenfalls im folgenden Kapitel beleuchtet.

Durchführung der Literaturanalyse und der Thematischen Analyse

Um eine relevante und aussagekräftige Datenbasis zu entwickeln, welche als Grundbaustein für die weitere Kodierung und Themenfeldgenerierung dient, wurde eine systematische Literaturanalyse, basierend auf der Methodik von Wolfswinkel et al. (2013), durchgeführt. Die Literaturanalyse fundiert dabei auf der „Grounded Theory", welche an sich bereits einen induktiven Charakter besitzt (Wolfswinkel et al., 2013).

Mithilfe der Grounded Theory ist es möglich, die Schlüsselkonzepte während des analytischen Prozesses zu identifizieren, anstatt sie vorher deduktiv abzuleiten. Daher bietet sich das Vorgehen nach Wolfswinkel et al. (2013) an, um den ersten Schritt der thematischen Analyse, das Vertrautwerden mit den Daten, zu vollziehen. Das Vorgehen gliedert sich dabei in fünf Phasen und ist ebenfalls iterativer Natur. In der ersten Phase, der Definitionsphase, werden vorerst Kriterien für den Ein- oder Ausschluss von Literatur definiert sowie der eigentliche Umfang der Literaturanalyse festgelegt. Erst in der zweiten Phase erfolgt die eigentliche Suche in den gewählten Datenbanken. Die Selektionsphase ist die dritte Phase des Vorgehens und beschreibt im Grunde die engere Auswahl der zu untersuchende Artikel. Die Analyse- und Präsentationsphase bilden den Abschluss des Vorgehens und umfassen die Kodierung, sowie die Darstellung der Ergebnisse.

In der Definitionsphase ist es erforderlich den Umfang des Forschungsgegenstandes zu bestimmen. Das Themengebiet rund um Plattformen und Ökosysteme ist breit gefächert und nicht immer einheitlich definiert (Jacobides et al., 2024), weshalb eine Eingrenzung erforderlich ist. Die Eingrenzung lässt sich mittels definierter Suchstrings in den Datenbanken bewerkstelligen (Wolfswinkel et al., 2013). Das erste Kriterium bildet das Publikationsjahr der wissenschaftlichen Artikel. Aufgrund der Aktualität des Themas sollen hierbei nur Artikel betrachtet werden, welche nach oder im Jahr 2010 veröffentlicht wurden. Außerdem wurden als weitere Kriterien nur englisch- und deutschsprachige Literatur und zusätzlich nur Zeitschriftenartikel und Konferenzartikel berücksichtigt. Eine Einschränkung der Artikel auf bestimmte Journals oder Gütekriterien, wie beispielsweise die Anzahl an Zitationen, wurde in diesem Schritt nicht vorgenommen. Grund ist die eventuelle Exklusion von relevanten Artikeln, welche noch zu neu sind, sodass sie gegebenenfalls festgelegte Gütekriterien nicht erfüllen.

Ein weiterer Schritt der Definitionsphase ist die Auswahl von geeigneten Datenbanken, in denen die Suche nach den Artikeln stattfindet. In dieser Arbeit wurde sich auf die folgenden Datenbanken festgelegt: Scopus, Web of Science und EBSCOhost. Web of Science, Scopus & EBSCOhost zählen zu den größten Datenbanken und garantieren dadurch eine gewisse Breite. Die Datenbanken umfassen dabei eine Vielzahl an potenziell relevanten Journals und Zeitfachschriften aus allen Branchen. Somit soll sichergestellt werden, dass eine große Anzahl möglich relevanter Artikel gefunden wird. Ein weiterer Aspekt ist die Reproduzierbarkeit des Suchverfahrens in den gewählten Datenbanken, da beispielsweise Google Scholar einen unbekannten Algorithmus verwendet und so die Reproduzierbarkeit nicht immer eingehalten werden kann.

Im nächsten Schritt erfolgte die Formulierung von möglichen Suchstrings. Der Suchstring setzt sich dabei aus drei wesentlichen Bestandteilen zusammen und wird in den Datenbanken mit einer „AND-Verknüpfung“ umgesetzt. Der erste Bestandteil umfasst die zentralen Begriffe „platformbased“ und „ecosystem(s)“. Der zweite Bestandteil beinhaltet den Begriff „orchestration“, sowie weitere Synonyme wie beispielsweise „coordination“ oder „governance“. Den letzten Bestandteil bildet das Themengebiet rund um die Wertschöpfung, also zentrale Aspekte wie „value creation“, „value proposition“, „value discovery“, etc.

Die zweite Phase, die Suchphase, ist die eigentliche Implementation der Suchstrings in den gewählten Datenbanken. Die Suche in den Datenbanken erstreckte sich dabei über den Zeitraum von Mai bis Juni 2023. Die finale Festlegung der Suchstrings erfolgte iterativ, wobei diese schrittweise weiter spezifiziert und eingegrenzt worden sind. Teilweise wurden die Bausteine der Suchstrings modifiziert, da eine unzureichende Anzahl an Artikeln gefunden wurde oder am Thema vorbei gingen.

Der finale Suchstring für Web of Science ergab 88 Artikel und lautet wie folgt:

ALL=("ecosystems" AND ("orchestration" OR "emergence") AND ("value creation" OR "value co-creation" OR "value discovery"))

Der Suchstring für EBSOhost ergab 46 Artikel und setzt sich wie folgt zusammen:

“ecosystems" AND ("orchestration" OR "emergence") AND ("value creation" OR "value co-creation" OR "value discovery"))

Scopus lieferte 288 Artikel mit folgendem Suchstring, wobei dieser String die größten Änderungsiterationen durchlaufen ist:

("value creation" OR "value co-creation" OR "value discovery") AND ("ecosystems") AND ("orchestration" OR "emergence") AND "digital platform" AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (EXACTKEYWORD , "ecosystems") OR LIMIT-TO

(EXACTKEYWORD , "digital platforms") OR LIMIT-TO (EXACTKEYWORD , "platform ecosystems") OR LIMIT-TO (EXACTKEYWORD , "value creation") OR LIMIT-TO (EXACTKEYWORD , "value co-creation") OR LIMIT-TO (EXACTKEYWORD , "ecosystem") OR LIMIT-TO (EXACTKEYWORD , "platform") OR LIMIT-TO (EXACTKEYWORD , "value co creations") OR LIMIT-TO (EXACTKEYWORD , "co-creation") OR LIMIT-TO (EXACTKEYWORD , "platforms") OR LIMIT-TO (EXACTKEYWORD , "platform ecosystem") OR LIMIT-TO (EXACTKEYWORD , "value proposition") OR LIMIT-TO (EXACTKEYWORD , "digital platform ecosystem") OR LIMIT-TO (EXACTKEYWORD , "orchestration") OR LIMIT-TO (EXACTKEYWORD , "ecosystem emergence"))

In der dritten Phase, der Selektionsphase, werden die Ergebnisse aus den Datenbanken zentral gesammelt und Duplikate gefiltert. Abbildung 1 stellt den Prozess der Selektion graphisch dar. Die Datenbasis bestand nun aus möglichen 366 Artikeln, welche in einem weiteren Schritt auf Grundlage des Titels, Abstracts und der Keywords inhaltlich auf den Forschungsgegenstand gescreent wurden. 169 Artikel konnten ausgeschlossen werden aufgrund des fehlenden inhaltlichen Bezugs. Der fehlende Bezug zu einem digitalen Ökosystem führte zu einem Ausschluss von weiteren 32 Artikeln. Bei den restlichen 12 Artikel handelte es sich um Ökosysteme, jedoch fehlte hier der Bezug zur Plattform.

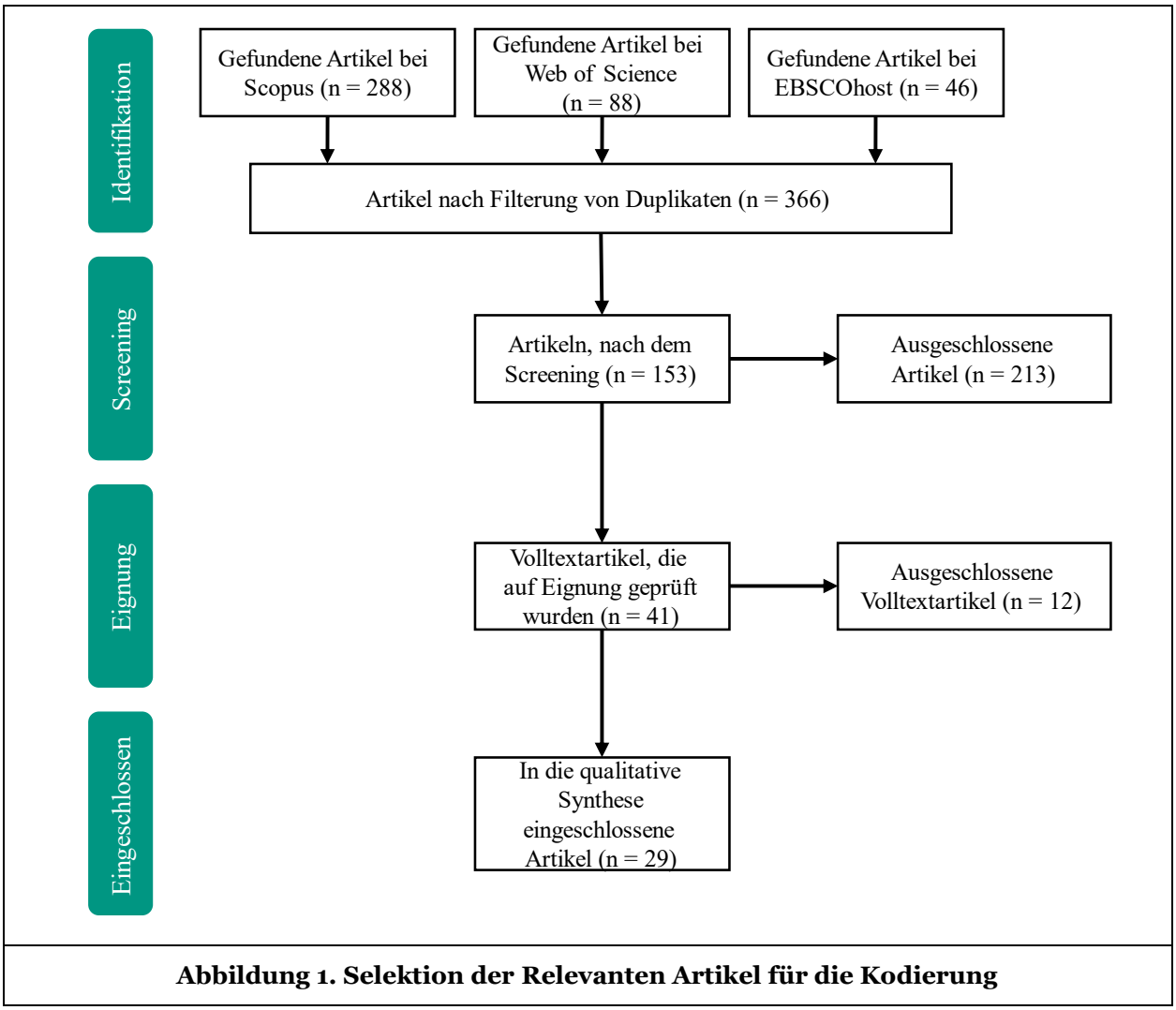

Abbildung 1. Selektion der Relevanten Artikel für die Kodierung

Die Datenbasis setzte sich nach dem Screening aus 41 möglichen Artikeln zusammen. Diese wurden anschließend auf den Volltext gescannt, wodurch weitere 12 Artikel nachträglich ausgeschlossen wurden. Grund war hierbei der fehlende thematische Bezug. Dieser Prozess bildet im Wesentlichen den ersten Schritt der thematischen Analyse nach Saunders et al. (2019).

Der zweite wesentliche Schritt der thematischen Analyse stellt die Kodierung dar. Die Kodierung erfolgte dabei nach Anleitung von Saunders et al. (2019) und Wolfswinkel et al. (2013). Zuerst wurden die 29 Artikel sorgfältig gelesen und daraus Auszüge generiert, welche aus einzelnen Textabschnitten bestanden. Jeder Auszug wurde anschließend in einen deskriptiven Code umgewandelt, der eine Art Zusammenfassung oder ein übergeordnetes Konzept darstellt. Insgesamt wurden 178 Codes generiert, welche von 176 Textstellen der eingeschlossenen Artikel entstammen. Dieser Schritt wird auch als offene Kodierung bezeichnet und kann als erster Abstraktionsschritt gesehen werden (Wolfswinkel et al., 2013). Die Codes wurden dabei tabellarisch organisiert und es wurde dokumentiert, aus welchen Artikeln sie entstanden. Die Generierung der Codes ist dadurch datengestützt (Saunders et al., 2019). So wurde exemplarisch anhand der Textstelle „A competitive advantage in platform markets can be achieved by establishing a strong identity and distinctiveness, making the platform unique." (Nerbel & Kreutzer, 2023, S. 9), folgender Code generiert „Unverwechselbarkeit und Einzigartigkeit der Plattform sind ein Wettbewerbsvorteil".

Nachdem die Kodierung abgeschlossen wurde, wurden Codes mit ähnlicher Bedeutung gruppiert und übergeordnete Themenfelder daraus abgeleitet. Das Verbinden von Codes und die Identifikation von Beziehungen wird auch axiale Kodierung genannt (Wolfswinkel et al., 2013). Die Themenidentifizierung war hierbei ein iterativer Prozess. Innerhalb dieses Prozesses wurden Codes mehrmals umbenannt, umgruppiert und neue hinzugefügt oder entfernt. Das Ergebnis bildete eine Anzahl von 17 Themenfeldern, welche die Zusammenhänge der einzelnen Codes darstellen. Beispielsweise wurden die beiden Codes „Standardisierung von Schnittstellen/Grenzressourcen für einen vereinfachten Eintritt" und „Grenzressourcen sollten vor der Integration von Komplementierenden bereitstehen" zu dem Themenfeld „Grenzressourcen" aggregiert. Abbildung 2 stellt die Aggregation der einzelnen Codes zu den Themenfeldern dar. So konnte dem Themenfeld „Rollenverteilung" insgesamt 29 Codes zugewiesen werden, während dem Themenfeld „Nutzung von Technologien" nur zwei Codes entsprechen.

Das selektive Kodieren bildet den nächsten Schritt und damit auch den Abschluss der thematischen Analyse (Wolfswinkel et al., 2013). In diesem Schritt wurden die Themenfelder verfeinert und übergeordnete Kategorien definiert, welche wiederum als aggregierte Themenfelder angesehen werden können. Hierbei gliedert sich z.B. die Kategorie „Technische Aspekte" aus den fünf Themenfeldern „Grenzressourcen", „Grad der Offenheit", „Nutzung von Technologien", „Plattformarchitektur" und „Technische Standards". Das Resultat stellt eine strukturierte und hierarchische Darstellung von Themenfelder und Konzepten dar, die sich in übergeordnete Kategorien und untergeordnete Themenfelder gliedert.

Zentrale Aspekte der Orchestrierung Innerhalb der Initialen Phase

Die Ergebnisse der Literaturanalyse legen nahe, dass die identifizierten Themenfelder in drei Kategorien eingeteilt werden, aus deren Schnittmenge sich ein effizienter Orchestrierungsansatz innerhalb der initialen Phase ergibt (vgl. Tabelle 1).

Die Kategorie der Orchestrierung beschreibt dabei sowohl die Herausforderungen, denen sich die einzelnen Agierenden und insbesondere die Orchestrierenden ausgesetzt sehen als auch die Mechanismen, die zu einer erfolgreichen Zusammenarbeit innerhalb des Ökosystems führen. Dazu zählen alle Aktivitäten, die darauf ausgerichtet sind, die Attraktivität des Ökosystems für neue Teilnehmende zu steigern und gleichzeitig das Handeln aller Teilnehmende auf ein gemeinsames Werteversprechen auszurichten. So zum Beispiel das Definieren von Regeln und Standards durch Orchestrierende und die sich durch die Aktivitäten der einzelnen Agierenden ergebende Rollenverteilung innerhalb des Ökosystems.

Daran anknüpfend lassen sich innerhalb der Kategorie der Wertschöpfung alle Themenfelder einordnen, die die Ausrichtung auf das gemeinsame Ziel, die Vision des Ökosystems, und damit verbunden ein gemeinsames Werteversprechen gegenüber der Kundschaft konkretisieren.

Dazu zählen neben der Vision selbst die Abhängigkeiten zwischen den einzelnen Agierenden, die Entwicklung und die Aktivitäten zur Umsetzung des Werteversprechens und die wesentlichen Eigenschaften des Werteversprechens, die dieses von der Konkurrenz abhebt. Dadurch erreichen die durch das Ökosystem verbundenen Agierenden ihren spezifischen Wettbewerbsvorteil gegenüber ihrer Konkurrenz.

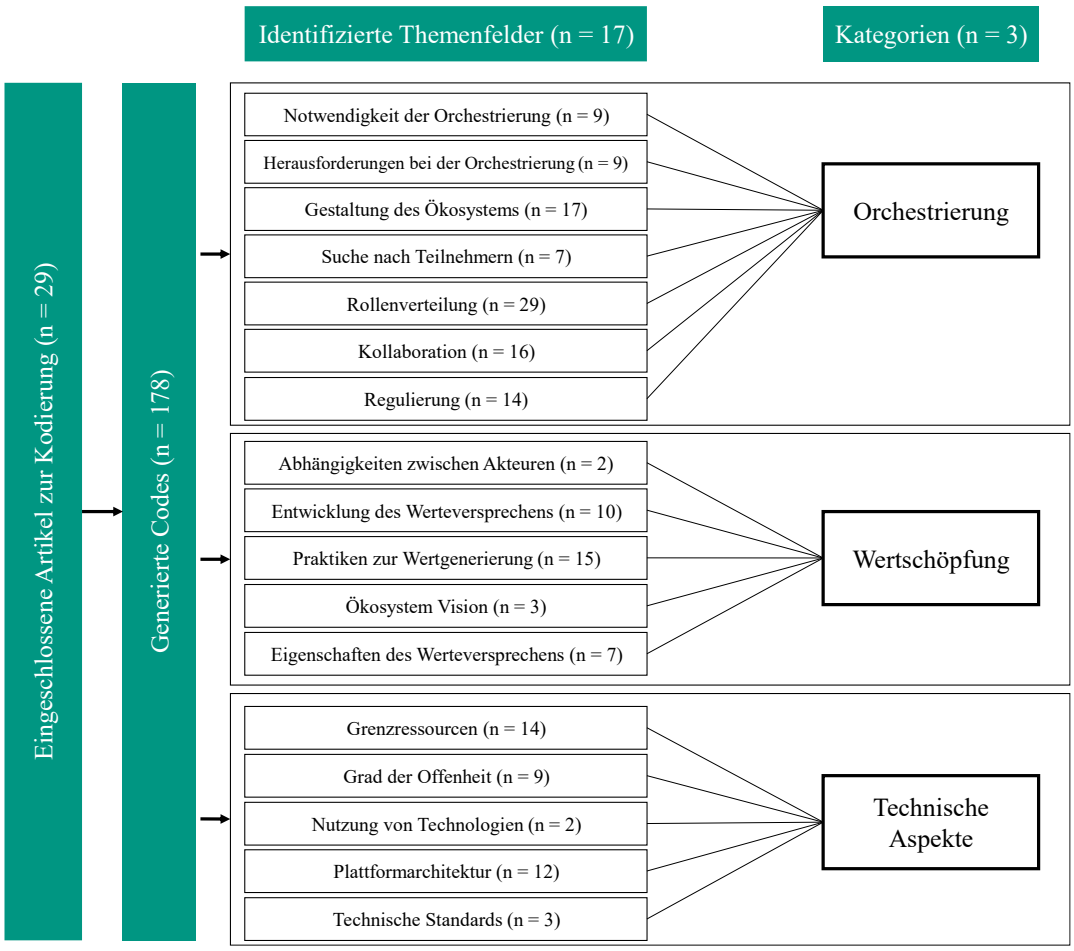

Abbildung 2. Übersicht der Identifizierten Themenfelder und der Übergeordneten Kategorien

Unter den technischen Aspekten werden die strukturellen und technologischen Eigenschaften der Plattform zusammengefasst. Dazu gehören grundlegende Entscheidungen der Plattformbesitzenden, die im Zuge der Plattformentwicklung die Plattformstruktur festlegen, den Grad der Offenheit bestimmen und die technischen Standards definieren. Dadurch prägen die technischen Aspekte die Plattform und infolgedessen auch das Ökosystem maßgeblich.

In den nachfolgenden Unterkapiteln werden die Themenfelder genauer beschrieben, Interdependenzen aufgezeigt, die sich durch das Zusammenspiel der einzelnen Aspekte ergeben und ihre jeweiligen Auswirkungen auf die zugehörige Kategorie dargelegt.

Orchestrierung

Notwendigkeit der Orchestrierung

Plattformbasierte Ökosysteme tragen durch einen modularen Aufbau dazu bei, miteinander verbundene Organisationen zu koordinieren, die über eine erhebliche Autonomie verfügen (Jacobides et al., 2018). So können diese weitestgehend frei über ihre Preisgestaltung, das Design und den Inhalt ihrer jeweiligen Produkte entscheiden, solange diese den übergeordneten Architekturparametern des Ökosystems entsprechen und die bereitgestellten Grenzressourcen genutzt werden, um eine einheitliche Kommunikation gewährleisten zu können.

Allerdings können die fehlenden hierarchischen Strukturen, das Fehlen klarer Definition von Rollen und Verantwortungen und die teilweise gegensätzlichen Interessen der einzelnen Agierenden zu Koordinationsproblemen führen (Hodapp et al., 2019; Jacobides et al., 2018; Ofe & Sandberg, 2019), woraus sich ein Orchestrierungsbedarf innerhalb des Ökosystems ergibt. Diese Aufgabe übernimmt in der Regel die plattformbesitzende Partei, welche somit als orchestrierende Partei auftritt und die Prozesse, Regeln und Standards festlegt bzw. anpasst, die innerhalb des Ökosystems für alle verbindlich gelten.

Herausforderungen bei der Orchestrierung

Aus der Notwendigkeit einer Orchestrierung und den damit einhergehenden Regeln und Prozessen ergeben sich oft auch Herausforderungen sowohl in Bezug auf die Rollen- und Machtverteilung innerhalb des Ökosystems als auch in Bezug auf die Orchestrierung an sich. Als Beispiel kann hier der Fall „Epic Games vs. Apple“ angeführt werden, bei dem sich Epic Games als app-anbietendes Unternehmen durch das Umgehen der festgeschriebenen Regeln zur Abwicklung von Bezahlvorgängen über den App Store nicht länger an die Standards und Regeln, die Apple als plattformbesitzende Partei festgelegt hat, halten wollte (Ambrasaitė & Smagurauskaitė, 2021). Weitere Faktoren, welche Orchestrierende vor Herausforderungen bei der Gestaltung und Orchestrierung stellen, sind ein fehlendes ganzheitliches und tiefergehendes Verständnis des Aufbaus und der Dynamik des Ökosystems (Han et al., 2022), das Fehlen klarer Definitionen von Rollen und Verantwortungen für jede agierende Partei (Hodapp et al., 2019) sowie teils gegensätzliche Ziele der einzelnen Agierenden (Ofe & Sandberg, 2019).

Um diesen Herausforderungen zu begegnen, sollten Orchestrierende sowohl ein besonderes Augenmerk auf die eigenen Fähigkeiten zur Orchestrierung des Ökosystems (Helfat & Raubitschek, 2018) als auch auf die Kooperationsbereitschaft zwischen den Agierenden legen (Ben Letaifa, 2014; Helfat & Raubitschek, 2018). Hierzu ist insbesondere die Identifikation von Schlüsselfaktoren wichtig, die die gemeinsamen Anstrengungen aller Agierenden und die Kooperationsbereitschaft zwischen den Agierenden fördern, um eine stabile und nachhaltige Entwicklung des Ökosystems zu ermöglichen (Ben Letaifa, 2014; Xu et al., 2023).

Gestaltung des Ökosystems

Ein zentraler Bestandteil der Orchestrierung ist die Gestaltung des Ökosystems. Ein weit gefasstes Gebiet, das im Wesentlichen jedoch auf die Allokation solcher Ressourcen beschränkt ist, die nicht organisationaler Natur sind. Wiedergegeben wird die Gestaltung durch drei Aspekte: die Attraktivität des Ökosystems, Innovationsorchestrierung und die Ökosystemarchitektur. Die Attraktivität ist ein wichtiges Kriterium, um externe Agierende bzw. Nutzende für das Ökosystem zu gewinnen und interne im Ökosystem zu halten. Notwendig dafür sind monetäre und nichtmonetäre Anreize (Ofe & Sandberg, 2019). Neben dem Werteversprechen sind z.B. die Kosten der Austauschbarkeit von Teilnehmenden von Bedeutung (Jacobides et al., 2018), d.h. zu Beginn sollten diese niedrig gehalten werden, um die Hürden für neue Teilnehmende zu reduzieren. Gleichzeitig ist die Erzeugung von Lock-In Effekten essenziell, wobei schon während der Initialisierung der Fokus auf beziehungsorientierte Lock-In Effekte, wie z.B. ein gemeinsamer Vertrieb über die Plattform, gesetzt werden sollte (Schreieck et al., 2021). Diese bieten gegenüber technischen Lock-In Effekten den Vorteil, dass sie nicht nur den Wert des Ökosystems steigern, sondern auch für alle Teilnehmende und insbesondere für die Plattformbesitzenden gewinnbringend sind. Der zweite Aspekt betrifft das Ökosystem als Lernumgebung um Innovationen zu fördern. Neben der Einrichtung eng gekoppelter Koordinationsmechanismen, wie kollektives Lernen (Han et al., 2022), ist auch die Verwaltung des Wissensaustauschs (de Vasconcelos Gomes et al., 2022) Teil der Orchestrierung. Zuletzt wird im Zuge der Gestaltung der grundlegende Aufbau des Ökosystems, d.h. die Architektur, festgelegt. Dazu werden die Schnittstellen auf organisatorischer Ebene koordiniert (Nerbel & Kreutzer, 2023), Regeln bzw. Standards definiert (Jacobides et al., 2018) und entsprechend der Breite des Wertangebotes die Ressourcen diversifiziert (Blaschke et al., 2018). Dadurch kommt es zu Überschneidungen zwischen der Rollenverteilung, der Machtverteilung, der Suche nach Teilnehmenden und der Regulierung mit der Gestaltung des Ökosystems. Allerdings steht bei diesen mehr die Interaktion der Agierenden untereinander im Vordergrund und weniger das Ökosystem als Entität.

Kategorie	Themenfelder	Beschreibung
Orchestrierung	Notwendigkeit der Orchestrierung	Erläutert die essenzielle Rolle der Koordination in plattformbasierten Ökosystemen, um trotz fehlender Hierarchien eine effektive Zusammenarbeit sicherzustellen.
	Herausforderungen bei der Orchestrierung	Beschreibt Herausforderungen in Bezug auf die Orchestrierung für eine oder mehrere Rollen innerhalb des Ökosystems.
	Gestaltung des Ökosystems	Umfasst Attraktivität, Innovationsorchestrierung und Ökosystemarchitektur.
	Suche nach Teilnehmenden	Erläutert die verschiedenen Anforderungen an den Suchprozess nach neuen Teilnehmenden.
	Rollenverteilung	Beschreibt die Fähigkeiten, die Teilnehmende benötigen und die Aufgaben/Rollen, die diese innerhalb des Ökosystems wahrnehmen.
	Kollaboration	Stellt die Art der Zusammenarbeit der Agierenden innerhalb des Ökosystems dar.
	Regulierung	Beschreibt formal die Zusammenarbeit und beinhaltet Regeln, welche die Autonomie der Teilnehmenden je nach Ausmaß einschränken.
Wertschöpfung	Abhängigkeiten zwischen Agierenden	Beschreibt auf abstrakter Ebene die (Wirk-)Beziehungen zwischen den Teilnehmenden des Ökosystems.
	Entwicklung des Werteversprechens	Beschreibt den Prozess von anfänglichen Geschäftsmöglichkeiten über Komplementierenden-Anziehung bis hin zum etablierten Werteversprechen.
	Praktiken zur Wertgenerierung	Umfasst allgemeine Prozesse und Aktivitäten, die einen gewissen Wert innerhalb oder außerhalb des Ökosystems generieren.
	Ökosystem Vision	Erläutert den Nutzen einer übergeordneten, gemeinsamen Vision der einzelnen Agierenden innerhalb des Ökosystems.
	Eigenschaften des Werteversprechens	Beschreibt auf welchen Eigenschaften das Werteversprechen basiert.
Technische Aspekte	Grenzressourcen	Sind essenzielle technische Instrumente innerhalb von Plattform-Ökosystemen, bestehend aus APIs, Softwaretools und Richtlinien, welche die Grundlage für Komplementierende bilden.
	Grad der Offenheit	Erläutert, wie vertikale und horizontale Offenheit der Plattform das Ökosystem beeinflusst.
	Nutzung von Technologien	Beschreibt, wie moderne Technologien sinnvoll genutzt werden können.
	Plattformarchitektur	Stellt den Aufbau der Plattform und dessen entscheidende Charakteristika dar.
	Technische Standards	Erläutert, wie solche Standards Interoperabilität gewährleisten können.

Tabelle 1. Beschreibung der Themenfelder

Suche nach Teilnehmenden

Eng an die Attraktivität des Ökosystems geknüpft ist die Suche nach neuen Teilnehmenden. In der initialen Phase bilden sich die Kernpartnerschaften heraus, d.h. im ersten Schritt wird nach strategischen Agierenden gesucht, wenn sie nicht schon aus dem bisherigen Geschäft vorhanden sind (Lechowski & Krzywdzinski, 2022; Schreieck et al., 2021; Tsai et al., 2022). Anschließend wird im zweiten Schritt nach möglichst vielen weiteren Agierenden, z.B. nach innovativen Start-Ups, gesucht, um das Ökosystem weiterzuentwickeln (Tsai et al., 2022). Die Anzahl an Teilnehmenden insgesamt verläuft dementsprechend progressiv (de Vasconcelos Gomes et al., 2022). Da die Teilnahme am Ökosystem freiwillig ist, werden von Beginn an die Auswirkungen eines Austritts mitberücksichtigt (Mbanefo & Saartijie Grobbelaar, 2022). Bei der Suche ist zu beachten, dass der Ansatz, wie die Suche durchgeführt wird, mit der Art des Ökosystems in Verbindung steht. Murthy und Madhok (2021) entdecken beispielsweise, dass mit steigender Komplexität des Zwecks eines Ökosystems mehr Aktivitäten bei den Plattformbesitzenden verbleiben sollten und die Suche gezielter abläuft. Bei einem einfachen Zweck und gleichzeitig geringen Experimentierkosten werden Agierende hingegen über ein Trial and Error Verfahren am Markt gefunden. Besitzt das Ökosystem eine orchestrierende Partei, so fällt die Suche in seinen Aufgabenbereich.

Rollenverteilung

Je nach strategischer Ausrichtung nehmen Firmen innerhalb des Ökosystems verschiedene Rollen ein (Biedebach & Hanelt, 2020). Dabei beschreibt die Rollenverteilung zum einen die Fähigkeiten, welche die Teilnehmenden zur Ausführung ihrer Rolle benötigen, aber auch die konkreten Aufgaben, die sie als Agierende innerhalb des Ökosystems wahrnehmen. So lassen sich im Kontext eines digitalen plattformbasierten Ökosystems vor allem drei wesentliche Rollen unterscheiden: die Nutzenden, ein oder mehrere Plattformanbietende und die Entrepreneure (Acs et al., 2021; Helfat & Raubitschek, 2018). Die Nutzenden lassen sich wiederum in zwei Gruppen aufteilen: Nutzende, welche die Dienstleistungen der Anbietenden in Anspruch nehmen und Nutzende, welche Dienstleistungen über die Plattform anbieten. Hier unterscheidet sich die Rollenverteilung insbesondere in den benötigten Fähigkeiten bzw. Ausstattungen. Uber-Fahrende zählen beispielsweise zu den anbietenden Nutzenden (Acs et al., 2021), welche die Plattform, die Uber als Plattformanbietende zur Verfügung stellt, nutzen, um wiederum Nutzende zu finden, die ihre Dienstleistung in Anspruch nehmen wollen. Die Fahrenden bieten damit ihre Fähigkeit ein Auto zu fahren und ihr Auto als Ausstattung denjenigen Personen an, die bereit sind hierfür Geld zu bezahlen. Zu den Entrepreneuren zählen beispielsweise App-Entwickelnde, welche das Ökosystem um weitere Funktionen und Angebote erweitern und damit durch Prozess- und Produktinnovationen dem Ökosystem einen Mehrwert bieten.

Die Plattformanbietenden treten meist als zentrale Agierende auf, weshalb diese auch als „Plattformführende“ (Mukhopadhyay & Bouwman, 2019), „Architekt“ (Gulati et al., 2012; Helfat & Raubitschek, 2018) oder „Hub“ (Jacobides et al., 2018) bezeichnet werden. Diese stellen die Plattform bereit und legen die grundsätzlichen Regeln und Standards des Ökosystems fest (Helfat & Raubitschek, 2018). Damit übernimmen die Plattformanbietenden in der Regel auch die Orchestrierung der Plattform (Helfat & Raubitschek, 2018; Ofe & Sandberg, 2019). Dadurch entsteht im Vergleich zu den übrigen Agierenden ein deutliches Machtgefälle, denn diese müssen sich trotz ihrer formalen Unabhängigkeit den Regelungen der Plattform anpassen (Jacobides et al., 2018), um Teil des Ökosystems zu sein. Aufgrund dieses Machtgefälles können, insbesondere durch die gegenseitigen Abhängigkeiten, Herausforderungen in der Orchestrierung und damit auch innerhalb der Kollaboration der Agierenden entstehen.

Kollaboration

Erst durch die Zusammenarbeit aller Agierenden innerhalb des Ökosystems entfalten sich die Synergien. Die Kollaboration umfasst daher verschiedene Vereinbarungen, die eine kooperative, wechselseitige Interaktion sicherstellt. Hilfreich dafür ist die Ausrichtung der Ziele aller Agierenden (Stonig et al., 2022; Wulfert et al., 2022) und die Etablierung einer eigenen Ökosystemkultur (Zeng et al., 2023). Aufgrund der neuartigen Weise zusammenzuarbeiten, d.h. nicht in der klassischen Auftraggebende-Auftragnehmende Beziehung, ist besonders in den frühen Phasen, die durch hohe Dynamik gekennzeichnet sind und vertragliche Flexibilität erfordern, gegenseitiges Vertrauen grundlegend (Hodapp et al., 2019). Im Laufe der Zeit wird dieses allerdings durch Verträge abgesichert. Bezüglich der Plattform und der Verarbeitung von unternehmenskritischen Daten ist neben dem Vertrauen auch ein offener und kontinuierlicher

Austausch von Informationen und Wissen notwendig (Ofe & Sandberg, 2019), was auch allgemein die Beziehung der Agierenden untereinander stärkt (Blaschke et al., 2018).

Regulierung

Abschließend gilt es im Zuge der Orchestrierung gewisse Formalitäten zu definieren und einzuhalten, um das Ökosystem zu regulieren. Ein wesentlicher Aspekt ist der Umgang mit Eigentumsrechten, da neue Agierende ihr Eigentum schützen wollen (Huang et al., 2013). Außerdem sollte Einigkeit über sämtliche Regeln und Kriterien bestehen, wie z.B. den Eintritt in das Ökosystem. Mit Regulierungen gehen auch Kontrollen einher. Diese sichern z.B. die Qualität (Schreieck et al., 2021), dienen mehr der Unterstützung als der starren Kontrolle (Zeng et al., 2023) und werden von einem zentralen Organ übernommen (Blaschke et al., 2018). Je nach Ausmaß der Regulierung und Kontrolle wird die Autonomie der Teilnehmenden stärker eingeschränkt, weshalb ein Gleichgewicht beider Teile hergestellt werden muss (Mukhopadhyay & Bouwman, 2019). Falls dennoch Konflikte auftreten, besitzen Ökosysteme den Vorteil, dass diese untereinander direkt gelöst werden können und kein zentrales Konfliktmanagement notwendig ist (van Vulpen et al., 2022). Die Regulierung ordnet sich neben der Kollaboration ein, indem sie formell die Zusammenarbeit regelt.

Wertschöpfung

Abhängigkeiten Zwischen Agierenden

Die Agierenden innerhalb eines Ökosystems weisen oft eine starke Abhängigkeit zu den Kernkompetenzen anderer auf. Denn innerhalb eines Ökosystems wird der Wert erst zusammen mit den Kompetenzen bzw. Produkten der anderen Agierenden entwickelt und generiert (Linde et al., 2021). Dabei sind diese Abhängigkeiten nicht nur zwischen den einzelnen Agierenden untereinander sichtbar, sondern es besteht auch zwischen den Plattformanbietenden und den Komplementierenden ein Abhängigkeitsverhältnis.

Bei der Werteentwicklung entdecken die Agierenden im Ökosystem oft verschiedene neue Praktiken. Die Plattformanbietenden müssen sich daher dauernd an die Bedürfnisse der Agierenden anpassen und somit immer unter Unsicherheit der technischen Entwicklungsrichtung der Plattform handeln (Ofe & Sandberg, 2019). Die Abhängigkeit zwischen den Agierenden stellt damit eine Grundlage für die effektive Entwicklung des Werteversprechens und die Wertgenerierung dar.

Entwicklung des Werteversprechens

In der ersten Phase der Ökosystementstehung finden die zukünftigen Plattformbesitzenden eine Geschäftsmöglichkeit für ein plattformbasiertes Produkt oder eine Dienstleistung und gründeen daraufhin die digitale Plattform. Diese Geschäftsmöglichkeit wird anschließend allokiert und es wird versucht über Netzwerkeffekte Komplementierende anzuziehen (Nerbel & Kreutzer, 2023). Insbesondere bei gerade aufstrebenden Plattformen ist der Wert teilweise noch unsicher und undurchsichtig, da diese oft eine instabile Nutzendenbasis aufweisen und der Wert mit dieser Nutzendenbasis eng verbunden ist (Ofe & Sandberg, 2019). Daher ist es vor allem bei der Entwicklung des Werteversprechens für den Orchestrierende wichtig, fortlaufend im Detail zu verstehen, welche Probleme die Kundschaft beschäftigen, was sie benötigen und was für sie am wichtigsten ist (Linde et al., 2021).

In der weiteren Entwicklung des Werteversprechens ist zu beobachten, dass dieses mit der Reifung des Ökosystems immer wieder überarbeitet wird. Diese Überarbeitung wird oftmals von neuen Komplementierenden vorangetrieben, die die Fortschritte in digitalen Technologien nutzen, um neue Geschäftsmodelle zu ermöglichen (Kohtamäki et al., 2020; Miehé et al., 2023; Thomas et al., 2022). Dabei ist es sinnvoll, gemeinsam mit der orchestrierenden Partei neue Geschäftsmöglichkeiten zu erkunden und dadurch das Werteversprechen des Ökosystems zu erweitern (Miehé et al., 2023).

Praktiken zur Wertgenerierung

Diese Kategorie befasst sich mit allgemeinen Aktivitäten, die zur Schaffung und zur Steigerung von Wert in digitalen plattformbasierten Ökosystemen führen. Innerhalb eines Ökosystems ist Innovation essenziell für die Wertgenerierung und diese kann nur erreicht werden, wenn alle Agierenden am Informationsfluss

innerhalb des Ökosystems teilnehmen (Biedebach & Hanelt, 2020; Yoo et al., 2010). Die Plattform ermöglicht dabei den Gegenstand des Austauschs festzulegen, die Lieferung und Bezahlung zu erleichtern und eine institutionelle Infrastruktur bereitzustellen (Ghazawneh & Henfridsson, 2015; Hodapp et al., 2019).

Neben dem Informationsfluss auf der Plattform und der Plattforminfrastruktur sind die Erweiterungen der Komplementierenden und somit auch deren aktive Teilnahme und das Management der Erweiterungen wesentlich für die Wertgenerierung (McIntyre & Srinivasan, 2017; Murthy & Madhok, 2021). Jeder neue Komplementierende soll die Leistungen der Plattform steigern und somit den Wert erhöhen (Miehé et al., 2023). Die Praktiken zur Wertgenerierung haben enge Verknüpfungen sowohl zur technischen als auch zur orchestrierenden Dimension.

Ökosystem Vision

Das Ziel der zukunftsorientierten Betrachtung des Ökosystems ist es den Komplementierenden mit einer gemeinsam geteilten Vision oder einem Markenimage der Plattform einen Anreiz zu bieten, mehr zur Plattform und somit zum Ökosystem beizutragen (Schreieck et al., 2021). Hierbei ist wesentlich, dass diese Vision von allen Ökosystemteilnehmenden geteilt wird und eine starke Identität sowie gewisse Alleinstellungsmerkmale aufweist, um möglichst effektiv im Markt konkurrieren zu können (Mukhopadhyay & Bouwman, 2019; Nerbel & Kreutzer, 2023).

Die Ökosystem Vision ist eng mit der Gestaltung des Ökosystems beziehungsweise mit dessen Attraktivität verknüpft, da die Vision eine Inspiration und somit auch einen gewissen Anreiz und eine Motivation für die Agierenden darstellt.

Eigenschaften des Werteversprechens

Bei digitalen plattformbasierten Ökosystemen basiert das Werteversprechen hauptsächlich auf der Teilnahme der Nutzenden und den dadurch entstehenden Netzwerkeffekten (Chen et al., 2019; Nambisan et al., 2019; Zeng et al., 2023). Wenn ein neuer Komplementierender dem Ökosystem beitritt, wird das Werteversprechen des Ökosystems beeinflusst. Entweder wird es vertieft und gestärkt oder es verändert sich und weitet sich aus (Miehé et al., 2023).

Die Eigenschaften des Werteversprechens sollten deshalb eng mit der Suche nach Teilnehmenden gekoppelt sein, da die dadurch gefundenen Teilnehmenden wiederum das Werteversprechen selbst beeinflussen können.

Technische Aspekte

Grenzressourcen

Grenzressourcen sind ein zentrales technisches Steuerelement und umfassen APIs, Softwareentwicklungstools, Leitlinien für die Anwendungsentwicklung und technische Dokumentation (Mukhopadhyay & Bouwman, 2019), die ein gemeinsames Toolset und einen Interpretationsstandard für Komplementierende schaffen, auf dem diese aufbauen können (Benlian et al., 2015; Nerbel & Kreutzer, 2023).

Damit Drittanbietende problemlos Module beisteuern können, müssen die eingeführten Grenzressourcen klare Arbeitsabläufe vorgeben (Wulfert et al., 2022) und festlegen, nach welchen Richtlinien die zukünftigen Komplementierenden interagieren. Das führt dazu, dass über die Grenzressourcen Governance-Aspekte in die Umgebung der Plattform getragen werden (Nerbel & Kreutzer, 2023). Dabei ist wichtig, dass die Grenzressourcen auf die Bedürfnisse der Ökosystemteilnehmenden zugeschnitten sind (Wulfert et al., 2022). Dementsprechend muss das fokale Unternehmen in der Lage sein zu erkennen, welche zusätzlichen Grenzressourcen noch fehlen und zur Plattform hinzugefügt werden sollten (Nerbel & Kreutzer, 2023).

Bezüglich des Grades der Offenheit wird für viele Ökosysteme eine offene Architektur empfohlen (Wulfert et al., 2022), um die Generativität zu steigern. Gleichzeitig versuchen Orchestrierende jedoch unter anderem mithilfe der Grenzressourcen die Kontrolle über die Komplementierenden aufrecht zu erhalten

(Mukhopadhyay & Bouwman, 2019). Dementsprechend gibt es einen Zusammenhang zwischen den Grenzressourcen, dem Grad der Offenheit, der Plattformarchitektur und der Vision eines Ökosystems.

Grad der Offenheit

Beim Grad der Offenheit wird oftmals zwischen vertikaler und horizontaler Offenheit differenziert. Die vertikale Offenheit bezieht sich auf den Grad der Zugänglichkeit und Transparenz der Grenzressourcen für die Komplementierenden. Bei der horizontalen Offenheit geht es mehr um die Kompatibilität mit anderen Plattformen und die Bereitschaft zum Teilen des Plattformeigentums mit anderen (Weiss et al., 2020).

Über verschiedene Zugriffsgrade können die Plattformressourcen so konfiguriert werden, dass den Entwickelnden, je nach den von ihnen erstellten Diensten, unterschiedliche Anreize für den Zugang zur Plattform geboten werden (Ofe & Sandberg, 2019). Ein hoher Grad an Offenheit der Plattform führt zum Beispiel dazu, dass mehr Komplementierende dem Ökosystem beitreten und sich somit der Gesamtwert der Plattform erhöht (Nerbel & Kreutzer, 2023).

Sowohl die vertikale als auch die horizontale Offenheit trägt dazu bei, dass der Plattformeigentümer über potenzielle Plattformverbesserungen lernt. Dabei ist der Grad der Offenheit keinesfalls fest, sondern kann im Verlauf der Ökosystementwicklung angepasst werden. So senkte beispielsweise Google nach einiger Zeit den Grad der Offenheit seines Ökosystems, um mehr Kontrolle über die Plattform zu erlangen, nachdem sich bereits eine große Kundschaft und viele Komplementierende an Google gebunden hatten (Weiss et al.). Somit herrscht ein enger Zusammenhang zwischen dem Grad der Offenheit und den Grenzressourcen, da die Offenheit größtenteils über die Grenzressourcen reguliert werden kann.

Nutzung von Technologien

Innerhalb digitaler plattformbasierter Ökosysteme spielt die Nutzung von Technologien eine entscheidende Rolle. Tencent beispielsweise setzt auf die Nutzung von KI-Technologien, um sein Geschäft weiter auszubauen und neue Projekte zu integrieren. Das ermöglicht ihnen effizient auf große Datenbanken und intelligente Algorithmen zuzugreifen, um schneller Kenntnisse über die Agierenden innerhalb des Ökosystems zu gewinnen, einfacher Urteile zu fällen und auf bestimmte Anfragen schneller reagieren zu können (Zeng et al., 2023).

Die Nutzung von Technologie steht in einer engen Verbindung zu den Themenfeldern der Orchestrierung, da diese es erst ermöglichen ein Ökosystem effektiv zu skalieren und zu verwalten.

Plattformarchitektur

Die Plattformarchitektur beschreibt den Aufbau der Plattform. Die entscheidenden Charakteristika der Plattformarchitektur sind die Modularität und die Komplexität. Die Hauptaufgabe der Module ist es die Kernfunktionalität der Plattform zu verbessern, wobei jedes Modul eine beabsichtigte Funktion beinhaltet. Die Komplexität der Plattform kommt durch die intermodulare Interaktion zustande. Je mehr die Module interagieren, desto höher wird die Plattformkomplexität (Cennamo et al., 2018; Kapoor et al., 2021; Singaraju et al., 2016).

Eine modulare Architektur senkt die Koordinierungskosten, wodurch die vorhandenen Ressourcen spezialisierter eingesetzt werden können. Gleichzeitig kann eine zu hohe Modularität jedoch auch dazu führen, dass die Einzigartigkeit der einzelnen Module beeinträchtigt wird. Darüber hinaus kann es zu einer Verringerung des gegenseitigen Lernens kommen, da in der Regel mit einer hohen Modularität eine Verringerung der Interaktion zwischen diesen einhergeht (Mukhopadhyay & Bouwman, 2019).

Die Plattformarchitektur stellt die Grundlage für die gesamte Entwicklung und Orchestrierung eines plattformbasierten Ökosystems dar. Besonders entscheidend ist die Architektur für die effektive Implementierung der Grenzressourcen, sodass diese für zukünftige Komplementierende nutzbar sind (Cennamo et al., 2018; Nerbel & Kreutzer, 2023; Tiwana, 2014; Weiss et al., 2022).

Technische Standards

Technische Standards haben häufig das primäre Ziel Interoperabilität zu gewährleisten. Alle Komplementierenden, aus denen ein Kunde wählt, nutzen die gleichen technischen Standards und sind somit eng miteinander verbunden und voneinander abhängig (Jacobides et al., 2018).

Technische Standards sind ebenfalls ein Mittel zur Bewältigung von Herausforderungen in der Infrastruktur. Denn durch gemeinsame Standards können Engpässe beseitigt werden (Lechowski & Krzywdzinski, 2022). Dabei fällt eine enge Verknüpfung zu den Themenfeldern der Abhängigkeit zwischen den Agierenden und der Regulierung auf, denn die technischen Standards stellen ein bindendes und festgelegtes Element zwischen den Agierenden dar. Allerdings dienen technischen Standards in erster Linie der technischen Zusammenarbeit und der Förderung der Interoperabilität und nicht der Durchsetzung von Regularien.

Fazit

Diskussion

Innerhalb dieser Arbeit werden die Orchestrierungsaktivitäten während der gemeinsamen Wertschöpfungsfindung im Zuge der Initialisierung eines digitalen plattformbasierten Ökosystems betrachtet. Dabei konnten durch eine systematische Literaturanalyse die drei Kategorien Orchestrierung, Wertschöpfung und technische Aspekte abgeleitet werden, die einen Rahmen um die verschiedenen Aktivitäten innerhalb des Ökosystems bilden. In Bezug auf die Forschungsfrage zeigen die Ergebnisse, dass nicht ein einheitlicher und damit bestmöglicher Orchestrierungsansatz existiert. Viel mehr setzt sich ein effizienter Orchestrierungsansatz daraus zusammen, dass Bestandteile aus allen drei Kategorien extrahiert, individuell ausgestaltet und anschließend zu einem Ansatz aggregiert werden.

Maßnahmen, die in Verbindung zur Koordinierung der Zusammenarbeit stehen, fallen unter die Kategorie Orchestrierung. Es zeigt sich, dass aufgrund unterschiedlicher Ökosystemvoraussetzungen zwar keine einheitlichen Rollen erkennbar sind, aber dennoch eine definierte Aufgabenverteilung stattfindet. So sind beispielsweise Plattformbesitzende, die einen großen Anteil der Aufgaben übernehmen, für die progressive Suche nach neuen Teilnehmenden verantwortlich und besitzen dementsprechend auch mehr Macht im Ökosystem. Gleichzeitig sind Plattformbesitzende oft auch treibende Kraft während der Initialisierungsphase, in welcher diese von Komplementierenden unterstützt wird, deren Rolle z.B. Start-Ups einnehmen.

Ein weiterer Bestandteil der Orchestrierung stellt die Gestaltung des Ökosystems dar. Diese umfasst unter anderem die Reduzierung von Hürden während des Eintritts neuer Komplementierenden, gemeinsame Innovationen und die Ökosystemarchitektur. Dabei besteht das Hauptproblem der Orchestrierenden darin, die unterschiedlichen Ziele aller Teilnehmenden anzugleichen, wofür es einer besonderen Ökosystemkultur und vor allem einer gemeinsamen Vision bedarf. Sie definiert gemeinsame Werte und Ziele, stärkt damit das Vertrauen ineinander und wird der dynamischen Umwelt zu Beginn eines Ökosystems gerecht. Nichtsdestotrotz sind Regulierungen notwendig, um Eigentum zu schützen und die Qualität des Ökosystems zu gewährleisten.

Der zweite Bereich, die Wertschöpfung, umfasst überwiegend Aktivitäten bezüglich des Werte-versprechens. Dieses setzt sich zusammen aus den Synergien der Teilnehmenden und deren Abhängigkeiten zueinander sowie den Netzwerkeffekten. Begleitet wird es durch einen effizienten Informationsaustausch und die Nutzung der durch die Plattform bereitgestellten Infrastruktur.

In der Initialisierungsphase des Ökosystems ist das Werteversprechen noch sehr dynamisch aufgrund der instabilen Nutzendenbasis. Zur Überwindung der Instabilität ist eine gewisse Flexibilität der Plattformbesitzenden und ein ausgeprägtes Nutzendenverständnis erforderlich, um die Attraktivität des Ökosystems zu steigern. Die Attraktivität wird dabei durch die Etablierung einer Ökosystemvision unterstützt, welche die Festigung der Identität des Ökosystems und damit die Schaffung eines Markenimages ermöglicht. Denn ein attraktives Markenimage ermöglicht weiteren Wert zu generieren. Ergänzt und weiterentwickelt wird das Wertangebot durch neue Komplementierende, deren Suche auf den Bedarf, d.h. entweder die Vertiefung oder die Diversifizierung des Wertangebots, abgestimmt wird.

Um einen Orchestrierungsansatz zu vervollständigen, bedarf es neben der Koordination der Zusammenarbeit und der Entwicklung eines gemeinsamen Werteversprechens zusätzlich einer Orchestrierung bzgl. technischer Aspekte. Diese bilden das Fundament des Ökosystems und stellen damit eine zentrale Rolle dar. Für die Plattformtechnologie, als Teil des Werteversprechens, wird ein Trade-Off getroffen zwischen Senkung der Koordinierungskosten (hohe Modularität) und Erhöhung der Innovationskraft (geringe Modularität).

Technische Ressourcen werden primär durch technische Standards aufeinander abgestimmt. Zuvor ist jedoch die Implementierung von Grenzressourcen von essenzieller Bedeutung, da diese die Schnittstellen regulieren und den Eintritt neuer Teilnehmenden ermöglichen. Das Ausmaß und die Art der Grenzressourcen werden dabei über den Grad der Offenheit geregelt, der sich beispielsweise darin widerspiegelt, inwieweit Open-Source Anwendungen verwendet werden.

Nachdem innerhalb der initialen Phase des Ökosystems die ermittelte Schnittmenge der drei Bereiche den individuellen Orchestrierungsansatz definieren, entwickelt sich dieser in der nächsten Lebensphase, der Expansionsphase, weiter.

Implikationen

Die Einteilung der Facetten eines Orchestrierungsansatzes in drei Kategorien bietet einen strukturierten Ansatz, der durch bestehende Literatur gestützt wird (Autio, 2022). Gleichzeitig führen die Erkenntnisse dieser Arbeit die bisherigen Forschungsansätze weiter aus, indem verschiedene Aspekte innerhalb dieser Kategorien durch die identifizierten Themenfelder näher konkretisiert und Interdependenzen zwischen den Kategorien aufgezeigt werden. Beispielsweise ist nicht nur die Plattformarchitektur als technische Komponente relevant, sondern aufgrund dessen Einbettung in das Ökosystems auch die Gestaltung des gesamten Ökosystems ein Bestandteil der Orchestrierung. Als weiteres Beispiel kann die Rollenverteilung bezüglich der Wertgenerierung genannt werden, die eine Schnittstelle zwischen Orchestrierung und Wertschöpfung darstellt. Da die Thematik zu möglichen Interdependenzen allerdings nur bedingt innerhalb dieser Arbeit untersucht werden konnte, sollte weiterführende Forschung die Wechselwirkungen zwischen den Kategorien und deren Auswirkungen näher untersuchen.

In der Praxis ermöglicht der aufgezeigte Ansatz den Orchestrierungsaktivitäten innerhalb eines dynamischen digitalen plattformbasierten Ökosystems Struktur zu verleihen. Dies erlaubt es den Teilnehmenden, die Herausforderung eines kollektiven Werteversprechens strukturiert zu bewältigen, indem sie zuerst Praktiken zur Wertgenerierung bestimmen und anschließend die Eigenschaften des Werteversprechens charakterisieren und kontinuierlich weiterentwickeln. Um die verschiedenen Ziele in einem Ökosystem auszurichten, sollten die Teilnehmenden zudem eine gemeinsame Ökosystemvision entwickeln, die eine Identifikation mit dem Ökosystem ermöglicht und dadurch die Kollektivität des Werteversprechens unterstützt.

Da das Werteversprechen in das Ökosystem eingebettet ist, sind weitere Aktivitäten zur Orchestrierung rund um das Werteversprechen notwendig. Durch eine klare Rollenverteilung, unterstützt durch geeignete Regulierungen, lassen sich Zuständigkeiten und Arbeitsweisen definieren, was die klassische Vertragskonstellation ersetzt. Eine ausgeprägte Ökosystemkultur kann die kollaborative Arbeitsweise zusätzlich fördern.

Ein zentrales praktisches Problem innerhalb eines digitalen plattformbasierten Ökosystems besteht in der Gewinnung neuer Teilnehmenden. Dies erfordert nicht nur Orchestrierungsaktivitäten zur Steigerung der Attraktivität des Ökosystems und zur Optimierung der Teilnehmendensuche, sondern auch die Berücksichtigung technischer Aspekte, die das Fundament bilden. Denn ohne entsprechende Grenzressourcen ist es neuen Teilnehmenden nicht möglich dem Ökosystem beizutreten. Dementsprechend sind Grenzressourcen ein kritischer Faktor für die erfolgreiche Initialisierung und Erweiterung eines digitalen plattformbasierten Ökosystems.

Limitationen

An ihre Grenzen stoßen die Ergebnisse in Bezug auf eine zeitliche Abgrenzung der verschiedenen Entwicklungsstadien von Ökosystemen. Denn die Phase der Initialisierung selbst ist nicht einheitlich abgegrenzt und der Übergang zur Expansionsphase stellt sich als fließend heraus. Zusätzlich können sich

Aktivitäten über mehrere Phasen erstrecken, wodurch manche Ergebnisse die Initialisierung überschreiten, andere wiederum langfristige Effekte nicht abbilden können. Weiter zeigen die Ergebnisse ein gewisses Ungleichgewicht innerhalb der Literatur bezüglich der Orchestrierung von Rollen, sodass überwiegend die Rolle der Plattformbesitzenden eine zentrale Position einnimmt und die Bedeutung anderer Agierender vernachlässigt wird. Dies hängt auch mit den Grenzen der strukturierten Literaturrecherche zusammen, da sie keine vollständige Abdeckung der Literatur garantieren kann und nur eine gute Approximation darstellt. Zudem reagiert sie sensibel auf kleine Änderungen im Suchstring, wodurch teilweise relevante Literatur unbemerkt ausgeschlossen werden könnte. Für die Codierung der thematischen Analyse gilt im Idealfall, dass auf jeder Abstraktionsebene alle Themen das gleiche Niveau erreichen, um die Transparenz der logischen Struktur zu gewährleisten. In der Praxis ist das allerdings nahezu unmöglich und stellt somit auch hier ausschließlich eine Approximation dar.

Ausblick

Wie im vorherigen Abschnitt angesprochen, benötigen vernachlässigte Ökosystemkonstellationen bzgl. der Rollenverteilung der Teilnehmenden weitere Untersuchungen. Denn nicht immer muss die Orchestrierung und das damit einhergehende Machtgefälle eindeutig von einer zentralen orchestrierenden Partei ausgehen, sondern es sind auch Ökosystemkonstellationen denkbar, welche die Aufgaben und Herausforderungen über mehrere oder alle Agierende aufteilen. Hierbei stellt sich die Frage, inwiefern sich in einem solchen Fall ein Orchestrierungsansatz von den in dieser Arbeit untersuchten Ansätzen unterscheidet.

Hinzu kommt die Fragestellung, inwiefern sich schnell entwickelnde disruptive Technologien, wie z.B. künstliche Intelligenz, auf die in dieser Arbeit genannten Themenfelder unter Beachtung der speziellen Anforderungen plattformbasierter Ökosysteme, im Vergleich zu anderen Ökosystemen, auswirken können. Insbesondere, da Daten einen immer höheren Stellenwert auf einer Plattform einnehmen und es damit zu einer weiteren Verschiebung von Machtanteilen in Richtung der Plattformbesitzenden kommen könnte (Clough & Wu, 2022), ist es von zentraler Bedeutung ein Verständnis der Wechselwirkungen zwischen der Technologie- und der Ökosystementwicklung zu gewinnen.

Literaturverzeichnis

Acs, Z. J., Song, A. K., Szerb, L., Audretsch, D. B., & Komlósi, É. (2021). The evolution of the global digital platform economy: 1971–2021. Small Business Economics, 57(4), 1629–1659. **https://doi.org/10.1007/s11187-021-00561-x**

Ambrasaitė, P., & Smagurauskaitė, A. (2021). Epic Games v. Apple: Fortnite battle that can change the industry. Vilnius University Open Series, 6–25. **https://doi.org/10.15388/TMP.2021.1**

Autio, E. (2022). Orchestrating ecosystems: A multi-layered framework. Innovation, 24(1), 96–109. **https://doi.org/10.1080/14479338.2021.1919120**

Ben Letaifa, S. (2014). The uneasy transition from supply chains to ecosystems. Management Decision, 52(2), 278–295. **https://doi.org/10.1108/MD-06-2013-0329**

Benlian, A., Hilkert, D., & Hess, T. (2015). How open is this Platform? The Meaning and Measurement of Platform Openness from the Complementers' Perspective. Journal of Information Technology, 30(3), 209–228. **https://doi.org/10.1057/jit.2015.6**

Biedebach, M., & Hanelt, A. (2020). Towards a Typology of Ecosystem Roles in the Era of Digital Innovation - An Inductive Empirical Analysis. ICIS 2020 Proceedings, Artikel 7. **https://aisel.aisnet.org/icis2020/general_topics/general_topics/7**

Birkinshaw, J. (2018). How is technological change affecting the nature of the corporation? Journal of the British Academy, 6(s1), 185–214. **https://doi.org/10.5871/jba/006s1.185**

Blaschke, M., Haki, K., Aier, S., & Winter, R. (2018). Capabilities for Digital Platform Survival: Insights from a Business-to-Business Digital Platform. ICIS 2018 Proceedings, Artikel 2. **https://aisel.aisnet.org/icis2018/service/Presentations/2**

Cennamo, C., Ozalp, H., & Kretschmer, T. (2018). Platform Architecture and Quality Trade-offs of Multihoming Complements. Information Systems Research, 29(2), 461–478. **https://doi.org/10.1287/isre.2018.0779**

Chen, L., Shaheer, N., Yi, J., & Li, S. (2019). The international penetration of ibusiness firms: Network effects, liabilities of outsidership and country clout. Journal of International Business Studies, 50(2), 172–192. **https://doi.org/10.1057/s41267-018-0176-2**

Clough, D. R., & Wu, A. (2022). Artificial Intelligence, Data-Driven Learning, and the Decentralized Structure of Platform Ecosystems. Academy of Management Review, 47(1), 184–189. **https://doi.org/10.5465/amr.2020.0222**

de Vasconcelos Gomes, L. A., Facin, A. L. F., Leal, L. F., Zancul, E. de S., Salerno, M. S., & Borini, F. M. (2022). The emergence of the ecosystem management function in B2B firms. Industrial Marketing Management, 102, 465–487. **https://doi.org/10.1016/j.indmarman.2021.12.015**

Ghazawneh, A., & Henfridsson, O. (2015). A Paradigmatic Analysis of Digital Application Marketplaces. Journal of Information Technology, 30(3), 198–208. **https://doi.org/10.1057/jit.2015.16**

Gulati, R., Puranam, P., & Tushman, M. (2012). Meta-organization design: Rethinking design in interorganizational and community contexts. Strategic Management Journal, 33(6), 571–586. **https://doi.org/10.1002/smj.1975**

Han, J., Zhou, H., Lowik, S., & de Weerd-Nederhof, P. (2022). Enhancing the understanding of ecosystems under innovation management context: Aggregating conceptual boundaries of ecosystems. Industrial Marketing Management, 106, 112–138. **https://doi.org/10.1016/j.indmarman.2022.08.008**

Helfat, C. E., & Raubitschek, R. (2018). Dynamic and Integrative Capabilities for Profiting From Innovation in Digital Platform-Based Ecosystems. SSRN Electronic Journal. **https://doi.org/10.2139/ssrn.3122046**

Hodapp, D., Hawlitschek, F., & Kramer, D. (2019). Value Co-Creation in Nascent Platform Ecosystems: A Delphi Study in the Context of the Internet of Things. Proceedings of the 40th International Conference on Information Systems (ICIS). Artikel 1838.

Huang, P., Ceccagnoli, M., Forman, C., & Wu, D. J. (2013). Appropriability Mechanisms and the Platform Partnership Decision: Evidence from Enterprise Software. Management Science, 59(1), 102–121. **https://doi.org/10.1287/mnsc.1120.1618**

Iansiti, M., & Levien, R. (2004). Strategy as Ecology. Harvard Business Review, 82(3), 68–78. **https://www.ncbi.nlm.nih.gov/pubmed/15029791**

Iyer, B., Lee, C.-H., & Venkatraman, N. (2006). Managing in a "Small World Ecosystem": Lessons from the Software Sector. California Management Review, 48(3), 28–47. **https://doi.org/10.2307/41166348**

Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. Strategic Management Journal, 39(8), 2255–2276. **https://doi.org/10.1002/smj.2904**

Jacobides, M. G., Cennamo, C., & Gawer, A. (2024). Externalities and complementarities in platforms and ecosystems: From structural solutions to endogenous failures. Research Policy, 53(1), Artikel 104906. **https://doi.org/10.1016/j.respol.2023.104906**

Kapoor, K., Ziaee Bigdeli, A., Dwivedi, Y. K., Schroeder, A., Beltagui, A., & Baines, T. (2021). A socio-technical view of platform ecosystems: Systematic review and research agenda. Journal of Business Research, 128, 94–108. **https://doi.org/10.1016/j.jbusres.2021.01.060**

Kohtamäki, M., Parida, V., Patel, P. C., & Gebauer, H. (2020). The relationship between digitalization and servitization: The role of servitization in capturing the financial potential of digitalization. Technological Forecasting and Social Change, 151, Artikel 119804. **https://doi.org/10.1016/j.techfore.2019.119804**

Lechowski, G., & Krzywdzinski, M. (2022). Emerging positions of German firms in the industrial internet of things: A global technological ecosystem perspective. Global Networks, 22(4), 666–683. **https://doi.org/10.1111/glob.12380**

Linde, L., Sjödin, D., Parida, V., & Wincent, J. (2021). Dynamic capabilities for ecosystem orchestration A capability-based framework for smart city innovation initiatives. Technological Forecasting and Social Change, 166, 120614. **https://doi.org/10.1016/j.techfore.2021.120614**

Liu, G., & Rong, K. (2015). The Nature of the Co-Evolutionary Process: Complex Product Development in the Mobile Computing Industry's Business Ecosystem. Group & Organization Management, 40(6), 809–842. **https://doi.org/10.1177/1059601115593830**

Mbanefo, C., & Saartijie Grobbelaar, S. S. (2022). A Systematic Review of Concepts and Future Directions of Platform Ecosystem Development. 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference, 1–11. **https://doi.org/10.1109/ICE/ITMC-IAMOT55089.2022.10033234**

McIntyre, D. P., & Srinivasan, A. (2017). Networks, platforms, and strategy: Emerging views and next steps. Strategic Management Journal, 38(1), 141–160. **https://doi.org/10.1002/smj.2596**

Miehé, L., Palmié, M., & Oghazi, P. (2023). Connection successfully established: How complementors use connectivity technologies to join existing ecosystems – Four archetype strategies from the mobility sector. Technovation, 122, Artikel 102660. **https://doi.org/10.1016/j.technovation.2022.102660**

Mukhopadhyay, S., & Bouwman, H. (2019). Orchestration and governance in digital platform ecosystems: A literature review and trends. Digital Policy, Regulation and Governance, 21(4), 329–351. **https://doi.org/10.1108/DPRG-11-2018-0067**

Murthy, R. K., & Madhok, A. (2021). Overcoming the Early–stage Conundrum of Digital Platform Ecosystem Emergence: A Problem–Solving Perspective. Journal of Management Studies, 58(7), 1899–1932. **https://doi.org/10.1111/joms.12748**

Nambisan, S., Zahra, S. A., & Luo, Y. (2019). Global platforms and ecosystems: Implications for international business theories. Journal of International Business Studies, 50(9), 1464–1486. **https://doi.org/10.1057/s41267-019-00262-4**

Nerbel, J. F., & Kreutzer, M. (2023). Digital platform ecosystems in flux: From proprietary digital platforms to wide-spanning ecosystems. Electronic Markets, 33(1), 6. **https://doi.org/10.1007/s12525-023-00625-8**

Ofe, H., & Sandberg, J. (2019). Digital Platform Establishment: Navigating Competing Concerns in Emerging Ecosystems. Proceedings of the 52nd Hawaii International Conference on System Sciences, 52, 1425–1434. **https://doi.org/10.24251/HICSS.2019.173**

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2019). Research methods for business students (Eighth Edition). Pearson.

Schreieck, M., Wiesche, M., & Krcmar, H. (2021). Capabilities for value co-creation and value capture in emergent platform ecosystems: A longitudinal case study of SAP's cloud platform. Journal of Information Technology, 36(4), 365–390. **https://doi.org/10.1177/02683962211023780**

Singaraju, S. P., Nguyen, Q. A., Niininen, O., & Sullivan-Mort, G. (2016). Social media and value co-creation in multi-stakeholder systems: A resource integration approach. Industrial Marketing Management, 54, 44–55. **https://doi.org/10.1016/j.indmarman.2015.12.009**

Snihur, Y., Thomas, L. D. W., & Burgelman, R. A. (2018). An Ecosystem-Level Process Model of Business Model Disruption: The Disruptor's Gambit. Journal of Management Studies, 55(7), 1278–1316. **https://doi.org/10.1111/joms.12343**

Stonig, J., Schmid, T., & Müller–Stewens, G. (2022). From product system to ecosystem: How firms adapt to provide an integrated value proposition. Strategic Management Journal, 43(9), 1927–1957. **https://doi.org/10.1002/smj.3390**

Thomas, L. D. W., Autio, E., & Gann, D. M. (2022). Processes of ecosystem emergence. Technovation, 115, 102441. **https://doi.org/10.1016/j.technovation.2021.102441**

Tiwana, A. (2014). Platform ecosystems: Aligning architecture, governance, and strategy. Morgan Kaufmann.

Tsai, C. L., Ahn, J. M., & Mortara, L. (2022). Managing platform-based ecosystems in B2B markets – outbound open innovation perspective. International Journal of Technology Management, 89(3/4), 139. **https://doi.org/10.1504/IJTM.2022.123722**

Ulrich, K. (1995). The role of product architecture in the manufacturing firm. Research Policy, 24(3), 419–440. **https://doi.org/10.1016/0048-7333(94)00775-3**

van Vulpen, P., Jansen, S., & Brinkkemper, S. (2022). The orchestrator's partner management framework for software ecosystems. Science of Computer Programming, 213, Artikel 102722. **https://doi.org/10.1016/j.scico.2021.102722**

Weiss, N., Wiesche, M., Schreieck, M., & Krcmar, H. (2022). Learning to be a Platform Owner: How BMW Enhances App Development for Cars. IEEE Transactions on Engineering Management, 69(6), 4019–4035. **https://doi.org/10.1109/TEM.2020.3017051**

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. European Journal of Information Systems, 22(1), 45–55. **https://doi.org/10.1057/ejis.2011.51**

Wulfert, T., Woroch, R., Strobel, G., Seufert, S., & Möller, F. (2022). Developing design principles to standardize e-commerce ecosystems: A systematic literature review and multi-case study of boundary resources. Electronic markets, 32(4), 1813–1842. **https://doi.org/10.1007/s12525-022-00558-8**

Xu, Y., Sun, H., & Lyu, X. (2023). Analysis of decision–making for value co–creation in digital innovation systems: An evolutionary game model of complex networks. Managerial and Decision Economics, 44(5), 2869–2884. **https://doi.org/10.1002/mde.3852**

Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research. Information Systems Research, 21(4), 724–735. **https://doi.org/10.1287/isre.1100.0322**

Zeng, J., Yang, Y., & Lee, S. H. (2023). Resource Orchestration and Scaling–up of Platform–Based Entrepreneurial Firms: The Logic of Dialectic Tuning. Journal of Management Studies, 60(3), 605–638. **https://doi.org/10.1111/joms.12854**

The Social and Ethical Impacts of the Metaverse on the Younger Generation

Selected Issues in Critical Information Infrastructure, Summer Term 2023

Elena Fantino
Master Student
Karlsruhe Institute of Technology
uyqel@student.kit.edu

Katharina Fischer
Master Student
Karlsruhe Institute of Technology
k.h.s.fischer@gmail.com

Linda Günder
Master Student
Karlsruhe Institute of Technology
l.guender@gmail.com

Ahmed Zekri
Master Student
Karlsruhe Institute of Technology
ahmed.zekri@student.kit.edu

Abstract

Background: *The rapid rise of the "Metaverse" could profoundly influence interpersonal interactions, especially for the younger generation, who are already growing up in a digitally influenced world. This study aims to analyze the social and ethical implications of the Metaverse on youth, addressing both its potential benefits such as enhanced connectivity and creativity, as well as concerns like addiction risks and privacy issues.*

Objective: *This paper examines the advantages and risks of increased use of the Metaverse for young people. It begins with an overview of existing research, followed by an in-depth analysis of the pros and cons.*

Methods: *This study outlines a systematic method for reviewing literature on the social and ethical implications of the metaverse for the younger generation. To identify relevant studies and articles an extensive internet search was conducted. Various databases including EBSCO and the Wiley Online Library were used to capture a wide range of scholarly publications. The search strategy combined keywords such as 'metaverse', 'young people', 'youth' and 'adolescents' with specific themes such as education, social interaction and psychological effects. Following the title and abstract screening, relevant articles were selected and subjected to a thorough full-text analysis, to ensure inclusion of qualitative and thematically relevant research.*

Results: *Enhanced networking opportunities, from which young people already benefited during the Covid-19 pandemic, are identified as advantages. Visual learning content allows for increased absorption capacity for young people, and personal development is promoted through creative self-expression. Additionally, the Metaverse could become more accessible and inclusive than the real world through the right measures. However, dependency and addiction are identified as health risks for young people. The loss of connection to reality also poses a serious risk. The reduction of social inhibitions, on the other hand, can be both an advantage and a risk. It is necessary to regulate platform operators and establish universal standards for providers to ensure*

data protection and privacy. It is also concluded that the focus on the younger generation must be deepened while simultaneously examining the impacts on other generations.

Conclusion: *This paper provides a basis for the development of appropriate measures to prevent the negative effects of the Metaverse and to optimally utilize the positive aspects. The goal is to enable a safe, transparent, and inclusive future in a digital world for young people.*

Keywords: Metaverse, digital influence, young generation, interpersonal interactions, networking opportunities, visual learning, personal development, data protection, addiction risks, accessibility

Einleitung

Das rasante Wachstum der digitalen Technologie hat eine neue Ära eingeläutet, die das Potential hat, unsere Gesellschaft und unser zwischenmenschliches Miteinander durch das Eintauchen in eine virtuelle Welt grundlegend zu verändern. Dabei tritt das sogenannte "Metaverse" immer stärker in den Fokus. Es repräsentiert eine virtuelle, erweiterte Realität, in der Nutzende in eine vernetzte digitale Welt eintauchen können (Luber, 2022).

Das Metaverse verankert sich zunehmend im alltäglichen Leben, welches z.B. auch durch die steigende Beliebtheit von VR-Brillen (Rinaldi, 2022) deutlich wird. In den sozialen Medien wie Facebook, Instagram, Snapchat und Co. werden bereits mit Avataren, also „fiktiven, softwarebasierten Bildschirmgestalten“ (Kollmann, 2018), erste Berührpunkte mit dem Metaverse aufgebaut. Auch Videospiele und Gaming-Plattformen entwickeln sich immer stärker in Richtung Metaverse und ermöglichen den Nutzenden, durch die Spielerfahrung mit allen Sinnen, noch tiefer in virtuelle Räume einzutauchen (Wieser, 2022). Die wachsende Popularität des Metaverse erinnert an die rapide Entwicklung des Trends von sozialen Medien. Innerhalb weniger Jahre sind soziale Medien zu einem festen Bestandteil des täglichen Lebens geworden und haben nun großen Einfluss auf die Gesellschaft (Ortiz-Ospina, 2019). Das Metaverse und soziale Medien weisen in einigen Punkten große Ähnlichkeiten auf. Auf beiden Plattformen werden digitale Inhalte konsumiert und digitale Interaktion und Vernetzung spielt jeweils eine große Rolle. Es wird vermutet, dass das Metaverse aufgrund seiner tiefen Eintauchmöglichkeit in die digitale Welt mit allen Sinnen noch stärkere Auswirkungen auf die Gesellschaft haben kann. Deshalb ist es notwendig, die sozialen und ethischen Auswirkungen des Metaverse genauer zu untersuchen. Besonders interessant ist die heutige junge Generation, die in einer Zeit aufwächst, in der digitale Technologien und virtuelle Welten allgegenwärtig sind. Diese jungen Menschen könnten sich in naher Zukunft zu den Hauptnutzenden des Metaverse entwickeln (Sortlist - Data Hub, 2023) und werden von seinen Vorzügen und Herausforderungen gleichermaßen beeinflusst. Vorteile, also die positiven Effekte des Metaverse, können z.B. verbesserte Vernetzungsmöglichkeiten zwischen Jugendlichen, neue Bildungs- und Arbeitsformate, die Förderung von Kreativität und Selbstentfaltung oder neue Inklusionsmöglichkeiten sein. Herausforderungen können die Sucht- und Abhängigkeitsgefahren, aber auch Sicherheits- und Privacy-Bedenken sowie soziale Isolation sein.

Daraus ergibt sich die Forschungsfrage: „Welche sozialen und ethischen Auswirkungen hat das Metaverse auf die junge Generation?“ Diese Seminararbeit analysiert diese Fragestellung und trägt somit zu einer Beantwortung bei. Dies ist von großer Relevanz, da entsprechende Antworten dazu beitragen können, angemessene Maßnahmen zu entwickeln, um potenzielle negative Effekte des Metaverse zu minimieren und die positiven Aspekte optimal zu nutzen. Dabei werden zum einen die positiven und zum anderen die negativen Konsequenzen des Metaverse betrachtet. Durch eine Analyse dieser Themenfelder kann ein umfassendes Bild der sozialen und ethischen Auswirkungen des Metaverse auf die junge Generation gezeichnet werden. Zunächst folgt eine Übersicht zum Thema Metaverse sowie weitere Begriffsdefinition, die in dieser Arbeit genutzt werden. Im Weiteren wird die Methodik erläutert, gefolgt von einem Hauptteil, in dem Vor- und Nachteile diskutiert werden. Schließlich werden die Ergebnisse zusammengefasst und die Arbeit mit den Limitationen und dem weiteren Forschungsbedarf abgeschlossen.

Metaverse im Überblick und weitere Definitionen

Die Arbeit befasst sich mit dem Metaverse und den daraus folgenden Implikationen für die junge Generation. Nachfolgend werden die Begriffe „Metaversum", „Augmented Reality", „Virtual Reality" und „virtuelle Welten" als Synonyme für Metaverse benutzt.

Der Begriff „Metaverse" ist eine Zusammensetzung aus den Wörtern „Meta", was Transzendenz und Virtualität, und „Verse", was Welt und Universum bedeutet (Kye et al. 2021, S. 1) und wurde von Neal Stephenson in seinem Science-Fiction-Roman mit dem Titel „Snow Crash" aus dem Jahr 1992 eingeführt (Wang et al. 2023a, S. 1). In diesem Roman treten die Menschen der physischen Welt in das Metaverse über Virtual-Reality-Geräte ein und beginnen ein Leben dort. Das Metaverse ist das postreale Universum, eine immerwährende und beständige Multiuser-Umgebung, in der die physische Realität mit der digitalen Virtualität fusioniert (Mystakidis, 2022, S. 1). Es basiert auf der Konvergenz von Technologien, die multisensorische Interaktionen mit virtuellen Umgebungen, digitalen Objekten und Menschen ermöglichen, wie z.B. Virtual Reality (VR) und Augmented Reality (AR) (Mystakidis 2022, S. 1). VR ist eine alternative, völlig eigenständige, digital geschaffene künstliche Umgebung, die computergestützte Technologien nutzt, um realistische (oder auch nicht realistische) Anwendungen zu schaffen und zu simulieren (Pellas et al. 2020, S. 5; Pellas et al. 2021, S. 2). Die Nutzenden fühlen sich in der VR in eine andere Welt versetzt und agieren in ähnlicher Weise wie in der physischen Umgebung (Slater und Sanchez-Vives 2016, S. 2). Jeder Benutzende hat die Illusion „vor Ort" zu sein, umgeben von einer dreidimensionalen (3D) Umgebung in einem 360-Grad-Umfeld, welches er frei erkunden, dort mit visuellen Objekten interagieren und an praktischen Versuchen teilnehmen kann (Pellas et al. 2021, S. 1; Pellas et al. 2020, S. 2). AR hingegen wendet eine einzigartige Strategie für physische Räume an, indem sie digitale Eingaben und virtuelle Elemente in die reale Umgebung integriert und diese dadurch bereichert. Diese Technologie verschmilzt nahtlos die physische und die virtuelle Welt in einem räumlichen Kontext (Mystakidis 2022, S. 2). Eine weitere wichtige Entität im Metaverse ist der Avatar (Park & Kim, 2022, S. 10). Die Gestaltung einer Online-Identität innerhalb des Metaverse wird in erster Linie durch die Schaffung einer digitalen Repräsentation der eigenen Person durch Avatare in der virtuellen Welt erreicht. Diese ermöglichen dabei ein besseres Selbstgefühl, da die Nutzenden ihre Avatare kontrollieren (Messinger et al., 2008, S. 3). Die Nutzenden haben die kreative Freiheit, ihre Avatare zu personalisieren, was ihnen ein breites Spektrum der Selbstdarstellung ermöglicht. Diese Selbstrepräsentationen können von realistischen menschenähnlichen Darstellungen bis hin zu völlig fantasievollen Formen reichen (Mystakidis, 2022, S. 6).

In den letzten Jahren hat das Konzept des Metaversums aufgrund seines Anspruchs auf Realitätsnähe und des gleichzeitigen technologischen Fortschritts stark an Bedeutung gewonnen. Die angestrebte Verschmelzung von physischen und virtuellen Welten innerhalb eines digitalen Realitätsrahmens hat bei globalen Unternehmen, Technologiefans und der allgemeinen Bevölkerung große Aufmerksamkeit erregt (Y. Wang et al., 2023, S. 1). Einige der führenden Tech-Giganten haben bereits das Potential des Metaverse erkannt und ihren Einstieg angekündigt (Y. Wang et al., 2023, S. 1). Allen voran ist Facebook, das kürzlich seine Neuausrichtung und Umbenennung in "Meta" bekanntgegeben hat. Dieser Schritt verdeutlicht das Engagement des Unternehmens für den Aufbau und die Gestaltung des zukünftigen Metaverse (Isaac, 2021; Y. Wang et al., 2023, S. 1). Meta hat bereits bedeutende Investitionen getätigt und plant, seine bestehenden Technologien wie VR und AR weiterzuentwickeln, um die Grundlagen für das Metaverse zu schaffen (Bosworth, 2022).

Aber Meta ist nicht allein. Andere Technologieriesen wie Microsoft, Tencent und NVIDIA haben ebenfalls angekündigt, in das Metaverse einzusteigen (Ball, 2022). Microsoft hat bereits seine Mixed-Reality-Plattform "Microsoft Mesh" vorgestellt, die auf den Prinzipien des Metaverse basiert. Mit Microsoft Mesh können Unternehmen, benutzendendefinierte, realitätsnahe Erlebnisse schaffen, die eine neue Art der Kommunikation ermöglichen, beispielsweise können Reiseunternehmen digitale Besichtigungen von Sehenswürdigkeiten anbieten (Kubach, 2021). Zudem haben Microsoft Teams-Nutzende die Möglichkeit, Microsoft Mesh nahtlos zu integrieren, wodurch Teams-Meetings in Mixed-Reality-Umgebungen stattfinden können. Die Teilnehmenden können an diesen Meetings mit ihren Microsoft Mesh-Avataren teilnehmen, ohne ihre Kameras aktivieren zu müssen (Pleasant, 2023).

Ein weiterer Aspekt, der die Bedeutung des Metaversums verdeutlicht, liegt in der Analyse des Gartner Hype Cycles. Gartner, das als eines der weltweit führenden Unternehmen auf dem Gebiet der Unternehmensanalyse gilt, veröffentlicht jedes Jahr eine grafische Darstellung, die transformative und

bahnbrechende Technologien beschreibt, die einer aufmerksamen Prüfung bedürfen, und die als "Hype Cycle" bezeichnet wird (Skarredghost, 2022). Im August 2022 wurde im Rahmen des Gartner Hype Cycle 2022 auf mehrere immersive Technologien hingewiesen, insbesondere auf das Metaverse (Skarredghost, 2022). Gartner prognostiziert, dass bis zum Jahr 2026 schätzungsweise 30 Prozent der Unternehmen Produkte und Dienstleistungen entwickelt haben werden, die sich in das Metaverse integrieren lassen (Gartner, 2022a). Darüber hinaus deuten Prognosen darauf hin, dass der gesamte adressierbare Markt für das Metaverse bis 2030 eine beachtliche Spanne von 8 bis 13 Billionen US-Dollar umfassen könnte, während die Metaverse-Nutzendenbasis nach Erkenntnissen der Citi etwa 5 Milliarden Menschen umfassen könnte (Citi GPS, 2022; Lasserre, 2022).

Die Vielfalt des Metaverse kommt bereits in unterschiedlichen Formen und Anwendungen zum Ausdruck (L.-H. Lee et al., 2021, S. 34). Neben Simulationen, die das reale Leben detailgetreu nachahmen, manifestieren sich schon heute vielseitige und für jeden zugängliche Anwendungen innerhalb dieses digitalen Universums (Park & Kim, 2022, S. 16).

Vor allem im Bereich der Unterhaltung ist das Metaverse äußerst dynamisch. Spiele spielen dabei eine führende Rolle, um das Metaverse einem breiteren Publikum bekannt zu machen. Durch die Erfassung von Echtzeitnutzendenbewegungen in spielerischen Umgebungen können komplexe Bewegungsabläufe ausgeführt werden, die zur Selbstverbesserung und Förderung des Wettbewerbs anregen. Ein prominentes Beispiel dafür ist die Spielplattform "Roblox", die am 10. März 2021 erstmals das Konzept des Metaverse in ihr Prospekt aufnahm und erfolgreich an der New Yorker Börse gelistet wurde (H. Wang et al., 2023, S. 3). So umfasst das Roblox Metaverse eine Vielzahl virtueller Spiele und immersiver Umgebungen. Zusätzlich zum Spielen können Nutzende an verschiedenen Veranstaltungen teilnehmen, wie zum Beispiel Konzerten und Zeremonien (Ibrahim, 2022). Darüber hinaus ermöglicht Roblox den Nutzenden, sich voll und ganz auszudrücken, indem sie ihre Avatare individuell gestalten und digitale Accessoires erwerben, darunter Luxusartikel von Marken wie Gucci (Ibrahim, 2022).

Diese Arbeit untersucht den potenziellen Einfluss des Metaversums auf junge Menschen im Alter von 15-30 Jahren. Der Begriff "junge Generation" umfasst in diesem Zusammenhang Jugendliche und junge Erwachsene. Eine gemeinsame Generation teilt oft ähnliche Erfahrungen, Einstellungen und Verhaltensweisen, wodurch sie eine ähnliche Wertekultur entwickeln (Liebsch, 2020).

Der Fokus auf diese Gruppe ist aus mehreren Gründen wichtig. Zum einen gehören junge Menschen zu den ersten Anwendenden des Metaverse. Die jüngeren Generationen verbringen immer mehr Zeit auf sozialen Plattformen. So verbringt die „Generation Z“, das sind Menschen, die zwischen 1997 und 2012 geboren wurden, etwa 2 Stunden und 55 Minuten pro Tag in den sozialen Medien (Oh, Kim, Chang, Park & Lee, 2023, S. 1; World Economic Forum, 2019). Außerdem stellen jüngere Menschen den größten Teil der Bevölkerung im Metaverse dar, so sind z.B. etwa 83% der Nutzenden auf Roblox unter 25 Jahre alt (Statista, 2022). Auch das Nutzendenengagement ist bei diesem Spiel seit Ende 2018 um sechs Mal gestiegen (Statista, 2023) und beschreibt somit einen wachsenden Trend.

Zum anderen nehmen Heranwachsende bestimmte Aspekte sozialer virtueller Welten im Vergleich zur Offline-Welt als weniger riskant wahr, was interessante Herausforderungen und Möglichkeiten für die Identitätskonstruktion im Metaverse schafft (Maloney, 2021).

Nachdem in diesem Grundlagenteil die wesentlichen Konzepte und Hintergründe zum Metaverse erörtert wurden, soll im folgenden Abschnitt die angewandte Methodik zur Untersuchung der sozialen und ethischen Auswirkungen des Metaverse auf die junge Generation dargelegt werden.

Methodik

Die vorliegende Methodik beschreibt einen systematischen Ansatz zur Durchführung einer Literaturrecherche über die sozialen und ethischen Auswirkungen des Metaverse auf die junge Generation. Die Recherche wurde sowohl vor dem Schreibprozess als auch währenddessen durchgeführt, um sicherzustellen, dass alle relevanten Informationen und Perspektiven abgedeckt werden.

Für die Zusammenstellung von Literatur und Artikeln wurde eine umfangreiche Internetrecherche durchgeführt. Aufgrund der Vielfalt an Suchanfragen im Zusammenhang mit der Forschungsfrage wurden sämtliche verfügbaren Informationsquellen genutzt. Zur Recherche dienten unter anderem die Online-Bibliotheken EBSCO, GoogleScholar und Wiley Online Library, um eine breite Palette relevanter

Veröffentlichungen zu erfassen. Da es sich hierbei um ein breites Themenspektrum handelt, wurden für jede Hypothese alle relevanten Schlüsselwortkombinationen verwendet, wie in Tabelle 1 beschrieben. Die Suchstrategie kombiniert die identifizierten Schlüsselbegriffe unter Verwendung von Booleschen Operatoren. Die verwendeten Schlüsselwörter umfassen stets "Metaverse", "young people", "youth" und "adolescent", um eine Basis für die Suche zu schaffen. Je nach Thematik der aufgestellten Hypothesen wurden zusätzliche Schlüsselbegriffe verwendet, um die benötigten Aspekte abzudecken. Beispiel: ("Metaverse") AND ("young people" OR " youth " OR "adolescent") AND zusätzliche Schlüsselbegriffe. Die Gesamtheit aller Schlüsselwortkombinationen sind in Tabelle 1 vorzufinden.

Nachdem die initiale Suche abgeschlossen wurde, ist das Titel- und Abstract-Screening erfolgt. Dabei wurden die gefundenen Artikel stets anhand ihrer Titel und Abstracts überprüft, um jene auszuschließen, die nicht im Einklang mit dem Forschungsthema stehen. Es wurden die Artikel ausgewählt, die potenziell relevante Informationen über die sozialen und ethischen Auswirkungen des Metaverse auf die junge Generation enthalten könnten. Im gleichen Zuge wurde eine Rückwärtsrecherche durchgeführt. Diese Artikel wurden in der nächsten Phase der Auswertung berücksichtigt. Die im vorherigen Schritt ausgewählten Artikel wurden einer umfassenden Volltext-Analyse unterzogen. In dieser Phase wurden die vollständigen Artikel beschafft und anhand zuvor festgelegter Qualitätskriterien eingehend bewertet. Diese Kriterien umfassten die angewandte Forschungsmethodik, die Relevanz der Ergebnisse und die Überzeugungskraft der Argumentation. Diese kritische Prüfung stellte sicher, dass nur qualitativ hochwertige und thematisch relevante Artikel in die weitere Analyse einbezogen wurden.

Hypothesen	**Keyword Basis**		**Keyword Zusatz**	**Kapitel**
H1: Das Metaverse bietet eine verbesserte Kommunikation und eine größere Vernetzungsmöglichkeit für die junge Generation.	{Metaverse} x	{young people, youth, adolescent} x	{connecting, networking, linking up}	4.1.1
H2: Durch erweiterte Bildungsformate im Metaverse, können junge Menschen besser lernen.	{Metaverse} x	{young people, youth, adolescent} x	{education, learning, studying}	4.1.2
H3: Das Metaverse fördert Kreativität und Selbstentfaltung bei der jungen Generation.	{Metaverse} x	{young people, youth, adolescent} x	{self-discrepancy, creativity}	4.1.3
H4: Durch entsprechende Maßnahmen kann das Metaverse ein frei zugänglicher und inklusiver Raum werden.	{Metaverse} x	{young people, youth, adolescent} x	{inclusion, diversity, equality }	4.1.4
H5: Durch Metaverse besteht einer höhere Sucht- und Abhängigkeitsgefahr bei jungen Menschen.	{Metaverse} x	{young people, youth, adolescent} x	{dependence, addiction}	4.2.1
H6: Metaverse fördert Entfremdung von der Realität und sozialer Isolation.	{Metaverse} x	{young people, youth, adolescent} x	{reality, hyperreality, physical world, alienation, perception, view, attitude}	4.2.2
H7: Metaverse ist eine Gefahr für Sicherheit- und Data Privacy.	{Metaverse} x	{young people, youth, adolescent} x	{privacy, security}	4.2.3
H8: Metaverse ist sehr komplex zu regulieren und hat Bedarf Verbraucherschutz.	{Metaverse} x	{young people, youth, adolescent} x	{regulation, safe, policies}	4.2.4

Tabelle 1: Übersicht der Hypothesen und Keywords

Die Suchparametrierung wurde bewusst offen gestaltet und nicht auf einen definierten zeitlichen Rahmen begrenzt. Die verwendeten Suchbegriffe wurden gezielt so ausgewählt, dass sie im Titel der abgefassten Arbeiten auftauchen. Die Anzahl der erzielten Treffer variierte abhängig von der konkreten Fragestellung und bewegte sich zwischen 10 und 200 Ergebnissen je Hypothese. Um die anfängliche Ergebnismenge zu verfeinern, erfolgte eine Analyse der Zusammenfassungen der identifizierten Arbeiten. Auf diese Weise wurde beurteilt, inwiefern diese im Kontext der Forschungsfragen relevant sind. Daraufhin wurde eine engere Auswahl getroffen, die einer intensiveren Untersuchung unterzogen wurden. Die Ergebnisse dieser Recherche führten zu einer endgültigen Auswahl von 110 Arbeiten, die als Grundlage für die umfassende Literaturrecherche dienten.

Die erste Forschungsfrage stützt sich auf fünf wissenschaftlichen Arbeiten, die Zweite auf 7 und die Dritte sowie Vierte auf acht. Die fünfte Hypothese stützt sich auf 18 Quellen, die sechste auf 17 und die Siebte auf 15 relevante Arbeiten. Schließlich fokussiert sich die Hypothese acht auf 12 Quellen.

Das Metaverse als aufstrebende digitale Realität wirft wichtige Fragen hinsichtlich seiner Konsequenzen für junge Menschen auf. Eine gezielte Recherche ist unerlässlich, um ein umfassendes Verständnis dieser Thematik zu entwickeln.

Vor- und Nachteile des Metaverse

Positive Soziale Auswirkungen des Metaverse auf die Junge Generation

Verbesserte Kommunikation und Vernetzung

Aufbauend auf Fortschritten in der Kommunikation, Vernetzung und anderen Technologien wird das Metaverse zu einem Ort für Interaktion, Austausch, Spiele und eine Vielzahl von Diensten, die in virtuellen Welten funktionieren (Bajić, Saeedi-Bajić & Saeedi-Bajić, 2023, S. 1).

Insbesondere während der langanhaltenden COVID-19-Pandemie rückte das Konzept des Metaverse weiter in den Fokus (Bibri & Allam, 2022, S. 2; H. J. Lee & Gu, 2022, S. 2). Besonders junge Menschen waren von den Auswirkungen der Pandemie betroffen, u. a. durch Verschlechterung der psychischen Gesundheit wegen sozialer Isolation (Schlack et al., 2023, S. 24) (Vergleich 4.2.2), und konnten von virtuellen Technologien profitieren. In Zeiten, in denen physische Interaktionen und persönliche Begegnungen eingeschränkt waren, gewann das Metaverse als digitales Reich der Möglichkeiten an Bedeutung. Im Metaverse können sich nämlich Hunderttausende oder Millionen von Menschen versammeln, um ein Festival zu veranstalten oder ein Konzert ihres Lieblingssängers zu sehen (La Capra, 2021). Virtuelle Realitäten, wie Roblox und Zepeto, boten Menschen, die aufgrund von Covid-19 nicht ausgehen konnten, einen neuen sozialen Raum zur Begegnung und Entspannung, in dem sie miteinander kommunizieren konnten. Ein beispielhaftes Ereignis dieser Art war die Durchführung einer virtuellen Autogrammstunde durch die kunstschaffende Gruppe „Blackpink“. Hierbei nahmen mehr als 46 Millionen Nutzende teil, um digitale Autogramme zu erhalten und Selfies mit ihren bevorzugten Kunstschaffenden anzufertigen (Kye, Han, Kim, Park & Jo, 2021, S. 9) und konnten von virtuellen Technologien profitieren. Diese Beispiele illustrieren die Möglichkeiten, die das Metaverse bietet, um soziale Bindungen zu schaffen und den Bedürfnissen der Menschen nach Interaktion in einer zunehmend digital geprägten Welt gerecht zu werden (Kye et al., 2021, S. 9–10).

Erweiterte Bildungsmöglichkeiten

Das Metaverse stellt des Weiteren ein bedeutendes Bildungspotential dar, indem es die zur Wissensaneignung erforderlichen Informationen und Funktionen effizient ausweitet und gleichzeitig eine reale Welt in spiegelbildlicher Form wiedergibt (Arruzza & Chau, 2021, S. 9).

Innerhalb des Metaverse manifestieren sich vielfältige Vorzüge, darunter die Möglichkeit zur bereits genannten Interaktion, die Förderung von Authentizität und Inklusion (Vergleich Kapitel 4.1.4) sowie die Übertragbarkeit von Wissen. Dies erfordert demnach eine Neugestaltung des Bildungssystems, um dessen Zugänglichkeit zu wahren und seine Relevanz aufrechtzuerhalten (Lin, Wan, Gan, Chen & Chao, 2022, S. 3).

Ein Vorteil der Bildung im Metaverse ist z.B., dass es die Einschränkungen herkömmlicher webbasierter Unterrichtsmethoden überwindet (Lin et al., 2022, S. 4). Es eröffnet eine neue Dimension des Lernens, die über die traditionelle Online-Lernumgebung hinausgeht. Durch die Nutzung immersiver Technologien und virtueller Räume können Lernende in realistische Lernsituationen eintauchen, die das Engagement und die Interaktion steigern (Hirsh-Pasek et al., 2022, S. 11; Kye et al., 2021, S. 8). Dies schafft eine Lernerfahrung, die nicht nur auf passiver Wissensaufnahme basiert, sondern die aktive Teilnahme und praktische Anwendung von Konzepten und Fähigkeiten betont (Kye et al., 2021, S. 11; Lin et al., 2022, S. 4).

Gemäß den digitalen Technologien kann das Metaverse den Lernenden helfen, Dinge zu sehen, die in der realen Welt kaum mit den Augen zu sehen sind, wie z.B. Moleküle oder biologische Zellen in einer mikroskopischen Ansicht (Lin et al., 2022, S. 4; Thompson et al., 2021, S. 12). Zusätzlich eröffnet es die Chance, ideale Bedingungen in der Physik zu simulieren und abstrakte Theorien greifbar zu machen (Lin et al., 2022, S. 4). Diese vertieften Einblicke und besseren Visualisierungen ermöglichen ein besseres Verständnis von Inhalten, die im Text schwer zu beobachten oder zu erklären sind (Kye et al., 2021, S. 8). Insbesondere in Fachgebieten wie Chemie und Physik erfordert die Lehre oft experimentelle Ansätze. Nichtsdestotrotz kann durch die Implementierung von Digitalisierung, ein integraler Bestandteil der Metaverse-Bildung, eine umfassende Simulation dieser Experimente realisiert werden. Dies trägt dazu bei, den Ressourcenverbrauch zu verringern. Gleiches trifft auf die praktische Schulung in riskanten Experimenten zu, beispielsweise im Umgang mit entzündlichen oder explosiven chemischen Substanzen (Lin et al., 2022, S. 4). In solchen Fällen reduziert das Metaverse-Training das potenzielle operationelle Risiko für die Lernenden erheblich (Lin et al. 2022, S. 4).

Eine Facette der Bildung in virtuellen Realitäten ist ihre zeitliche Unabhängigkeit, die eine Emanzipation von zeitlichen Zwängen bedeutet (Lin et al., 2022, S. 4). Dies wird durch die Fähigkeit veranschaulicht, historische Ereignisse zu rekonstruieren und in sie einzutauchen, wodurch die Lernenden nicht mehr auf ihre Vorstellungskraft oder herkömmlichen Medien, wie Textquellen oder audiovisuelle Hilfsmittel angewiesen sind, um sie zu verstehen (Kye et al., 2021, S. 8; Lin et al., 2022, S. 4). Auch sind geografische Grenzen in diesem Fall nicht gegeben (Lin et al., 2022, S. 4).

Jedoch gilt es zu erwähnen, dass einige Studien bereits besagen, dass das Aufnahmevermögen von jungen Menschen durch digitale Medien verringert werden und zu einer kürzeren Aufmerksamkeitsspanne führen kann (Dutilleux & Chang, 2022, S. 4727; Hunt, 2023). Man müsste somit näher erforschen, ob durch virtuelle Welten wirklich eine Verbesserung in der Lernfähigkeit entstünde oder, ob diese nicht sogar vermindert werden würde.

Förderung der Kreativität und Selbstentfaltung

Die schnelle Entwicklung digitaler Technologien hat eine neue Ära der virtuellen Realität eingeläutet, die durch das Konzept des Metaverse verkörpert wird. Insbesondere für junge Menschen, eröffnet das Metaverse eine facettenreiche Welt der kreativen Entfaltung und individuellen Entwicklung. Im Folgenden werden positive Auswirkungen auf die Förderung von Kreativität und Selbstentfaltung bei dieser Altersgruppe aufgezeigt. Die beiden zentralen Aspekte, die in diesem Kapitel behandelt werden, sind die Nutzung virtueller Welten als künstlerische Ausdrucksform und die individuellen Gestaltungsmöglichkeiten (Sagar, 2023).

Das Metaverse zeichnet sich durch die Anregung und Entfaltung der kreativen Vorstellungskraft der jungen Generation aus. Virtuelle Umgebungen bieten ein anpassungsfähiges und interaktives Umfeld, in dem junge Menschen ihre kreativen Ideen zum Ausdruck bringen können. Laut eines Berichts von UNICEF (The Metaverse, Extended Reality and Children, 2023) ermöglichen virtuelle Umgebungen den Nutzenden das Erstellen von Avataren, das Gestalten von Objekten und den Aufbau einer eigenen Welt. Diese Möglichkeiten zur kreativen Gestaltung unterstützen das Spiel und regen die Fantasie von jungen Erwachsenen an. Zusätzlich ermöglichen virtuelle Umgebungen die Entwicklung einer Identität in einer spielerischen Umgebung, die in der realen Welt nicht möglich ist. Durch die Interaktion mit der virtuellen Umgebung und anderen Nutzenden können Individuen ihre Kreativität innovativ und interaktiv präsentieren (The Metaverse, Extended Reality and Children, 2023). Zudem besteht auch die Option, ihre Gedanken und Ideen global zu teilen und sich mit anderen zu vernetzen (The Metaverse, Extended Reality and Children, 2023). Das Metaverse bietet auch neue Formen der sozialen Interaktion, wie gemeinsame Unterhaltungserlebnisse bei virtuellen Konzerten oder die Zusammenarbeit bei Spielen und kreativen Übungen, selbst wenn die Teilnehmenden räumlich getrennt sind (Deloitte Deutschland, 2023). Beliebte

virtuelle Events sind beispielsweise das Konzert von Ariana Grande in Fortnite, an dem 78 Millionen Spielende teilnahmen, und das Konzert von Lil Nas in Roblox mit 33 Millionen Nutzenden (The Metaverse, Extended Reality and Children, 2023). Eine Studie des World Economic Forums betont, dass die jüngeren Generationen voraussichtlich mehr Zeit im Metaversum verbringen werden (World Economic Forum, 2023c). Deshalb konzentriert sich Meta auf die jüngere Generation als Hauptzielgruppe, um sie zu ermutigen das Metaverse zu entdecken. Dazu wird mit dem Slogan „It's Your World" (Explore your interests & connections with Meta, 2023) geworben und es zeigt durch prominente Persönlichkeiten und vereinzelt normale Individuen, dass es eine Welt für jeden ist und jeder das sein kann, was er sein möchte. Das Nahbare weckt in der jungen Generation Neugier und Interesse. Das Metaverse fördert somit nicht nur die Kreativität, sondern könnte auch zu einem gestärkten Selbstvertrauen und einem klareren Identitätsgefühl führen.

Jedes Individuum hat das Bedürfnis, sich selbst zu finden und auszudrücken. Verschiedene Wege stehen hierbei zur Verfügung, um diesen Ausdruck zu finden. Dieses Bedürfnis ist ein fundamentaler Bestandteil menschlicher Natur (Endriss, 2021, S. 112). Manche Menschen finden ihre Identität und ihren Ausdruck durch kreative Tätigkeiten wie Malen, Schreiben oder Musik. Andere identifizieren sich durch ihre sozialen Interaktionen oder durch die Ausübung von Berufen, die ihren Fähigkeiten und Interessen entsprechen (Endriss, 2021, S. 112). Das Metaverse als virtuelle Umgebung bietet eine Grundlage, um das Individuum objektiv in den Fokus zu rücken. Insbesondere für die junge Generation ist das Metaverse geeignet, um ihre Kreativität und Selbstentfaltung zu fördern. Ein konkretes Beispiel hierfür ist die digitale Kollaboration zur Förderung der Kreativität im Bereich der Mode (Gastautor, 2022) oder auch im Gaming-Umfeld (Puscher, 2022). Digitale Plattformen ermöglichen es Menschen, ihre Identität gezielt zu formen, indem sie personalisierte Erfahrungen und Selbstdarstellungsmöglichkeiten bereitstellen.

Zusammenfassend bietet das Metaverse der jungen Generation eine umfangreiche Plattform für kreative Entfaltung. Durch die Möglichkeit, selbst erstellte Avatare und Welten zu gestalten sowie interaktive Optionen zu nutzen, wird die Fantasie angeregt und eine individuelle Identität geformt. Die vielseitigen Chancen des Metaverse ermöglichen somit nicht nur globale Vernetzung und kollaboratives Schaffen, sondern fördern auch die Kreativität und das Selbstvertrauen. In einer Ära, in der das Metaverse immer relevanter wird, unterstreicht Meta mit "It's Your World" passenderweise die individuelle Entfaltung, so dass diese virtuelle Sphäre zum Nährboden dafür und für den persönlichen Ausdruck wird.

Inklusion

In Deutschland erfahren im Jahr 2022 immer noch sehr viele Menschen Diskriminierung auf Grundlage der Merkmale des Allgemeinen Gleichbehandlungsgesetz (Ataman, 2023, S. 25). Das Metaverse bietet als eine „ganz neue Welt" die Möglichkeit neue Heuristiken und Paradigmen zu schaffen (Zallio & Clarkson, 2022, S. 5). Dabei müssen die Technologien auf Diversität ausgerichtet sein, sodass sowohl unterrepräsentierte als auch vulnerable Gruppen die Technologien nutzen können (Zallio & Clarkson, 2022, S. 5), damit Diskriminierung entgegengewirkt werden kann. Denn Inklusion, Vielfalt, Gerechtigkeit und Zugänglichkeit sind, als von den United Nations festgelegten Zielen, Grundpfeiler der realen Welt, die auch in der digitalen Welt angestrebt und umgesetzt werden sollten (Zallio & Clarkson, 2022, S. 10). Wenn diese Ziele von Anfang an ins Metaverse inkludiert werden, kann dort ein Raum entstehen, in dem die Jugendlichen Inklusion erleben und lernen können und im besten Fall in die reale Welt adaptieren können. Ein Paradebeispiel für Inklusion bei digitalen Medien war die Covid-19 Pandemie. Der Zugang zu Internet und digitalen Medien war bestimmten Personen und Personengruppen aus unterschiedlichen Gründen verwehrt (Ramsetty & Adams, 2020, S. 1148), wodurch massive Nachteile entstanden sind. Diese Situation kann ebenfalls im Metaverse auftreten, wodurch bestimmte Bevölkerungsgruppen ausgeschlossen und benachteiligt werden (Dwivedi et al., 2023, S. 11).

Den Zugriff zum Metaverse zu gewährleisten, ist zwar notwendig und unumgänglich (Zallio & Clarkson, 2022, S. 7), jedoch kann sich nicht jede Person den Zugang zu virtuellen Technologien, wie z.B. VRBrillen leisten. Die Zugänglichkeit für die breite Masse, kann dadurch aktuell nicht sichergestellt werden. Die Fachleute sehen außerdem eine Chance darin, gleich zu Beginn auf Inklusion im Metaverse zu setzen, um dadurch eine neue Perspektive zu schaffen, die es so in der Realität noch nicht gibt (Zallio & Clarkson, 2022, S. 7). Als Beispiel hierfür sind barrierefreie Zugänge für Menschen mit körperlichen Behinderungen zu nennen (Zallio & Clarkson, 2022, S. 7). Inklusion bedeutetet jedoch auch Zugang für Menschen zu schaffen, die unabhängig von diesen Gründen noch keinen Zugang zum Metaverse oder virtuellen Technologien

haben, um ihnen unbekannte Erfahrungen zu schenken (Zallio & Clarkson, 2022, S. 8). Das Narrativ des Metaverse muss so gestaltet werden, dass Inklusion und Zugänglichkeit von Anfang an und automatisch umgesetzt werden kann (Zallio & Clarkson, 2022, S. 10). Um das sicherzustellen, sollten Werkzeuge, Richtlinien und Regelwerke entworfen werden (K. J. Lee & Law, 2023, S. 3; Sakkas et al., 2008, S. 12; Zallio & Clarkson, 2022, S. 10). Dabei ist es wichtig, dass die Gruppen, die diese Richtlinien aufstellen, ebenfalls Diversität repräsentieren (K. J. Lee & Law, 2023, S. 3; Zallio & Clarkson, 2022, S. 10) und pragmatische sowie durchsetzbare Lösungen finden (Sakkas et al., 2008, S. 66). Zusätzlich spielt auch die Gestaltung von Avataren eine Rolle. Jede nutzende Person hat das Recht dazu, ihren Avatar so zu gestalten, wie sie möchte. Um ein inklusives Metaverse zu schaffen, muss es jedoch möglich sein, den Avatar nach Vorbild des physischen Ichs zu gestalten, was Geschlecht, Repräsentation der Kultur, Kleidung und Behinderungen miteinschließt (Mahlich, 2021). Wenn neben der freien Gestaltung auch die anderen erläuterten Punkte so früh wie möglich umgesetzt werden, hat das Metaverse das Potential, einen inklusiveren Raum als die reale Welt zu bieten und sich positiv auf die Gesellschaft auszuwirken. Insbesondere jüngere Menschen lernen und adaptieren solche neuen Prinzipien schneller als ältere Menschen (Watkins, 2020).

Negative Soziale Auswirkungen des Metaverse auf die Junge Generation

Suchtgefahr, Abhängigkeit und Gesundheitliche Folgen

Neben den Chancen und positiven Aspekten gibt es jedoch, wie bei den meisten Technologien, auch einige Risiken und negative Auswirkungen, die beachtet und minimiert werden sollten. Eine große Gefahr, die bereits aus dem Gebrauch von sozialen Medien bekannt ist, ist die einer Abhängigkeit von einer solchen Technologie (Kırcaburun & Griffiths, 2019, S. 910; Mann & Blumberg, 2022, S. 1; Oberst, Wegmann, Stodt, Brand & Chamarro, 2017, S. 1; Wells, Horwitz & Seetharaman, 2021). Viele Fachleute sehen in der Metaverse-Nutzung eine mindestens gleich starke, wenn nicht sogar schlimmere Suchtgefahr als bei dem Nutzen sozialer Medien oder Computerspielen (Han, Bergs & Moorhouse, 2022, S. 1449; Huddleston, 2022; Madiega, Car & Niestadt, 2022, S. 10; Muslihati et al., 2023, S. 33–34). Bestimmte Studien widersprechen diesen Ergebnissen (Barreda-Ángeles & Hartmann, 2022, S. 6–7), warnen jedoch trotzdem vor zukünftigen Entwicklungen.

Besonders betroffen scheinen dabei die jüngeren Generationen zu sein, die in einer bereits digitalisierten Welt aufwachsen und daher solchen Technologien von jung an ausgesetzt sind (Madiega et al., 2022, S. 9–10) und zudem materialistischer, als die Vorgänger-Generation sind (Ameen, Hosany & Taheri, 2023, S. 2). Letzteres kann zu einer höheren Konsumabhängigkeit, zu der auch digitale Angebote zählen, führen (Mason, Zamparo, Marini & Ameen, 2022, S. 1–3). Laut einer durchgeführten Studie spielen vorwiegend männliche Kinder und Jugendliche Metaverse-Spiele wie Fortnite, Roblox und Minecraft, verglichen zu den üblichen Computerspielen (Melcher, 2022). Das junge Alter dieser Spielenden ergibt sich wahrscheinlich daraus, dass diese Spiele auf Jüngere ausgelegt sind (Melcher, 2022). Dies hat zu Folge, dass das Metaverse, als bereits beliebte Spielplattform, von den jungen Generationen von früh an als Teil des Alltags wahrgenommen werden könnte, was wiederrum, ähnlich zu sozialen Medien, mögliche negative Konsequenzen auf deren Entwicklung haben könnte (Dutilleux & Chang, 2022, S. 4726–4727; Korte, 2020, S. 102–106; Mann & Blumberg, 2022, S. 2).

In Anbetracht der noch relativ geringen Daten und Studien zum Metaverse-Einfluss auf junge Nutzende (Han et al., 2022, S. 1443; Muslihati et al., 2023, S. 33; Petrigna & Musumeci, 2022, S. 6) und angesichts der Ähnlichkeit zu anderen digitalen Angeboten, wie sozialen Medien oder Computerspielen, werden im Folgenden einige negative Aspekte eines hohen Konsums dieser letzteren präsentiert, die sich mit hoher Wahrscheinlichkeit in ähnlichem Maße auf alltägliche Nutzende des Metaverse widerspiegeln werden (Jian, Chen & Yan, 2022, S. 34; Muslihati et al., 2023, S. 33).

Auf der einen Seite belegen verschiedene Studien inzwischen, dass ein hoher Gebrauch von immersiven Computerspielen und sozialen Medien, vor allen Dingen bei Jugendlichen, zu verschiedenen psychologischen Problemen, wie Unsicherheit, Depression und erhöhtem sozialen Vergleich (Gupta & Sharma, 2021, S. 4888; Mann & Blumberg, 2022, S. 1–2; Oberst et al., 2017, S. 2; Schou Andreassen et al., 2016, S. 13; Usmani, Sharath & Mehendale, 2022, S. 4) sowie zu einer geringeren Lebenszufriedenheit (Błachnio, Przepiorka & Pantic, 2016, S. 701) führen kann. Wichtig wurde in diesem Zusammenhang das Konzept „Fear of missing out“ (FOMO), was die Angst, etwas zu verpassen und den Wunsch, ständig in Verbindung mit dem, was andere machen, zu sein beschreibt und durch soziale Medien verstärkt wird

(Gupta & Sharma, 2021, S. 4881; Oberst et al., 2017, S. 3; Przybylski, Murayama, DeHaan & Gladwell, 2013, S. 1). Auch ein Eskapismus-Phänomen, bei dem Nutzende moderner Technologien, in sozialen Medien und Computerspielen bis hin zum Smartphone oder zukünftig dem Metaverse, einen Zufluchtsort sehen, um der realen Welt und dessen Problemen zu entfliehen (Hartl & Berger, 2017, S. 2422; Qasem, Hmoud, Hajawi & Al Zoubi, 2022, S. 293), gewinnt zunehmend an Bedeutung (Han et al., 2022, S. 1449; Schou Andreassen et al., 2016, S. 16). Hierauf wird im folgenden Kapitel 4.2.2 noch tiefer eingegangen.

Auf der anderen Seite unterstreichen andere wissenschaftliche Arbeiten wiederum das Potential vom Metaverse beim Behandeln mentaler Krankheiten und das Lindern von Ängsten, Unsicherheiten und Stress (Madiega et al., 2022, S. 10; Petrigna & Musumeci, 2022, S. 6–7; Usmani et al., 2022, S. 3). Nichtsdestotrotz wird auch in diesen Arbeiten erwähnt, dass das Nutzen von virtuellen 3D-Welten Risiken mit sich trägt und Behandlungen daher von Fachleuten in dem Gebiet unterstützt werden sollte, um gezielte Resultate zu erlangen (Petrigna & Musumeci, 2022, S. 6–7).

Neben den psychischen Nebenwirkungen, die eine exzessive Nutzung des Metaverse hervorruft, gilt es auch die physischen Konsequenzen zu erwähnen. Einerseits belegen einige Studien, dass eine übermäßige Aussetzung gegenüber VR-Brillen oder generell Bildschirmen durch das künstliche Licht die Augen belastet (Ouyang et al., 2020, S. 2; Tosini, Ferguson & Tsubota, 2016, S. 61), andererseits können das verlängerte Sitzen und ein bewegungsarmer Lebensstil entscheidend zu Übergewicht und damit verbundenen Schwierigkeiten beitragen (Madiega et al., 2022, S. 10). All diese negativen Effekte sollten besonders für heranwachsende und sich noch entwickelnde Menschen näher untersucht werden, da sie einen maßgeblichen Einfluss auf das erfolgreiche Umsetzen virtueller Welten haben.

Soziale Isolation und Verlust des Realitätsbezugs

Neben Sucht und Abhängigkeit als potenzielle negative Risiken des Metaverse besteht die Vermutung, dass junge Menschen durch verstärkte Präsenz im Metaverse den Bezug zur Realität verlieren und den Kontakt zur physischen Welt vernachlässigen könnten. Außerdem gilt es zu untersuchen, ob sich der Aufenthalt im Metaverse negativ auf die sozialen Interaktionen der Nutzenden auswirken und im schlimmsten Fall zu sozialer Isolation führen könnte.

Menschen steuern ihre Avatare oft nach dem Vorbild in der physischen Welt (Blascovich et al., 2002, S. 113) und verarbeiten damit virtuelle Erfahrungen ähnlich wie reale Erfahrungen. Dabei tritt der sogenannte Proteus-Effekt auf, der besagt, dass Menschen oft ihre eigenen Einstellungen und Überzeugungen anhand ihres Verhaltens ableiten, ähnlich wie sie das Verhalten einer anderen Person beobachten würden (Fox, Bailenson & Tricase, 2013, S. 932). In virtuellen Umgebungen wird dieser Effekt verstärkt, da das Verhalten oft durch Avatare dargestellt wird.

Durch stärkere Nutzung des Avatars wächst zum einen die Verbindung zwischen Nutzenden und Avatar, zum anderen wächst auch die Verwirrung der Nutzenden bezüglich des Avatars und der Realität (Dwivedi et al., 2022, S. 9). Junge Menschen sind aufgrund von höherer Fantasie und dem stärkeren Wunsch ihren Rollenbildern zu entsprechen tendenziell stärker davon betroffen (Holsapple & Wu, 2007, S. 87). In einer kürzlich durchgeführten Studie wird außerdem darauf hingewiesen, dass das Suchtpotential des Metaverse die Faszination für virtuelle Welten verstärken kann (Jian, Chen & Yan, 2022, S. 38). Darüber hinaus nennen die Verfassenden negative Emotionen, wie z.B. Langeweile, Schmerz und Hilflosigkeit, die junge Menschen dazu veranlassen können, das Metaverse als Ausweg aus dem realen Leben zu nutzen und sich dabei immer tiefer in der virtuellen Realität zu verlieren (Jian et al., 2022, S. 39). Diese Ergebnisse werden durch eine qualitative Interviewstudie gestützt, die den Einfluss der Vertiefung in Video-Spiele während und nach dem Spielen auf Nutzenden zwischen 15 und 21 Jahren untersucht. Die Verfassenden finden heraus, dass einige Spielende von Video-Spielen beeinflusst werden (Ortiz de Gortari & Gackenbach, 2021, S. 14; Ortiz de Gotari, Aronsson & Griffiths, 2011, S. 17). Es zeigt sich zudem, dass starkes Spielverhalten zu negativen psychologischen, emotionalen oder verhaltensbezogenen Auswirkungen führen kann (Ortiz de Gotari et al., 2011, S. 17)

Paul Barroso diskutiert das Thema „Hyperrealität“ und die Wahrnehmung dieser in einer zunehmend technologischen Welt. Er argumentiert, dass die Technologien stärker, visueller und immer hyperrealer werden (Barroso, 2019, S. 55). Da das Metaverse ein Extrembeispiel von digitalen Technologien ist, unterstützt der Autor damit übergreifend ebenfalls die Hypothese, dass die Nutzung von virtuellen Welten die Entfremdung von der realen Welt fördert. In einem aktuellen Paper wird das Metaverse diesbezüglich

sogar als „Darkverse“ bezeichnet (Dwivedi et al., 2023). Auch hier unterstützen die Forschenden durch mehrere Argumentationen die Hypothese, dass das Metaverse die Entfremdung der Realität begünstigt. Sie führen das Beispiel auf, dass traditionelle Medien unsere Realitätswahrnehmung zwar durch Zensur beeinflussen können, Metaverse-Technologien aber aktiv Informationen aus der realen Welt zurückhalten und den Konsumierenden eine verkehrte Welt vorspielen können (Dwivedi et al., 2023, S. 10). Zusätzlich können alle menschlichen Sinne durch Metaverse-Technologien aktiviert werden (Dwivedi et al., 2023, S. 30). Dies verstärkt den Effekt, dass Nutzende den Bezug zur Realität verlieren, was besonders negativ auf Opfer wirkt, die z.B. Erfahrungen mit Cybermobbing machen, (Blackwell, Ellison, Elliott-Deflo & Schwartz, 2019, S. 23).

Die Entfremdung von der Realität birgt jedoch nicht nur Gefahren für die psychische Verfassung der Nutzenden. Durch Hackingangriffe könnte sogar die Realitätswahrnehmung von Nutzenden manipuliert werden mit dem Ziel, ihre Einstellungen und Verhaltensweisen zu beeinflussen (Dwivedi et al., 2023, S. 10). Neben der Entfremdung von der Realität stellt sich außerdem die Frage nach den Auswirkungen des Metaverse auf zwischenmenschliche und soziale Interaktionen. So belegen verschiedene Quellen die Aussage, dass gewalttätige Videospiele zu einer Verringerung des emotionalen Engagements bei der sozialen Verarbeitung und zu verstärktem Aggressionsverhalten führen können (C. A. Anderson & Bushman, 2001, S. 357; Craig A. Anderson, 2004, S. 120; Lai et al., 2019, S. 906). Daher besteht die Vermutung, dass dieses Phänomen auch durch die Nutzung von Metaverse Technologien auftreten könnte.

Allerdings lassen sich in Bezug auf weitere Metaverse-Technologien vor allem positive Konsequenzen auf das soziale Verhalten und Emotionen feststellen (Markowitz & Bailenson, 2021, S. 12–13). Herrera et al. (2018) finden mithilfe zweier Studien bzgl. Perspektivübernahme mit VR am Beispiel von Obdachlosigkeit heraus, dass der Aufbau von Empathie durch die Übernahme der Perspektive gefördert werden kann (Herrera, Bailenson, Weisz, Ogle & Zaki, 2018, 30). Darüber hinaus könnte die Anwendung von Metaverse-Technologien bei Millennials und der Generation Z, die zwischen 1980 und 2012 geboren wurden, zu einer Verringerung der Einsamkeit durch den sozialen Austausch mit anderen beitragen und den Abbau von sozialen Hemmungen begünstigen (Oh et al., 2023, S. 1). Auch in der Covid19 Pandemie wurden die positiven Auswirkungen, insbesondere auf die junge Generation, deutlich. In Form von Videokonferenzen konnten Konzerte oder sogenannte Watch Partys stattfinden (Kye et al., 2021, S. 9). Auch diese Form von Technologien gehört zum Metaverse [Typ „Mirror World“] (Kye et al., 2021, S. 4).

Zusammenfassend kann festgestellt werden, dass das Metaverse sowohl zu einer Entfremdung von der Realität als auch zum Abbau sozialer Hemmungen beitragen kann. Obwohl einige Jugendliche möglicherweise zunehmend von der physischen Welt entfremdet werden, profitieren andere Nutzende von den sozialen Möglichkeiten und der verbesserten Interaktion im Metaverse. Es sei jedoch darauf hingewiesen, dass der Abbau von sozialen Hemmungen auch negative Folgen haben kann. Das zeigt sich im Gaming Bereich durch verstärktes Aggressionsverhalten bei höherem Konsum von Videospielen (s.o. (C. A. Anderson & Bushman, 2001, S. 357; Craig A. Anderson, 2004, S. 120; Lai et al., 2019, S. 906). Es ist wichtig, dieses Phänomen weiter zu erforschen und Mechanismen zu entwickeln, um potentielle negative Auswirkungen zu minimieren und positive Aspekte zu fördern.

Privatsphäre und Datensicherheit

Ein weiteres sehr umstrittenes Thema ist die Frage der Sicherheit und der Privatsphäre, was neben den bereits genannten psychologischen und sozialen Problemen auch rechtliche und ethische Schwierigkeiten mit sich tragen kann. Auf der einen Seite müssen die offengelegten Daten bei Nutzung des Metaverse und der Schutz vor unbefugtem Zugriff geklärt werden. Auf der anderen Seite ist die Privatsphäre ein zentrales Thema, das, wie so oft, bei neuen Technologien geschützt werden muss, um potenzielle Verletzungen zu vermeiden. Da es sich bei dem Metaverse um ein Zusammenspiel verschiedener neuer Technologien, wie beispielsweise Digital Twin1, IoT2 oder Blockchain3 handelt, erscheint das Thema der Sicherheit hier noch etwas komplexer als bei anderen Technologien (Chen, Wu, Gan & Qi, 2022, S. 1–2; Y. Wang et al., 2022, S. 2). Das Entwenden von tragbaren Geräten oder unbefugter Zugang zu Cloud-Speicher sowie der Diebstahl von virtuellen Währungen sind bereits bekannte Probleme, die durch die Verflechtung verschiedener Technologien durch die Komplexitätszunahme weiter verstärkt werden könnten (Y. Wang et al., 2022, S. 2). Gleichzeitig entstehen durch die Nutzung des Metaverse neue Bedrohungen: Durch verschiedene eingebaute Sensoren, die die Nutzung ermöglichen, können große Mengen an verschiedenen Daten (wie Augen- und Handbewegungen, Gesichtsausdrücken, Gehirnwellenmustern, Sprache, biometrischen Daten

und Umgebungsdaten) gesammelt werden, die für eine präzisere Analyse oder zum Profiling der Nutzenden verwendet werden können (Chen et al., 2022, S. 5; Madiega et al., 2022, S. 5; Miller, Herrera, Jun, Landay & Bailenson, 2020, S. 8–9; Nair, Garrido & Song, 2022, S. 6; Y. Wang et al., 2022, S. 9–10). Da die Nutzenden im Metaverse außerdem oft eindeutig identifiziert werden müssen, können Headsets, VR-Brillen oder andere Geräte verwendet werden, um den tatsächlichen Standort der Nutzenden ohne ihr Wissen und ihrer Zustimmung zu lokalisieren (Shang, Chen, Wu & Yin, 2022, S. 434). Aus rechtlicher Sicht stellen solche Informationen in Europa nach der Datenschutzgrundverordnung (DSGVO) (Europäische Union, 2023) sensible personenbezogene Daten dar (Art. 4 DSGVO) und erfordern daher besondere Aufmerksamkeit und die ausdrückliche Zustimmung der Nutzenden im Falle ihrer Verarbeitung (Art. 9 DSGVO) (Madiega et al., 2022, S. 4).

Durch gesammelte Daten könnte das Risiko vermehrter Straftaten auf Basis privater Big Data bestehen, falls unberechtigte Personen darauf zugreifen (Y. Wang et al., 2022, S. 2; Zhao, Zhang, Zhu, Lan & Hua, 2023, S. 2). Ausschlaggebend ist dabei das Benutzen von Avataren, da diese engeren Beziehungen zu den Nutzenden haben als in anderen virtuellen Welten, wie Online-Spielen, und im hohen Maße untereinander interagieren (Zhao et al., 2023, S. 2). Dabei kann es für böswillige Dritte leichter sein, an sensible Daten der Opfer zu geraten und diese damit zu erpressen oder im Allgemeinen diese Daten zu verkaufen (Zhao et al., 2023, S. 3).

Ein weiterer Punkt der Sicherheit und Privatsphäre im Metaverse betrifft das unangemessene oder sogar bösartige Verhalten von einigen Nutzenden, was zu Belästigungen und Cybermobbing führen kann (Qasem et al., 2022, S. 293; Zhao et al., 2023, S. 5). Dies erfolgt, wie in sozialen Medien, einerseits verstärkt dank gewährleisteter Anonymität durch falsche Profile und Avatare, hinter denen sich Nutzende verstecken können, und andererseits aufgrund mangelnder Regulierungen in diesem Bereich (Qasem et al., 2022, S. 292).

Obwohl Sicherheits- und Privatsphäre-Aspekte sich, wie bei den meisten digitalen Technologien, auf alle Nutzenden beziehen, besteht bei Jugendlichen eine besonders hohe Gefahr, da diese sich noch in ihrer Entwicklung und Selbstfindungsphase befinden und dadurch in Bezug auf einige Straftaten vulnerabler und wehrloser als Erwachsene sind (Dutilleux & Chang, 2022; Wolak, J. Mitchell & Finkelhor, 2006, S. 10). So zeigte eine amerikanische Studie bereits 2006, dass einer von sieben Jugendlichen unerwünschten sexuellen Annäherungsversuchen und einer von drei Jugendlichen unerwünschtem Kontakt mit sexuellem Material im Internet ausgesetzt ist (Wolak et al., 2006, S. 10). Weitere Belästigungen oder bedrohliches und beleidigendes Verhalten betreffen hingegen einen von 11 Jugendlichen laut der gleichen Studie (Wolak et al., 2006, S. 10). Obwohl sich diese letzte auf die Internet-Nutzung allgemein bezieht, besteht bisher keine Gewährleistung, dass das Metaverse ein vor Mobbing und Beleidigungen sicherer „Ort“ ist (Qasem et al., 2022, S. 292). Grundsätzlich gehören zu den bereits genannten Gefahren auch ein verstärktes gewalttätiges Verhalten, negative Auswirkungen auf das psychische Wohlbefinden oder die Veröffentlichung privater Daten (Chesney, Coyne, Logan & Madden, 2009, S. 525; Dutilleux & Chang, 2022, S. 4727; Patchin & Hinduja, 2006, S. 148; Qasem et al., 2022; Usmani et al., 2022, S. 4). Besonders bezüglich des Preisgebens persönlicher Informationen, sollten Jugendliche und vor allem Minderjährige rechtlich geschützt sein. Wenn die Europäische Union in den letzten Jahren durch die DSGVO die Tendenz hatte, eine transparentere Datensammlung durch informierte Einwilligung zu ermöglichen, wird dies im Metaverse kaum möglich sein, da die Datensammlung hier ungewollt und kontinuierlich erfolgt, ohne die Möglichkeit einer konstanten Zustimmung (Aneja, 2022). Des Weiteren wird diese Menge an Daten, wie bereits erwähnt, eine besonders genaue Profilierung der Nutzenden ermöglichen (Miller et al., 2020), die ohne strikte Regulierungen zum Missbrauch dieser führen kann, wie z.B. bereits mit anderen Profildaten von Facebook-Nutzenden geschehen ist (Cadwalladr & Graham-Harrison, 2018). Je größer die gesammelte Menge an Daten, desto leichter auch die Möglichkeit der Überwachung, da sich Nutzende dadurch, dank des Einsetzens von KI, vorhersehbarer und kontrollierbarer machen und keinen Einfluss auf das verwendete System haben (Bibri & Allam, 2022, S. 8). Damit würde das Metaverse einer Ermöglichung des von Zuboff (Zuboff, 2020) beschriebenen „Überwachungskapitalismus“ erheblich beitragen (Bibri & Allam, 2022, S. 9). Dieser basiert auf hohen Mengen von Daten, die für prädiktive Modellierung genutzt werden und erzeugt eine noch nie zuvor gesehene Konzentration von Wissen und damit Macht beispielsweise bei einer geringen Anzahl von Technologieunternehmen (Zuboff, 2020, S. 78). Dies könnte auch missbraucht werden, um Minderheiten und Meinungsfreiheit einzuschränken (Bibri & Allam, 2022, S. 16).

Man kann zusammenfassend feststellen, dass das Metaverse erhebliche Sicherheits- und Datenschutzbedenken aufwirft, insbesondere in Bezug auf Jugendliche und Kinder. Es gilt daher, den Schutz von Minderjährigen durch Gesetze zu gewährleisten. Aus rechtlicher Sicht besteht die Herausforderung darin, wie bei anderen Technologien, klare Verantwortlichkeiten zu definieren. (Madiega et al., 2022, S. 5). Eine weitere große Herausforderung wird darin bestehen, weltweit einheitliche Regulierungen festzulegen und durchzusetzen und die Macht der IT-Unternehmen in Grenzen zu halten (Madiega et al., 2022, S. 5–6). Letztere beiden Punkte werden im folgenden Kapitel 4.2.4 und im Kapitel 5.1.2 näher betrachtet und erörtert.

Verantwortung der Plattformbetreibenden im Metaverse

Mit dem wachsenden Interesse am Metaverse stellt sich zunehmend die Frage nach den potenziellen Risiken in einer Umgebung, in der die Grenzen zwischen der physischen und virtuellen Welt verschwimmen. Im Zuge der Entstehung des Metaverse als digitale Parallelwelt übernehmen Betreibende von Plattformen eine wichtige Rolle in der Gestaltung und Verwaltung dieses virtuellen Universums. Da das Metaverse noch nicht vollständig in der Gesellschaft etabliert ist, fehlen entsprechende Regelungen und Gesetze, welche die Plattformbetreibenden in die Pflicht nehmen würden. Die Verantwortung erstreckt sich dabei auf verschiedene Bereiche, die berücksichtigt werden müssen. Eine angemessene Regulierung, Überwachung sowie die Gewährleistung von Sicherheit und ethischen Standards, insbesondere für die junge Generation, sind daher notwendig und werden im Folgenden analysiert. Zudem werden auch die Herausforderungen und zukünftigen Entwicklungen betrachtet (World Economic Forum, 2023a). Aufgrund der unbekannten Dimensionen des Metaverse in der Gesellschaft befindet sich die Diskussion bezüglich regulatorischer Maßnahmen in einem aktiven Stadium. Die Frage nach der Regulierung im Metaverse ist ein komplexes Thema, das verschiedene rechtliche, ethische und technologische Aspekte umfasst. Fachleute und Institutionen, wie das Weltwirtschaftsforum und das Europäische Parlament, beschäftigen sich mit der Herausforderung, die rasante Entwicklung des Metaverse mit angemessener Regulierung in Einklang zu bringen (Madiega et al., 2022; World Economic Forum, 2023b). Die Vielfalt der Aktivitäten im Metaverse, angefangen vom virtuellen Handel bis hin zu sozialen Interaktionen, erfordert eine differenzierte Herangehensweise, um Nutzendenrechte und ethische Standards zu schützen. „Can We Govern The Ungovernable?“ (Boyd, 2022), bei dieser Frage wird einem bewusst, wie schwierig es ist, eine Antwort zu formulieren beziehungsweise eine Lösung zu finden. Für eine lebensfähige Nutzung des Metaverse als Lebens- und Geschäftsraum braucht es reale Kontrollen, um die Nutzenden vor Missbrauch, Betrug und Verlust zu schützen, zu denen z.B. Probleme des Eigentums, illegale Kopien und der Handel von Gütern gehören (Zhao et al., 2023, S. 2). Eine Regulierung erfordert Zeit und ist schwierig, wenn sie global angewendet werden soll (Boyd, 2022). Jedoch können Entwickelnde des Metaversums proaktive Schritte unternehmen, um einen eigenen „Metacode of Conduct“ zu schaffen (Boyd, 2022). Auf diese Weise wird gewährleistet, dass der Inhalt leicht verständlich ist. Diese Regulierungsinitiativen sollen sicherstellen, dass das Metaverse eine sichere Umgebung bleibt und gleichzeitig kreativen Entfaltungsspielraum bietet (Boyd, 2022).

Die Speicherung detaillierter Nutzungsdaten von Metaverse-Nutzenden durch Plattformanbietende sollte eingeschränkt werden (Rosenberg, 2022, S. 25). Um das Profiling der Nutzenden zu verringern, sollten diese Daten lediglich über kurze Zeiträume aufbewahrt werden. Zusätzlich sollte die Öffentlichkeit über die gesammelten Informationen und die Dauer der Aufbewahrung informiert werden (Rosenberg, 2022, S. 25). Plattformanbietende im Metaverse sollten Zugriff auf Nutzendenaktivitäten haben, jedoch sollten nur kurzzeitige Daten gespeichert werden dürfen, die ausschließlich für die Vermittlung virtueller Erfahrungen benötigt werden (Rosenberg, 2022, S. 25). Dadurch würde das Profiling von Nutzendenverhalten reduziert werden. Zusätzlich sollten Anbietende verpflichtet sein, die Öffentlichkeit über erfasste Informationen und Speicherdauer zu informieren, wie beispielsweise die Verfolgung der Blickrichtungen in virtuellen Welten (Rosenberg, 2022, S. 25). Die emotionale Analyse von Metaverse-Nutzenden sollte eingeschränkt werden, da Werbealgorithmen persönliche Eigenschaften wie Gesichtsausdruck, Stimmlage, Körperhaltung und sogar physiologische Parameter wie Herzfrequenz, Atemfrequenz, Pupillenerweiterung und Hautleitfähigkeit in Echtzeit überwachen könnten, um Marketingbotschaften anzupassen und den Nutzenden in Echtzeit zu beeinflussen (Rosenberg, 2022, S. 25). Um diese invasive emotionale Analyse zu begrenzen, sollte eine angemessene Regulierung eingeführt werden, und die Nutzenden sollten immer klar darüber informiert werden, wenn ihre persönlichen Eigenschaften verfolgt und für Werbezwecke verwendet werden (Rosenberg, 2022, S. 25). Die Regulierung virtueller Produktplatzierungen im Metaverse

ist ebenso wichtig, um gezielte und authentische Werbung zu gewährleisten. Plattformen sollten ihre Nutzenden transparent über solche Platzierungen informieren und visuelle Hinweise einführen, um eine lückenlose Transparenz und Schutz vor Missbrauch sicherzustellen (Rosenberg, 2022, S. 25). Die Beschränkung simulierter Persönlichkeiten im Metaverse ist daher unerlässlich. Plattformbetreibende sollten Nutzenden klar darauf hinweisen, wenn sie mit künstlichen Agenten interagieren, die zum Beispiel Werbegespräche führen und dabei eine Echtzeit-Emotionsanalyse verwenden (Rosenberg, 2022, S. 25). Eine strikte Regulierung ist notwendig, um Konsumierende zu schützen und sicherzustellen, dass solche Interaktionen transparent sind (Rosenberg, 2022, S. 25–26).

Die Gewährleistung von Sicherheit und Ethik im Metaverse ist vor allem für die junge Generation von zentraler Bedeutung. Angesichts der Vielfalt der Plattformen und der unzähligen Nutzenden, erfordert dies technologische Lösungen, um Missbrauch zu identifizieren und zu bekämpfen. Eine Herausforderung besteht darin, schädliche Inhalte und Aktivitäten zu erkennen und zu unterbinden, ohne die Meinungsfreiheit und kreative Entfaltung einzuschränken. Gleichzeitig sollten jedoch auch Privatsphäre und Nutzendenfreiheit respektiert werden. Technologische Lösungen spielen eine Schlüsselrolle bei der Gestaltung des Metaversums. Die Entwicklung und Implementierung fortschrittlicher Modelle hinsichtlich künstlicher Intelligenz und maschinellem Lernen, ermöglichen die Erkennung und Entfernung potentieller Gefahren (Bosworth, 2022). Auch Unternehmen wie Facebook haben sich verpflichtet, das Metaversum verantwortungsbewusst zu gestalten und Maßnahmen zur Förderung der Sicherheit und Ethik zu ergreifen (Bosworth, 2022). Es geht nicht nur um den Schutz vor Cyberangriffen, sondern auch darum, sicherzustellen, dass die virtuelle Welt frei von Diskriminierung, Belästigung und anderen ethischen Verstößen bleibt. Die Dynamik des Metaverse und die rasche technologische Evolution stellen Plattformbetreibenden vor ständig neue Herausforderungen. Beispielsweise haben Forschungsarbeiten aufgezeigt, dass die "Ungovernability" des Metaverse eine effektive Regulierung erschwert (World Economic Forum, 2023a). Eine der größten Herausforderungen bei der Regulierung des Metaverse besteht darin, dass sie nicht im Besitz eines einzigen Unternehmens sein dürfen. (Amoils, 2023; Building the Metaverse Responsibly, 2021).

Die Schaffung vertrauenswürdiger Ökosysteme ist ein zentrales Anliegen bei der Entwicklung von Technologien für das Metaverse. Es geht dabei um die Integration von Algorithmen, Strukturen, Rahmenwerken, Regeln und Richtlinien in die Hardware- und Software-Entwicklungszyklen, um die Grundsätze der Sicherheit, des Datenschutzes und der Integrität tief in die DNA der Technologie zu integrieren (Amoils, 2023). Die Übertragung von Daten in virtuellen Welten erfordert eine sorgfältige Prüfung, um den Schutz der Privatsphäre zu gewährleisten. Eine zweite Dimension, die bei der Entwicklung des Metaverse berücksichtigt werden muss, ist die Beseitigung von Vorurteilen, die zu einer nicht inklusiven oder böswilligen Adaption der realen Welt führen könnten. Die Teilnahme am Metaverse basiert auf der Anwendung inklusiver neuer Technologien. Dies erfordert einen globalen, gründlichen und offenen Sicherheitsvalidierungsprozess zum Schutz gegen Verletzungen der Vertraulichkeit, Integrität und anderer Sicherheitsaspekte innerhalb der Umgebungen. Diese Ökosysteme des Vertrauens werden zur Schaffung einer stabilen, integrativen und zielgerichteten Existenz einer virtuellen und immersiven Existenz beitragen (World Economic Forum, 2023b). Dieses Forschungsprojekt (Yang, 2023, S. 1) betont die Bedeutung der Kooperation zwischen Regierungen, Normungsinstitutionen und Unternehmen, um die Entwicklung systematischer Normen zu beschleunigen und die effektive Umsetzung von Metanormen zu fördern.

“Um im Zuge der Entstehung des Metaversums ein umfassendes Sicherheitskonzept zu entwickeln, müssen wir mit anderen Akteuren in Regierung, Industrie, Wissenschaft und Zivilgesellschaft zusammenarbeiten. (ins Deutsche übersetzt)”, Antigone Davis, der Global Head of Safety von Meta (World Economic Forum, 2023b). Laut eines Reports des Europäischen Parlaments gibt es mögliche Implikationen, um das Metaverse zu regulieren (Parliament, S. 4). Die EU debattiert über eine Änderung der Fusionskontrollverordnung, um Fusionen und Übernahmen auf digitalen Märkten zu bekämpfen (Parliament, S. 4). In den USA wird eine Reform der Fusionskontrolle gefordert, um der Marktmacht von Datenverwertungsunternehmen entgegenzuwirken (Parliament, S. 4). Im Metaverse-Umfeld sollen Unternehmen dazu angehalten werden, die digitalen Gesetze und den Wettbewerbsrahmen einzuhalten (Parliament, S. 4). Zudem wird ein Verbot von Dark Patterns und die Einführung von Datensilos vorgeschlagen, um marktübergreifende Datenströme zu blockieren (Parliament, S. 4). Auch die Regulierung von Standards und Interoperabilität im Metaverse wird diskutiert (Parliament, S. 4), dabei fordern einige Fachleute, die Entwicklung offener Metaverse-Standards zu fördern, um ein gemeinschaftliches Metaverse zu unterstützen (Parliament, S. 4)

Zusammenfassend lässt sich sagen, dass sich die Selbstregulierung im Metaverse, analog zu externen Kontrollmechanismen, auf die Stärkung von Transparenz, Glaubwürdigkeit und Verantwortlichkeit auf der Basis von Best Practices konzentriert. Die zu erwartenden Herausforderungen werden als Chance gesehen, wertvolle Erkenntnisse zu gewinnen. Die kollektive Reaktion von Regierungen, Unternehmen und Konsumierenden wird zweifellos die Gestaltung des zukünftigen Metaversums beeinflussen, indem es unser reales Leben ergänzt und erweitert. Die Verantwortung von Plattformbetreibenden im Metaverse umfasst vielfältige rechtliche, ethische, technologische und soziale Aspekte. Es gibt Herausforderungen, aber auch vielversprechende Möglichkeiten. Eine ausgewogene Regulierung, effektive Überwachung und innovative Sicherheitsmaßnahmen sind entscheidend für die Schaffung eines sicheren, ethischen und erfolgreichen Metaverse.

Fazit

Zusammenfassung

Nach einer Veranschaulichung konkreter Vor- und Nachteile der Nutzung des Metaverse vonseiten jüngerer Generationen, kann geschlussfolgert werden, dass das Metaverse zwar starke Potenziale birgt, um einige gesellschaftliche Aspekte zu verbessern oder diese effizienter und inklusiver zu gestalten, es jedoch auch viele Risiken mit sich trägt.

So kann festgestellt werden, dass virtuelle Welten eine erhöhte Vernetzungsmöglichkeit bieten. Insbesondere während Sondersituationen, wie z. B. einer Pandemie, stellt dies einen ausschlaggebenden Mehrwert dar. Insbesondere junge Menschen, die unter der Isolation während der Covid-19-Pandemie am meisten gelitten haben, könnten davon profitieren. Jedoch ist dabei ein gleicher und gerechter Zugriff für alle zu gewährleisten. Auch im Bildungswesen kann das Metaverse maßgeblich zu einer schnelleren und besseren Aufnahmekapazität von Lernenden und Studierenden führen, da eine noch nie zuvor gebotene Visualisierung der Lerninhalte möglich ist. Jedoch sind sich hier nicht alle Forschende einig, ob ein ausschließlich digitales Lernformat, sich nicht doch negativ auf das Lernen und die Aufmerksamkeitsspanne auswirkt. Besonders hervorgehoben werden in mehreren Studien auch die Chancen der Kreativität und der Selbstentfaltung in virtuellen Welten. Diese können die Fantasie bei Heranwachsenden fördern und sie bei ihrer persönlichen Entwicklung durch vernünftigen Gebrauch dieser Technologie unterstützen, indem sie z.B. dazu beitragen, ihr Selbstvertrauen zu stärken.

Ein weiterer positiver Aspekt, der durch den Gebrauch des Metaverse ermöglicht werden könnte, ist der der Inklusion. So wäre es nämlich jedem, unabhängig von körperlichen oder geistigen Gegebenheiten, möglich, an Events oder Aktionen teilzunehmen, die normalerweise nicht für jeden zugänglich wären. Eine jetzige bestehende Herausforderung besteht auch hier darin, Minderheiten sicher zu inkludieren und jeder nutzenden Person, egal aus welchem sozialen Milieu und mit welchen wirtschaftlichen Möglichkeiten, den Zugang zu erlauben.

Obwohl es positive Auswirkungen gibt, müssen auch wichtige negative Auswirkungen erkannt und vermieden werden. Dazu zählen beispielsweise die Abhängigkeit und Suchtgefahr, die sowohl die psychische als auch die physische Gesundheit der Nutzenden beeinträchtigen können. Besonders junge Menschen sind laut zahlreicher Studien betroffen und während ihrer Entwicklung noch vulnerabler und beeinflussbarer als Erwachsene. Da es jedoch auch Fachleute gibt, die im Metaverse Potential für die Bekämpfung bestimmter, vor allem psychischer, Krankheiten sehen, ist es von Bedeutung, die potenziellen Anwendungen und ihre Folgen eingehend zu untersuchen. Eine weitere negative Konsequenz ist die soziale Isolation und der Verlust des Bezugs zur Realität. Hierbei gilt es, im Gegensatz zu der Isolation, die durch besondere Situationen von oben induziert wird, eine Isolation von der realen Welt zu verhindern, die durch zu viel Zeit im Metaverse entsteht. Diese Isolation könnte zu Einsamkeit und im schlimmsten Fall zur Entfremdung von der realen Welt führen. Gleichzeitig besteht die Möglichkeit, dass Verbrauchende ihre sozialen Hemmungen durch eine bessere Interaktion mit anderen reduzieren können. Allerdings können im Extremfall auch negative Folgen auftreten, da laut anderer Studien beispielsweise Aggressionen durch eine verlängerte Nutzung gewalttätiger Videospiele verstärkt werden können. Einer der häufigsten und umstrittensten Nachteile des Metaverse ist der Aspekt der Datensicherheit und der Privatsphäre, wie es bei vielen digitalen Verfahren der Fall ist. Bei dieser aufkommenden neuen Technologie sollten wir beachten, dass Unmengen an Daten gesammelt werden, deren Zusammenführung zu einer noch nie dagewesenen präzisen Analyse der Konsumierenden führen kann. Es wird deutlich vor der Verletzung der Privatsphäre,

Intransparenz solcher Vorgänge und dem Verlust der Kontrolle über persönliche Informationen gewarnt. Auch die mögliche Überwachung seitens IT-Unternehmen oder Regierungen durch die Nutzung von prädiktiven Modellen, um Minderheiten und Meinungsfreiheit einzuschränken, ist als äußerst riskante Konsequenz beschrieben.

Plattformbetreibende sollten rechtlich reguliert werden und dies einheitlich für Anbietende virtueller Welten gelten, damit diese sicher und inklusiv gestaltet werden können. Dies wird jedoch die größte Herausforderung sein, da es schwierig sein wird, solche Gesetze auf globaler Ebene durchzusetzen, da das Vorhaben den Interessen großer und lobbystarken Industrien gegenübersteht. Es gibt bereits verschiedene Vorschläge, die es, unter Betrachtung der ganzen vorgestellten Aspekte, zusammenzubringen und umzusetzen gilt. Im folgenden Kapitel wird nochmal beleuchtet, welche zukünftigen Forschungsbereiche einer vertieften Untersuchung bedürfen.

Weitere Forschung und Einschränkungen

Die vorliegende Seminararbeit untersucht das Metaverse, eine aufstrebende digitale Realität, die die Interaktion von Individuen in virtuellen Welten revolutioniert. Obwohl dieses Forschungsfeld zweifellos ein hohes Potential birgt, um die Art und Weise zu verändern, wie Menschen miteinander agieren und Informationen konsumieren, müssen bestimmte Einschränkungen in Bezug auf Technologie und die vorliegende Seminararbeit berücksichtigt werden.

Eines der zentralen Hindernisse für die breite Implementierung des Metaverse ist die gesellschaftliche Akzeptanz. Die Vorstellung, einen bedeutenden Teil des sozialen und beruflichen Lebens in virtuellen Welten zu verbringen, stößt auf Vorbehalte und Bedenken bezüglich Datenschutzes, Privatsphäre und Verlust persönlichen Kontakts. Die vollständige Integration des Metaverse in die Gesellschaft erfordert daher eine umfassende Zustimmung, die möglicherweise Jahre oder Jahrzehnte dauern könnte. Ein kultureller Wandel ist dafür sicherlich erforderlich. Hinsichtlich der Regulierung ist das Metaverse äußerst komplex. Angesichts der globalen Reichweite und der Vielfalt der darin existierenden Aktivitäten ist die Schaffung und Durchsetzung effektiver Gesetze und Vorschriften eine enorme Herausforderung. Die Gesetzgebenden müssen den Schutz der Nutzenden, die Sicherheit der Plattformen und die Einhaltung ethischer Standards sicherstellen, ohne dabei die Innovationskraft und die Freiheit der kreativen Entfaltung zu beeinträchtigen. Diese ausgewogene Vorgehensweise erfordert umfangreiche Forschung und internationale fachübergreifende Zusammenarbeit. Es ist zu berücksichtigen, dass das Metaverse gemäß der Gartner-Hype-Zyklus-Klassifikation als aufkommende Technologie eingestuft wird und sich noch in einer Phase der Entwicklungsreife befindet. Demzufolge stehen viele der aktuellen Anwendungen und Konzepte dieser neuen Technologie noch am Anfang und haben noch weit entferntes Potential. Daher müssen einige der derzeitigen Annahmen und Vorhersagen über das Metaverse möglicherweise in der Zukunft revidiert werden, wenn sich die Technologie weiterentwickelt. Zudem basiert die Seminararbeit auf der verfügbaren Literatur zum Zeitpunkt der Forschung. Da das Metaverse ein sich schnell entwickelndes Gebiet ist, gibt es möglicherweise neuere Erkenntnisse und Entwicklungen, die in dieser Arbeit nicht berücksichtigt werden konnten. Es ist daher nochmal wichtig zu betonen, dass die Diskussion und Forschung zum Metaverse fortlaufend im Wandel sind. Im Weiteren gibt diese Arbeit einen Überblick über die Literatur zum Metaverse und nennt wichtige Themen und Herausforderungen. Es ist jedoch nicht möglich, alle Aspekte dieses komplexen Themas umfassend zu behandeln. Daher gibt es einen hohen Bedarf an weiterer Forschung und Vertiefung in spezifischen Bereichen dieser virtuellen Welt, um ein vollständiges Verständnis zu erlangen und Lösungen für die Fragen zu finden, die aufgeworfen wurden. Besonders interessant wäre es dabei, die Problematiken und Chancen für die junge Generation konkret weiter zu vertiefen, da diese, wie in dieser Ausarbeitung festgestellt, die Hauptzielgruppe virtueller Welten sein wird. Gleichzeitig resultiert das alleinige Fokalisieren auf die jüngere Gesellschaft als eine Einschränkung dieser Arbeit und es ist weitere Forschung in Bezug auf die Erwachsenen gefragt. Eine weitere Grenze besteht darin, sich rechtlich nur auf die Europäische Union bezogen zu haben, obwohl die meisten Investitionen in das Metaverse in anderen Kontinenten, wie Amerika und Asien, getätigt werden. Aus dieser Perspektive besteht eine große Notwendigkeit einer einheitlichen Gesetzgebung, was aber nicht alleinig das Metaverse betrifft. Dies erfordert nicht nur das Engagement der Plattformbetreibenden, sondern auch eine enge Zusammenarbeit innerhalb der globalen Gemeinschaft. Mit dem bevorstehenden Übergang von flachen sozialen Plattformen zu immersiven Metaverse-Umgebungen steht die Gesellschaft an einem Wendepunkt. Die technologische Ausrichtung des Metaversums auf das Verschwimmen von authentischen und künstlichen Erfahrungen birgt ein erhöhtes Missbrauchsrisiko. Während immersive Virtual-Reality- und

Augmented-Reality-Technologien das Potential haben, Kreativität zu fördern und unser Leben zu bereichern, sollten sowohl Regierungen als auch die Industrie aktiv über eine Regulierung nachdenken, um mögliche Probleme anzugehen, bevor sie fest in den Grundlagen und Geschäftsmodellen des Metaversums verankert sind. Schließlich könnten Schlüsselmaßnahmen wie die Einführung branchenweiter Standards und die Bereitstellung klarer Wahlmöglichkeiten für Nutzenden dazu beitragen, die Entwicklung des Metaversums in eine positive Richtung zu lenken. Ein unabhängiger Branchenverband könnte Standards setzen und Gütesiegel für selbstregulierende, sichere virtuelle Welten etablieren und damit verantwortungsvolle Praktiken fördern.

Literaturverzeichnis

Ameen, N., Hosany, S. & Taheri, B. (2023). Generation Z's psychology and new-age technologies: Implications for future research. *Psychology & Marketing*. **https://doi.org/10.1002/mar.21868**

Amoils, N. (2023, 1. März). How Regulation Will Apply To The Metaverse. *Forbes*. Verfügbar unter: **https://www.forbes.com/sites/nisaamoils/2023/03/01/how-regulation-will-apply-to-the-metaverse/**

Anderson, C. A. & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: a meta-analytic review of the scientific literature. *Psychological Science*, *12*(5), 353–359. **https://doi.org/10.1111/1467-9280.00366**

Anderson, C. A. (2004). An update on the effects of playing violent video games. *Journal of Adolescence*, *27*(1), 113–122. **https://doi.org/10.1016/j.adolescence.2003.10.009**

Aneja, U. (2022, 13. April). Opinion: The challenges of protecting data and rights in the metaverse. *Devex*. Verfügbar unter: **https://www.devex.com/news/sponsored/opinion-the-challenges-of-protecting-data-and-rights-in-the-metaverse-103026**

Arruzza, E. & Chau, M. (2021). The effectiveness of cultural competence education in enhancing knowledge acquisition, performance, attitudes, and student satisfaction among undergraduate health science students: a scoping review. *Journal of Educational Evaluation for Health Professions*, *18*, 3. **https://doi.org/10.3352/jeehp.2021.18.3**

Ataman, F. (2023). Jahresbericht 2022. *Antidiskriminierungsstelle des Bundes*.

Bajić, I. V., Saeedi-Bajić, T. & Saeedi-Bajić, K. (2023, 19. Juli). *Metaverse: A Young Gamer's Perspective*. Verfügbar unter: **https://arxiv.org/pdf/2307.10439**

Ball, M. (2022, 18. Juli). The Metaverse Will Reshape Our Lives. Let's Make Sure It's for the Better. *Time*. Verfügbar unter: **https://time.com/6197849/metaverse-future-matthew-ball/**

Barreda-Ángeles, M. & Hartmann, T. (2022). Hooked on the metaverse? Exploring the prevalence of addiction to virtual reality applications. *Frontiers in Virtual Reality*, *3,* Artikel 1031697. **https://doi.org/10.3389/frvir.2022.1031697**

Barroso, P. (2019). Hyperreality and virtual worlds: when the virtual is real. *Sphera Publica*, (2), 36–58.

Bibri, S. E. & Allam, Z. (2022). The Metaverse as a Virtual Form of Data-Driven Smart Urbanism: On Post-Pandemic Governance through the Prism of the Logic of Surveillance Capitalism. *Smart Cities*, *5*(2), 715–727. **https://doi.org/10.3390/smartcities5020037**

Błachnio, A., Przepiorka, A. & Pantic, I. (2016). Association between Facebook addiction, self-esteem and life satisfaction: A cross-sectional study. *Computers in Human Behavior*, *55*, 701–705. **https://doi.org/10.1016/j.chb.2015.10.026**

Blackwell, L., Ellison, N., Elliott-Deflo, N. & Schwartz, R. (2019). Harassment in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction*, *3(CSCW),* Artikel 100, 1–25. **https://doi.org/10.1145/3359202**

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L. & Bailenson, J. N. (2002). TARGET ARTICLE: Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology. *Psychological Inquiry*, *13*(2), 103–124. **https://doi.org/10.1207/S15327965PLI1302_01**

Bosworth, A. (2022, 19. Dezember). *Meta's Progress in Augmented and Virtual Reality*. Verfügbar unter: **https://about.fb.com/news/2022/12/metas-progress-in-augmented-and-virtual-reality/**

Boyd, M. (2022, 16. Mai). Regulating The Metaverse: Can We Govern The Ungovernable? *Forbes*. Verfügbar unter: **https://www.forbes.com/sites/martinboyd/2022/05/16/regulating-the-metaverse-can-we-govern-the-ungovernable/**

Building the Metaverse Responsibly (2021, 27. September). *Meta*. Verfügbar unter: https://about.fb.com/news/2021/09/building-the-metaverse-responsibly/

Cadwalladr, C. & Graham-Harrison, E. (2018, 17. März). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. Verfügbar unter: **https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election**

Chen, Z., Wu, J. [Jiayang], Gan, W. & Qi, Z. (2022, 27. November). *Metaverse Security and Privacy: An Overview*. Verfügbar unter: **http://arxiv.org/pdf/2211.14948v1**

Chesney, T., Coyne, I., Logan, B. & Madden, N. (2009). Griefing in virtual worlds: causes, casualties and coping strategies. *Information Systems Journal*, *19*(6), 525–548. **https://doi.org/10.1111/j.1365-2575.2009.00330.x**

Citi GPS. (2022). *Metaverse and Money,* Citi. Verfügbar unter: **https://icg.citi.com/icghome/what-we-think/citigps/insights/metaverse-and-money_20220330**

Deloitte Deutschland. (2023, 31. August). *Cross-Sector Briefing: Metaverse – Die Zukunft des Internets? | Deloitte Deutschland*. Verfügbar unter: **https://www2.deloitte.com/de/de/blog/sector-briefings/2022/cross-sector-briefing-metaverse-zukunft-des-internets.html**

Dutilleux, M. & Chang, K.-M. (2022, Februar). Metaverse. Future Addiction Concerned for Human-Being. *International Multilingual Journal of Science and Technology (IMJST)*, (7), 4724–4732. Verfügbar unter: **http://www.imjst.org/wp-content/uploads/2022/02/IMJSTP29120663.pdf**

Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M. et al. (2022). Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *66*, Artikel 102542. **https://doi.org/10.1016/j.ijinfomgt.2022.102542**

Dwivedi, Y. K., Kshetri, N., Hughes, L., Rana, N. P., Baabdullah, A. M., Kar, A. K. et al. (2023). Exploring the Darkverse: A Multi-Perspective Analysis of the Negative Societal Impacts of the Metaverse. *Information Systems Frontiers : a Journal of Research and Innovation*, 25, 2071-2114. **https://doi.org/10.1007/s10796-023-10400-x**

Endriss, L. (2021). *Aufblühen oder Verwelken?* Wiesbaden: Springer Fachmedien Wiesbaden. **https://doi.org/10.1007/978-3-658-34410-8**

Europäische Union. (27. April 2016). VERORDNUNG (EU) 2016/679 DES EUROPÄISCHEN PARLAMENTS UND DES RATES zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung). Verfügbar unter: **https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:32016R0679&qid=1693396254677**

Explore your interests & connections with Meta. (2023, 31. August). *Explore your interests & connections with Meta | Meta*. Verfügbar unter: **https://about.meta.com/meta-and-you/its-your-world/**

Fox, J., Bailenson, J. N. & Tricase, L. (2013). The embodiment of sexualized virtual selves: The Proteus effect and experiences of self-objectification via avatars. *Computers in Human Behavior*, *29*(3), 930–938. **https://doi.org/10.1016/j.chb.2012.12.027**

Gartner. (2022a). *Gartner Predicts 25% of People Will Spend At Least One Hour Per Day in the Metaverse by 2026*. Verfügbar unter: **https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026**

Gartner (Perri, L., Hrsg.). (2022b). *What's New in the 2022 Gartner Hype Cycle for Emerging Technologies*. Verfügbar unter: **https://www.gartner.com/en/articles/what-s-new-in-the-2022-gartner-hype-cycle-for-emerging-technologies**

Gastautor (2022, 15. Februar). Modedesign: Welche Fähigkeiten brauchen Kreative im Metaverse? *FashionUnited*. Verfügbar unter: **https://fashionunited.de/nachrichten/mode/modedesign-welche-faehigkeiten-brauchen-kreative-im-metaverse/2022021545186**

Gupta, M. & Sharma, A. (2021). Fear of missing out: A brief overview of origin, theoretical underpinnings and relationship with mental health. *World Journal of Clinical Cases*, *9*(19), 4881–4889. **https://doi.org/10.12998/wjcc.v9.i19.4881**

Han, D.-I. D., Bergs, Y. & Moorhouse, N. (2022). Virtual reality consumer experience escapes: preparing for the metaverse. *Virtual Reality*, *26*(4), 1443–1458. **https://doi.org/10.1007/s10055-022-00641-7**

Hartl, E. & Berger, B. *Escaping Reality: Examining the Role of Presence and Escapism in User Adoption of Virtual Reality Glasses*. Vortrag anlässlich European Conference on Information Systems (ECIS). Verfügbar unter: **https://www.researchgate.net/publication/318393107_Escaping_Reality_Examining_the_Role_of_Presence_and_Escapism_in_User_Adoption_of_Virtual_Reality_Glasses**

Herrera, F., Bailenson, J., Weisz, E., Ogle, E. & Zaki, J. (2018). Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PloS ONE*, *13*(10), Artikel e0204494. **https://doi.org/10.1371/journal.pone.0204494**

Hirsh-Pasek, K., Zosh, J. M., Hadani, H. S., Golinkoff, R. M., Clark, K., Donohue, C. et al. (2022). A Whole New World: Education Meets the Metaverse. Policy Brief. *Center for Universal Education at the Brookings Institution*. Verfügbar unter: **https://eric.ed.gov/?id=ED622316**

Holsapple, C. W. & Wu, J. [Jiming]. (2007). User acceptance of virtual worlds. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *38*(4), 86–89. **https://doi.org/10.1145/1314234.1314250**

Huddleston, T., JR. (2022, 31. Januar). 'This is creating more loneliness': The metaverse could be a serious problem for kids, experts say. *CNBC*. Verfügbar unter: **https://www.cnbc.com/2022/01/31/psychologists-metaverse-could-be-a-problem-for-kids-mental-health.html**

Hunt, E. (2023, 1. Januar). Is modern life ruining our powers of concentration? *The Guardian*. Verfügbar unter: **https://www.theguardian.com/technology/2023/jan/01/is-modern-life-ruining-our-powers-of-concentration**

Ibrahim, D. (2022). *Roblox Metaverse: An In-Depth Look [2023],* The Metaverse Insider. Verfügbar unter: **https://metaverseinsider.tech/2022/11/21/roblox-metaverse/**

Isaac, M. (2021, 28. Oktober). Facebook Renames Itself Meta. *The New York Times*. Verfügbar unter: **https://www.nytimes.com/2021/10/28/technology/facebook-meta-name-change.html**

Jian, S., Chen, X. & Yan, J. (2022). From Online Games to "Metaverse": The Expanding Impact of Virtual Reality in Daily Life. In M. Rauterberg (Hrsg.), *Culture and Computing* (Lecture Notes in Computer Science, Bd. 13324, S. 34–43). Cham: Springer International Publishing. **https://doi.org/10.1007/978-3-031-05434-1_3**

Kırcaburun, K. & Griffiths, M. D. [Mark D.]. (2019). Problematic Instagram Use: The Role of Perceived Feeling of Presence and Escapism. *International Journal of Mental Health and Addiction*, *17*(4), 909–921. **https://doi.org/10.1007/s11469-018-9895-7**

Kollmann, T. (2018). *Avatar*. Verfügbar unter: **https://wirtschaftslexikon.gabler.de/definition/avatar-31903/version-255451**

Korte, M. (2020). The impact of the digital revolution on human brain and behavior: where do we stand? *Dialogues in Clinical Neuroscience*, *22*(2), 101–111. **https://doi.org/10.31887/DCNS.2020.22.2/mkorte**

Kubach, B. (Microsoft, Hrsg.). (2021). *Microsoft erklärt: Was ist Microsoft Mesh? Definition & Funktionen | News Center Microsoft*. Verfügbar unter: **https://news.microsoft.com/de-de/microsoft-erklaert-was-ist-microsoft-mesh/**

Kye, B., Han, N., Kim, E., Park, Y. & Jo, S. (2021). Educational applications of metaverse: possibilities and limitations. *Journal of Educational Evaluation for Health Professions*, *18*, 32. **https://doi.org/10.3352/jeehp.2021.18.32**

La Capra, E. (2021, 11. Oktober). The Metaverse Concerts: Where Online Games and Music Performances Meet. *CoinMarketCap*. Verfügbar unter: **https://coinmarketcap.com/alexandria/article/the-metaverse-concerts-where-online-games-and-music-performances-meet**

Lai, C., Pellicano, G. R., Altavilla, D., Proietti, A., Lucarelli, G., Massaro, G. et al. (2019). Violence in video game produces a lower activation of limbic and temporal areas in response to social inclusion images. *Cognitive, Affective & Behavioral Neuroscience*, *19*(4), 898–909. **https://doi.org/10.3758/s13415-018-00683-y**

Lasserre, S. (2022). *The Metaverse Hype Cycle: how AR/VR for professionals will be impacted,* Techviz. Verfügbar unter: **https://blog.techviz.net/the-metaverse-hype-cycle-how-ar-vr-for-professionals-will-be-impacted**

Lee, H. J. & Gu, H. H. (2022). Empirical Research on the Metaverse User Experience of Digital Natives. *Sustainability*, *14*(22), Artikel 14747. **https://doi.org/10.3390/su142214747**

Lee, K. J. & Law, E. (2023). *Towards a More Inclusive Metaverse via Designing Tools That Support Collaborative Virtual World Building by Users With and Without Disabilities*. **https://doi.org/10.48550/arXiv.2305.04368**

Lee, L.-H., Braud, T., Zhou, P., Wang, L., Xu, D., Lin, Z. et al. (2021, 6. Oktober). *All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda*. Verfügbar unter: **https://arxiv.org/pdf/2110.05352**

Liebsch, K. (2020). *Generation*. Verfügbar unter: **https://www.staatslexikon-online.de/Lexikon/Generation**

Lin, H., Wan, S., Gan, W., Chen, J. & Chao, H.-C. (2022). Metaverse in Education: Vision, Opportunities, and Challenges. In S. Tsumoto (Hrsg.), *Proceedings, 2022 IEEE International Conference on Big Data. Dec 17 - Dec 20, 2022, Osaka, Japan* (S. 2857–2866) [Piscataway, New Jersey]: IEEE.

Luber, S. (2022). Definition: Zukunftstrend Metaversum und die Vision von Mark Zuckerberg Was ist Metaverse? *Cloudcomputing Insider*. Verfügbar unter: **https://www.cloudcomputing-insider.de/was-ist-metaverse-a-4986565d060780810 5d09dfcea005f42/#:~:text=Metaverse%20ist%20eine%20Vision%20f%C3%BCr,virtuellen%20Raum%20ohne%20innere%20Grenzen.**

Madiega, T., Car, P. & Niestadt, M. (2022, 24. Juni). Metaverse. Opportunities, risks and policy implications. *EPRS | European Parliamentary Research Service*. Verfügbar unter: **https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2022)733557**

Mahlich, H. (2021). How to Ensure That the Metaverse Is an Inclusive Space. *SHOWstudio*. Verfügbar unter: **https://www.showstudio.com/news/how-to-ensure-the-metaverse-is-an-inclusive-space**

Maloney, D. (2021). A Youthful Metaverse: Towards Designing Safe, Equitable, and Emotionally Fulfilling Social Virtual Reality Spaces for Younger Users. *All Dissertations*. Verfügbar unter: **https://tigerprints.clemson.edu/all_dissertations/2931**

Mann, R. B. & Blumberg, F. (2022). Adolescents and social media: The effects of frequency of use, self-presentation, social comparison, and self esteem on possible self imagery. *Acta Psychologica*, *228*, Artikel 103629. **https://doi.org/10.1016/j.actpsy.2022.103629**

Markowitz, D. M. & Bailenson, J. (2021). *Virtual Reality and Emotion: A 5-Year Systematic Review of Empirical Research (2015-2019)*. **https://doi.org/10.31234/osf.io/tpsmr**

Mason, M. C., Zamparo, G., Marini, A. & Ameen, N. (2022). Glued to your phone? Generation Z's smartphone addiction and online compulsive buying. *Computers in Human Behavior*, *136*, Artikel 107404. **https://doi.org/10.1016/j.chb.2022.107404**

Melcher, N. (Newzoo, Hrsg.). (2022, 19. Januar). *Deep Dive: Early Metaverse Players – Data on Demographics, Socializing, Playing, & Spending*. Verfügbar unter: **https://newzoo.com/resources/blog/deep-dive-metaverse-gamers-data-on-metaverse-demographics-socializing-playing-spending-2#:~:text=With%20an%20average%20age%20of,and%20are%20marketed%20to%20children.**

Messinger, P. R., Ge, X., Stroulia, E., Lyons, K., Smirnov, K. & Bone, M. (2008). On the Relationship between My Avatar and Myself. *Journal for Virtual Worlds Research*, *1*(2). **https://doi.org/10.4101/jvwr.v1i2.352**

The Metaverse, Extended Reality and Children. (2023, 31. August). Verfügbar unter: https://www.unicef.org/globalinsight/reports/metaverse-extended-reality-and-children

Miller, M. R., Herrera, F., Jun, H., Landay, J. A. & Bailenson, J. N. (2020). Personal identifiability of user tracking data during observation of 360-degree VR video. *Scientific Reports*, *10*(1), Artikel 17404. **https://doi.org/10.1038/s41598-020-74486-y**

Muslihati, M., Hotifah, Y., Hidayat, W. N., Sobri, A. Y., Valdez, A. V., Ilmi, A. M. et al. (2023). How to Prevent Student Mental Health Problems in Metaverse Era? *Jurnal Kajian Bimbingan dan Konseling*, *8*(1), 33–46. **https://doi.org/10.17977/um001v8i12023p33-46**

Mystakidis, S. (2022). Metaverse. *Encyclopedia*, *2*(1), 486–497. **https://doi.org/10.3390/encyclopedia2010031**

Nair, V., Garrido, G. M. & Song, D. (2022, 26. Juli). *Exploring the Unprecedented Privacy Risks of the Metaverse*. Verfügbar unter: **https://arxiv.org/pdf/2207.13176**

Oberst, U., Wegmann, E., Stodt, B., Brand, M. & Chamarro, A. (2017). Negative consequences from heavy social networking in adolescents: The mediating role of fear of missing out. *Journal of Adolescence*, *55*, 51–60. **https://doi.org/10.1016/j.adolescence.2016.12.008**

Oh, H. J., Kim, J. [Junghwan], Chang, J. J., Park, N. & Lee, S. (2023). Social benefits of living in the metaverse: The relationships among social presence, supportive interaction, social self-efficacy, and feelings of loneliness. *Computers in Human Behavior*, *139*, Artikel 107498. **https://doi.org/10.1016/j.chb.2022.107498**

Ortiz de Gortari, A. B. & Gackenbach, J. (2021). Game Transfer Phenomena and Problematic Interactive Media Use: Dispositional and Media Habit Factors. *Frontiers in Psychology*, *12*, Artikel 585547. **https://doi.org/10.3389/fpsyg.2021.585547**

Ortiz de Gotari, A., Aronsson, K. & Griffiths, M. D. [M. D.]. (2011). Game Transfer Phenomena in video game playing: A qualitative interview study. *International Journal of Cyber Behavior, Psychology and Learning*, 1(3), 15–33. Verfügbar unter: **https://www.researchgate.net/publication/235683755_Game_Transfer_Phenomena_in_video_game_playing_A_qualitative_interview_study**

Ortiz-Ospina, E. (2019). The rise of social media. Social media sites are used by more than two-thirds of internet users. How has social media grown over time? *OurWorldInData.org*. Verfügbar unter: **https://ourworldindata.org/rise-of-social-media**

Ouyang, X., Yang, J., Hong, Z., Wu, Y., Xie, Y. & Wang, G. (2020). Mechanisms of blue light-induced eye hazard and protective measures: a review. *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie*, *130*, Artikel 110577. **https://doi.org/10.1016/j.biopha.2020.110577**

Park, S.-M. & Kim, Y.-G. (2022). A Metaverse: Taxonomy, Components, Applications, and Open Challenges. *IEEE Access*, *10*, 4209–4251. **https://doi.org/10.1109/ACCESS.2021.3140175**

Parliament, E.. Metaverse. Verfügbar unter: **https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733557/EPRS_BRI(2022)733557_EN.pdf**

Patchin, J. W. & Hinduja, S. (2006). Bullies Move Beyond the Schoolyard. *Youth Violence and Juvenile Justice*, *4*(2), 148–169. **https://doi.org/10.1177/1541204006286288**

Petrigna, L. & Musumeci, G. (2022). The Metaverse: A New Challenge for the Healthcare System: A Scoping Review. *Journal of Functional Morphology and Kinesiology*, *7*(3), Artikel 63. **https://doi.org/10.3390/jfmk7030063**

Pleasant, R. (2023, 4. April). What is Microsoft Mesh? *UC Today*. Verfügbar unter: **https://www.uctoday.com/collaboration/what-is-microsoft-mesh/**

Przybylski, A. K., Murayama, K., DeHaan, C. R. & Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*, *29*(4), 1841–1848. **https://doi.org/10.1016/j.chb.2013.02.014**

Puscher, F. (2022). *Metaverse: "Wir müssen über Gaming und bunte Pixel hinausdenken&quot*. Verfügbar unter: **https://www.meedia.de/technologie/alexander-el-meligi-co-founder-und-managing-partner-metaverse-wir-muessen-ueber-gaming-und-bunte-pixel-hinausdenken-bd76318b44a4a57b098595d16ae11464**

Qasem, Z., Hmoud, H. Y., Hajawi, D. & Al Zoubi, J. Z. *The Effect of Technostress on Cyberbullying in Metaverse Social Platforms*. Vortrag anlässlich International Working Conference on Transfer and Diffusion of IT. Retrieved from **https://link.springer.com/chapter/10.1007/978-3-031-17968-6_22#Sec2**

Ramsetty, A. & Adams, C. (2020). Impact of the digital divide in the age of COVID-19. *Journal of the American Medical Informatics Association : JAMIA*, *27*(7), 1147–1148. **https://doi.org/10.1093/jamia/ocaa078**

Rinaldi, G. (2022). Virtuelle Welten - Leben wir bald im Metaverse? *Frankfurter Allgemeine Zeitung*.

Rosenberg, L. (2022). Regulation of the Metaverse: A Roadmap. In *Proceedings of the 6th International Conference on Virtual and Augmented Reality Simulations* (ACM Digital Library, S. 21–26). New York,NY,United States: Association for Computing Machinery.

Sagar, S. (2023, 8. August). Empowering Children to become Creators. *Wizklub Blogs*. Verfügbar unter: **https://blogs.wizklub.com/learning-journey/empowering-children-to-become-creators/**

Sakkas, N., Vlachaki, E., Futo, P., Toth, L., Melchiorri, M., Saracino, S. et al. (2008). ICT for All: Towards an e-Inclusive Society. Verfügbar unter: **https://www.researchgate.net/publication/281522604_ICT_for_All_Towards_an_e-Inclusive_Society**

Schlack, R., Neuperd, L., Junker, S., Eicher, S., Hölling, H., Thom, J. et al. (2023). Veränderungen der psychischen Gesundheit in der Kinder- und Jugendbevölkerung in Deutschland während der COVID-19- Pandemie – Ergebnisse eines Rapid Reviews. *Journal of Health Monitoring*, (S1), 1–74.

Schou Andreassen, C., Billieux, J., Griffiths, M. D. [Mark D.], Kuss, D. J., Demetrovics, Z., Mazzoni, E. et al. (2016). The relationship between addictive use of social media and video games and symptoms of psychiatric disorders: A large-scale cross-sectional study. *Psychology of Addictive Behaviors : Journal*

of the Society of Psychologists in Addictive Behaviors, *30*(2), 252–262. **https://doi.org/10.1037/adb0000160**

Shang, J., Chen, S., Wu, J. [Jie] & Yin, S. (2022). ARSpy: Breaking Location-Based Multi-Player Augmented Reality Application for User Location Tracking. *IEEE Transactions on Mobile Computing*, *21*(2), 433–447. **https://doi.org/10.1109/TMC.2020.3007740**

Skarredghost (2022, 20. August). The metaverse enters the Gartner hype cycle (but with a 10+ years outlook). *Skarredghost*. Verfügbar unter: **https://skarredghost.com/2022/08/20/metaverse-gartner-hype-cycle/**

Sortlist - Data Hub. (2023, 1. März). *Stand des Metaverse (Umfrage): So passen sich Unternehmen an*. Verfügbar unter: **https://www.sortlist.de/datahub/reports/metaverse-fuer-unternehmen/**

Statista. (2022). *Global Roblox game user distribution by age 2022 | Statista*. Verfügbar unter: **https://www.statista.com/statistics/1190869/roblox-games-users-global-distribution-age/**

Statista. (2023, 1. September). *Global Roblox games user engagement 2023 | Statista*. Verfügbar unter: **https://www.statista.com/statistics/1192663/user-engagement-global-roblox-games/**

Thompson, M., Uz-Bilgin, C., Tutwiler, M. S., Anteneh, M., Meija, J. C., Wang, A. et al. (2021). Immersion positively affects learning in virtual reality games compared to equally interactive 2d games. *Information and Learning Sciences*, *122*(7/8), 442–463. **https://doi.org/10.1108/ILS-12-2020-0252**

Tosini, G., Ferguson, I. & Tsubota, K. (2016). Effects of blue light on the circadian system and eye physiology. *Molecular Vision*, *22*, 61–72. Retrieved from **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4734149/**

Usmani, S. S., Sharath, M. & Mehendale, M. (2022). Future of mental health in the metaverse. *General Psychiatry*, *35*(4), Artikel e100825. **https://doi.org/10.1136/gpsych-2022-100825**

Wang, H., Ning, H., Lin, Y., Wang, W., Dhelim, S., Farha, F. et al. (2023). A Survey on the Metaverse: The State-of-the-Art, Technologies, Applications, and Challenges. *IEEE Internet of Things Journal*, *10*(16), 14671–14688. **https://doi.org/10.1109/JIOT.2023.3278329**

Wang, Y. [Yuntao], Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T. H. et al. (2022). *A Survey on Metaverse: Fundamentals, Security, and Privacy*. **https://doi.org/10.36227/techrxiv.19255058.v3**

Wang, Y. [Yuntao], Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T. H. et al. (2023). A Survey on Metaverse: Fundamentals, Security, and Privacy. *IEEE Communications Surveys & Tutorials*, *25*(1), 319–352. **https://doi.org/10.1109/COMST.2022.3202047**

Watkins, D. (2020). *Why Children Learn Faster than Adults and How You Can Learn Their Tricks*. Verfügbar unter: **https://irisreading.com/why-children-learn-faster-than-adults-and-how-you-can-learn-their-tricks/**

Wells, G., Horwitz, J. & Seetharaman, D. (2021, 14. September). Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. *The Wall Street Journal*. Verfügbar unter: **https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=hp_lead_pos7**

Wieser, A. (2022). *GAMING IM METAVERSE: EIN NEUER TREND VERÄNDERT DIE BRANCHE,* WAR DA. KOMME WIEDER. Verfügbar unter: **https://warda.at/magazin/gaming/gaming-im-metaverse-ein-neuer-trend-veraendert-die-branche/**

Wolak, J., J. Mitchell, K. & Finkelhor, D. (2006). Online victimization of youth: Five years later. *Crimes against Children Research Center*. Verfügbar unter: **https://www.unh.edu/ccrc/resource/online-victimization-youth-five-years-later**

World Economic Forum. (2019). *This graph tells us who's using social media the most*. Verfügbar unter: **https://www.weforum.org/agenda/2019/10/social-media-use-by-generation/**

World Economic Forum. (2023, 31. Augusta). *Here's how to make the metaverse safe for everyone*. Verfügbar unter: **https://www.weforum.org/agenda/2022/05/heres-how-to-make-the-metaverse-safe/**

World Economic Forum. (2023, 31. Augustb). *How to address digital safety in the metaverse*. Verfügbar unter: https://www.weforum.org/agenda/2022/01/metaverse-risks-challenges-digital-safety/

World Economic Forum. (2023, 31. Augustc). *Younger generations expect to spend a lot more time in the metaverse*. Verfügbar unter: **https://www.weforum.org/agenda/2022/08/metaverse-technology-virtual-future-people/**

Yang, L. (2023). Recommendations for metaverse governance based on technical standards. *Humanities and Social Sciences Communications*, *10*(1), 1–10. **https://doi.org/10.1057/s41599-023-01750-7**

Zallio, M. & Clarkson, P. J. (2022). Designing the metaverse: A study on inclusion, diversity, equity, accessibility and safety for digital immersive environments. *Telematics and Informatics*, *75*, Artikel 101909. **https://doi.org/10.1016/j.tele.2022.101909**

Zhao, R., Zhang, Y., Zhu, Y., Lan, R. & Hua, Z. (2023). Metaverse: Security and Privacy Concerns. *Journal of Metaverse*, *3*(2), 93–99. **https://doi.org/10.57019/jmv.1286526**

Zuboff, S. (2020). *The age of surveillance capitalism. The fight for a human future at the new frontier of power* (First trade paperback edition). New York, NY: PublicAffairs.

Towards Human Digital Twin in Healthcare: Technical, Legal, and Ethical Implications

Selected Issues in Critical Information Infrastructures, Summer Term 2023

Vasileios Xanthakis
Master Student
Karlsruhe Institute of Technology
vasileios.xanthakis@student.kit.edu

Jurek Muff
Master Student
Karlsruhe Institute of Technology
jurek.muff@student.kit.edu

Stephan Timpe
Master Student
Karlsruhe Institute of Technology
stephan.timpe@student.kit.edu

Sophia Weeber
Master Student
Karlsruhe Institute of Technology
sophia.weeber@student.kit.edu

Abstract

Background: *The clinical deployment of the Human Digital Twin (HDT) could bring huge benefits for patients, doctors, insurance, and other stakeholders of the healthcare system. Due to the complex and novel nature of the technologies involved in deploying an HDT, assessing the social and ethical implications of HDT along with the technical implementation is critical in order to properly evaluate the current and future state of deployment. Although these aspects have been examined in isolation, a combinatory approach is deemed vital in evaluating the transfer of HDT from theoretical and experimental research to real-world applications.*

Objective: *This work aims to provide a holistic overview of the technical, legal, and ethical aspects of HDT development and deployment.*

Methods: *Expert interviews were found to be an effective method to assess the current and future state of HDT development, as professionals in the field of medical informatics research, quality management, medical innovation, and personalized medicine have actively delved into relevant cases and aspects of this technology. The results generated through the interviews are analyzed by employing open, deductive, and axial coding techniques.*

Results: *Reflection upon the categorized results provides valuable insights into the examined implications of HDT technology. The interconnectiveness of the different areas becomes evident, as legal and ethical considerations are directly linked to the technical implementation of HDT. Furthermore, the use of advanced machine learning techniques, such as deep learning, may improve technical performance but it also creates significant regulatory barriers and ethical dangers due to the inherent lack of explainability.*

Conclusion: *Although the high importance of restrictions to prevent the abuse of HDT technology is established, the current regulatory framework in the EU and in particular in Germany is deemed overly restrictive, to a degree that it puts companies, research teams, and startups at a significant competitive disadvantage compared to their counterparts in the USA and in Asia. These insights contribute to a better understanding of HDT and its implementation while paving the way for more effective research approaches toward realistic, responsible, and effective real-world applications.*

Keywords: human digital twin, explainability, machine learning, clinical decision support systems, medicine, diagnostics

Introduction

Motivation

The field of healthcare has always played a vital role in society, but its importance has significantly increased in recent years due to a confluence of factors like advancements in technology, changing demographics, evolving healthcare needs, and lately, the COVID-19 pandemic (e.g. Agbo et al., 2019; Erol et al., 2020; Ghazal et al., 2021). Rapid developments in healthcare technology have opened up new possibilities for diagnosis, treatment, and healthcare delivery, leading to an increased reliance on and importance of healthcare services (Agbo et al., 2019; Yeganeh, 2019). At the same time, the emergence of global health challenges has placed healthcare at the forefront of societal concerns (Ferrara & Albano, 2020). The COVID-19 pandemic, for instance, has highlighted the importance of strong healthcare systems, disease surveillance, and effective public health interventions (e.g. Iyengar et al., 2020; Roy et al., 2021). In general, advancements in healthcare have led to longer life expectancies, resulting in a larger proportion of elderly individuals in many countries (United Nations & Social Affairs, 2019). According to the World Health Organization (2022a) between 2015 and 2050, the proportion of the world's population over the age of 60 years will nearly double from 12% to 22% and older adults typically require more healthcare services due to a higher prevalence of chronic conditions and age-related illnesses (Karlsson et al., 2018).

Apart from the prominent role healthcare is likely to play in the future, aspects such as the above mentioned global health challenges and demographic change are leading to ever-increasing healthcare expenditures, which pose challenges to governments across the globe (World Health Organization, 2022b). Health expenditures in Germany increased from 12% of GDP in 2019 to 13% in 2020 (World Health Organization, 2021). Furthermore, global health expenditures as a percentage of GDP rose from an average of 9.84 to 10.89 in 2020 (World Health Organization, 2021).

New technologies such as human digital twins (HDT) hold great potential in addressing the challenges associated with the increasing costs of healthcare (e.g. Y. Liu et al., 2019b; Okegbile et al., 2022). HDT is referred to as the replica of a physical-world human in the digital world (Lin et al., 2022). By creating a digital representation of an individual's unique physiological characteristics, health history, and genetic profile, HDT can contribute to more personalized and precise diagnosis and treatment, thus addressing prevailing challenges in healthcare Garg (2021). For instance, HDT enable targeted interventions, allowing healthcare professionals to identify the most effective and efficient actions by analyzing a patient's digital twin (Okegbile et al., 2022). The approach of targeted intervention can minimize healthcare costs by avoiding unnecessary treatments, reducing trial-and-error approaches, and thus improving patient outcomes (Haleem et al., 2023). Other key capabilities of HDT include personalized medicine, which facilitates the development of personalized treatment plans and prevention strategies tailored to an individual's specific needs based on their medical history (e.g. Erol et al., 2020; Sahal et al., 2022; Sun et al., 2023). In conjunction with other technologies such as AI and machine learning algorithms, HDT enable predictive analytics (J. Chen et al., 2023). By continuously monitoring a patient's digital twin, healthcare providers can identify early warning signs of potential health problems and proactively intervene (Sahal et al., 2022). This can help prevent costly hospitalizations, emergency visits, and disease progression, leading to a more efficient allocation of healthcare resources (Hassani et al., 2022). Furthermore, by connecting HDT to real-time data streams from wearable devices, remote sensors, and telemedicine platforms healthcare professionals can remotely assess and monitor patients, thus reducing unnecessary hospital visits (J. Chen et al., 2023; Sahal et al., 2022).

While a HDT does not yet fully exist in practice due to a complex intertwining of legal, ethical, and technical issues, there are already initial approaches by companies such as Siemens Healthineers or Babylonhealth, for example, which use 3D digital heart twins to simulate surgical interventions or enable personalized healthcare (Babylon Health, 2023; Siemens Healthineers, 2023). However, development is inhibited, partly because e.g., collecting, integrating, and managing the massive amounts of required data while respecting privacy regulations and ethical considerations is a significant hurdle. Furthermore, technological or legal factors such as liability and accountability are manifold and interdependent, thus limiting the development and implementation of HDT (Shengli, 2021). In order to effectively implement HDT in real-world

applications, there appears to be a notable gap in comprehensively understanding the intricate connections and trade-offs between technical, legal, and ethical considerations.

In parallel, research on HDT has advanced in recent years. In particular, in terms of technology development, concepts such as machine learning, Big Data, Cloud computing and the Internet of Things (IoT) have contributed to the advancement of the concept of HDT (Shengli, 2021). However, it is evident that often only specific aspects of HDT are investigated. In particular, technical details, specific application scenarios or a review of ethical and legal aspects are usually at the forefront of current research approaches (e.g. Chakshu et al., 2019a; Sahal et al., 2022; Sun et al., 2023). Frequent neglect of higher-level requirements, potential interdependencies, and trade-offs between seemingly independent aspects, including technical, legal, and ethical considerations, is a prevalent problem, which is indicative of where the current state of HDT in practice may stem from (e.g. Popa et al., 2021; Sahal et al., 2022; Sun et al., 2023). This refers in particular to the situations in which achieving improvement or optimization in one aspect involves sacrificing or compromising another. For example, detailed personal medical data might be used to provide highly accurate diagnoses, but at the same time, it increases the risk and impact of data protection incidents (see section "Positive and Negative Implications of HDT in Healthcare"). Due to the complexity of modeling and simulating various interconnected systems for HDT, these relationships may be of particular interest. Thus, this omission of potential interdependencies and trade-offs might be an impediment to achieving a comprehensive understanding of the subject matter and moving forward with the implementation of HDT in practice.

Objectives

This paper seeks to address the research gap regarding a holistic picture of the challenges for the adoption of HDT in healthcare. To achieve this objective, the paper aims to address the following research question:

What are the key challenges and opportunities for the adoption of HDT in clinical practice, and what are potential trade-offs of these aspects? (RQ)

To do so, the main objective is to shed light on the technical, legal and ethical challenges and opportunities of HDT. With the aim of enhancing the efficiency of the research approach, the research question was divided into distinguishable subsets. In addition to exploring the technical, ethical, and legal aspects of HDT in clinical practice, the paper seeks to investigate their interrelationships. Thereby, the goal is to obtain a holistic picture of the current landscape in order to identify interdependencies and potential trade-offs between the different aspects. By exploring and analyzing the multifaceted aspects associated with HDT on a superordinate level, this paper finally aims to contribute to the existing body of knowledge and provide valuable insights into the potential challenges, implications and interdependencies of HDT in healthcare settings.

Structure of the Work

The remainder of this paper is structured as follows. The next section provides a theoretical background, covering key concepts and definitions such as digital twin, HDT, and related research. This is followed by a discussion of the methodology used in the present paper, including expert interviews and the coding methodology. Subsequently, the results of the expert interviews and the data analysis are presented. The principal findings, future work, and limitations are then discussed. The paper concludes with a final summary.

Background

Digital Twin

The concept of a digital counterpart to a physical product or system was first introduced by professor Grieves in 2003 within the context of Product Lifecycle Management (Grieves, 2014). Initially referred to as the "Mirrored Spaces Model", this idea describes the connection between a virtual space and a real space, each housing a system's model facilitated by data and information flow. Later on, this concept was renamed the "Digital Twin" (Rosen et al., 2015). A digital twin is defined as a living model of the physical object that can continually adapt to various operational changes using real-time sensory data while forecasting the

future of the corresponding physical object (Z. Liu et al., 2018). However, there are multiple definitions of the digital twin concept. It can be described as a multi-domain simulation (Jaensch et al., 2018), a computerized counterpart of a physical system (Kritzinger et al., n.d.), a virtual representation of what has been produced (Grieves, 2014), a virtual substitute of real-world objects (Schluse et al., 2018), an integrated simulation, and forecasting tool (Negri et al., 2017) or a linked collection of digital artifacts (Boschert et al., 2018).

A digital twin system typically comprises three main components: a real-world entity, also known as a real-world twin, a digital twin, and an interchange component. The interchange component links the other two components together and enables the digital twin to be updated as the real-world entity acts within its environment (e.g. van der Valk et al., 2020; Sharotry, Jimenez, et al., 2020; Mendi et al., 2021). The digital twin then executes embedded models which are informed by the data it receives from the real-world twin to simulate one or more characteristics (e.g., structure or behavior) of the latter's function with the goal of determining changes in the real-world twin which might produce an improvement in its operation (Miller & Spatz, 2022). The Digital Twin can carry on simulation, validation, optimization, and evaluation, and give suggestions, predictions, and controls to the real entity for people to make decisions, improve the performance, and prolong the lifecycle of the physical entity (Shengli, 2021). Thus, this technology is most often employed in industrial settings for the digital representation of real-world equipment, as it enables predictive maintenance based on physics-based models and data-driven analytics (Z. Liu et al., 2018).

Human Digital Twin

HDT, also referred to as Patient Digital Twins, are currently defined as "computer models of humans tailored to any patient to allow researchers and clinicians to monitor the patient's health for providing and test treatment protocols" (Barricelli et al., 2020). Similarly to digital twins in other areas, the HDT is a digital representation of the corresponding real-world human. The HDT may exist purely as a mathematical representation of the modeled individual or class of individuals. Alternately, the HDT may exist as a virtual entity that can be rendered within a virtual or real-world system (Miller & Spatz, 2022). A HDT can therefore be understood as an integrated model that facilitates the description, prediction, or visualization of one or more characteristics of a human or class of humans as they perform within a real-world environment. The specific human attributes to be modeled for creating a HDT often depend on the use case (Miller & Spatz, 2022). Categories of such characteristics include physical (e.g., anthropometric attributes, biomechanics attributes), physiological (e.g., heart rate, heart rate variability, muscle tension, blood oxygen level), perceptual (auditory sensitivity, visual sensitivity, color sensitivity), cognitive (e.g., knowledge, skills, abilities or aptitudes), personality (e.g., personality type, propensity to trust), emotional (e.g., levels of depression, and anxiety), ethical, and behavioral (actions taken by the individual to interact with the system) (Lewis et al., 2019).

Human Digital Twin and Traditional Digital Twin

In the realm of digital twins, HDT emerge as a noteworthy subcategory, sharing numerous similarities with their counterparts in terms of their construction and function. However, it is crucial to acknowledge the existence of distinctive differences that warrant careful consideration. These differences are examined in Shengli (2021). In both cases, models are built to replicate a physical entity in the real world by employing similar technologies from the fields of big data analytics and Artificial Intelligence (AI). They also share the same primary structure, including an interchange module for the two-way communication between the real and the virtual entity. While digital twins primarily focus on physics-based models, HDT incorporate additional dimensions such as biology and physiology, for which less information is available. This complexity arises from the need to accurately model the intricacies of human anatomy, physiology, and behavior. Moreover, the influence of the environment on humans adds another layer of complexity to these models, as humans are constantly affected by external factors to a considerably higher degree compared to machines or manufacturing systems (e.g. He et al., 2019; Alam & El Saddik, 2017; Stark et al., 2019).

In order to address the increased complexity of human systems and enable the incorporation of the physical entity's interaction with its environment, Shengli (2021) proposes the classification of HDT under "Augmented Digital Twins", the most complex level of maturity for digital twins. The concept of an augmented digital twin extends beyond the conventional digital twin framework, encompassing a complex system that models the immediate surroundings of the real-world twin and its interactions with other

digital twins (Pool, 2021). In the virtual space, virtual surroundings and other digital twins are added to facilitate communication and social interactions, reflecting the social nature of humans. Additionally, the inclusion of genetic relationships, such as family ties, acknowledges the hereditary nature of genetic information among individuals. By incorporating these factors, HDT aim to provide a more comprehensive and accurate representation of human entities, their interactions, and their environment (Shengli, 2021).

Use Cases

The term "Human Digital Twin" has been applied in diverse fields, including medicine (e.g. Chakshu et al., 2019b; Corral-Acero et al., 2020; Hirschvogel et al., 2019; Y. Liu et al., 2019a; Lutze, 2020), sports performance (Barricelli et al., 2020), manufacturing ergonomics (e.g. Caputo et al., 2019; Greco et al., 2020; Sharotry, Jimenez, et al., 2020), and product design (e.g. Onan Demirel et al., 2021; Constantinescu et al., 2019; Ma et al., 2019). A notable distinction between the different applications of this technology is that HDT applied in settings related to medicine and sports performance prioritize the understanding of systems that are internal to the human, whereas HDT employed in manufacturing and product design applications aim to model behavioral characteristics related to the real-world entity's interaction with the system or the environment (Miller & Spatz, 2022). In the context of this thesis, emphasis will be laid on HDT applications in medicinal settings.

Medical Applications

HDT applications in the field of medicine have been gaining traction over the past years as they constitute a fitting solution to dealing with the increased complexity of human lifecycle management (e.g. He et al., 2019; Alam & El Saddik, 2017; Stark et al., 2019). This new medical simulation method is expected to enhance the current medical system by employing multi-science, multi-physics, and multi-scale mechanistic and statistical models to enable proactive, accurate, and efficient public health services. HDT applications appear particularly promising in the field of personalized healthcare, as HDT can assist in dealing with the high complexity of human full-cycle management (Haleem et al., 2023). Successful implementation could enable the individualization of therapy, care, and medication by employing data-driven virtual representations of the human body, thus improving the quality and efficiency of diagnosis and treatment (Shengli, 2021). Although numerous challenges lay ahead for the creation of HDT, mostly related to the nature and complexity of human data, it is believed that implementing this technology is possible, at least from a technical point of view. More significant challenges surrounding such applications will rather relate to the social and ethical issues as well as the public perception of the subject (Shengli, 2021). This realization further motivates an expanded, spherical consideration of HDT that is not limited to their technical implementation.

Related Research

In recent years, significant advancements in the field of HDT have spurred numerous endeavors aimed at comprehending and assessing the challenges and prospects associated with their implementation. These endeavors have, to some extent, been motivated by the novelty of this technology, as it is believed that shaping the trajectory of emerging technologies is more feasible during the initial stages of their development when the transition from research to practical applications has not yet been fully realized (Y. Chen, 2017).

The evaluation of the open challenges inherent in HDT is considered indispensable to facilitate proactive intervention, ensuring the responsible development and implementation of this innovative technology (e.g. Pidgeon & Rogers-Hayden, 2007; Wilsdon & Willis, 2004; Popa et al., 2021). This work attempts to shed light on the opportunities and challenges of the clinical adoption of HDT, doing so by considering the technical, legal, and ethical perspectives. These perspectives were chosen based on the understanding of HDT as a system that uses artificial intelligence in critical infrastructure. This thesis argues that such systems should be especially trustworthy. The integration of these three fields follows the argumentation of Floridi et al. (2018) and Thiebes et al. (2021) that trustworthy artificial intelligence should include the ethical and legal perspectives, besides the technical perspective.

Technical Characteristics

In order to obtain a spherical view of HDT applications, it is crucial to consider the nature of digital twins as a technological device, as the technical, ethical, and regulatory opportunities and risks associated with HDT are largely derived from the characteristics of this innovation. According to (Popa et al., 2021), the technical changes resulting from HDT follow a linear, evolutionary pattern, amplifying existing trends rather than instigating a revolutionary shift in direction. This phenomenon can be attributed, in part, to the inherent characteristics of the digital twin as a technological amalgamation, encompassing preexisting technologies (Popa et al., 2021). Current systems for assisting doctors are called clinical decision support systems (CDSS), which are intended to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information (Sutton et al., 2020).

Opportunities and Benefits. The opportunities associated with HDT are linked to an increase in the efficiency and effectiveness of the treatment development and delivery process. The digital twin promises to address an inherent limitation of current healthcare processes, namely the subjectivity of patient and diagnostic data, such as the reporting of symptoms during the anamnesis process (Harris, 2020). Simultaneously, detailed models of this nature play a crucial role in enabling a greater individualization of patient diagnosis and treatment (Popa et al., 2021). Consequently, new individualized processes have the potential to yield significantly better results, as care and therapy are directly tailored to the profile and needs of the treated person (Huang et al., 2022). Additionally, the increased efficiency in terms of equipment maintenance, time management, and drug development can substantially reduce the costs associated with developing and delivering therapeutic processes (Popa et al., 2021).

Barriers and Dangers. The creation and operation of digital twins, regardless of the sector, heavily rely on the availability and integration of sufficient and high-quality data. Therefore, it is not surprising that the primary concerns associated with HDT, as discussed in the examined literature, revolve around the acquisition, management, and protection of patients' personal data (e.g. Popa et al., 2021; Bruynseels et al., 2018; Huang et al., 2022). The inherent fragility of models employed in digital twins poses a significant technical challenge. Particularly in the early stages of development, artificial intelligence systems must rely on existing biomedical data for learning purposes (Popa et al., 2021). This means that, similarly to most machine learning applications, digital twins are vulnerable to the "garbage in – garbage out" principle, which dictates that the performance of good models can suffer significantly if the data used to train them is not up to par (Sanders, 2017). In the case of digital twins, the detrimental effects of poor data, flawed analysis, and subsequent inaccurate representation are exacerbated by the trust placed in these models (Popa et al., 2021). Even if the quality of the acquired data can be ensured, the implementation of digital twins brings forth an increased degree of complexity and the responsibility to store and manage large datasets. Moreover, practices such as hypercollection can significantly infringe upon the fundamental rights of privacy and autonomy, while the absence of a comprehensive understanding regarding the extent of data collection frequently results in the lack of meaningful informed consent (Huang et al., 2022).

While the objectification of clinical data in the context of HDT may appear advantageous from a technical perspective, it is essential to acknowledge the potential risks associated with disregarding patients' personal viewpoints and experiential knowledge (Huang et al., 2022). Despite the progress in modeling techniques, these models remain limited by the incomplete understanding and biases of their human creators, making them susceptible to human interpretation errors and biases (e.g. Huang et al., 2022; Rich et al., 2020). Consequently, the widespread implementation of HDT in healthcare could lead to distorted perceptions of health and contribute to the exacerbation of epistemic injustice (e.g. Huang et al., 2022; Prainsack, 2017; Ruckenstein & Schu¨ll, 2017).

Even after ensuring the quality and performance of a HDT, training, updating, and operating such extensive models incur significant resource utilization, which can pose a technical and financial obstacle in the extensive adoption of digital twins (Drummond & Coulet, 2022). Additionally, even with the push for environmentally friendly artificial intelligence models, the widespread implementation of HDT would lead to a substantial volume of greenhouse gas emissions (Drummond & Coulet, 2022).

In addition to influencing the technical implementation of HDT in healthcare, the aforementioned challenges assume greater significance when their socioethical and regulatory dimensions, which will be discussed in the following sections, are taken into account.

Regulatory Characteristics

The regulatory aspects of HDT implementation are directly derived from the technical requirements for the deployment of the technology and possible complications that can arise therethrough. Contrary to the technological advancements in the fields, which can be viewed as significant drivers for the creation of HDT-powered applications, regulatory standards mostly serve to protect the relevant stakeholders from a potential misuse of this technology. At the same time, difficulty in regulating the different aspects and of HDT based on the specific use case brings about further regulatory consideration and calls for increased caution from the side of the regulators.

Barriers and Dangers. As presented in the previous section, access to a sufficient amount of relevant data is essential for the development and operation of digital twins. However, fulfilling this requirement may prove difficult in the case of HDT, as personal data are protected by relevant legislation -especially in the case of health dataand access or use of such data is often highly restricted (e.g. Bagaria et al., 2020; Yuan & Li, 2019).

The delicate nature of the collected data and the difficulty of tracking it raises the significant question of where the responsibility of protecting said data lies. Especially as twins are created for specific use cases, assigning legal responsibility becomes significantly more complex (Popa et al., 2021). Furthermore, the use of HDT applications in a medicinal context raises vital questions on where the responsibility lies during the diagnostic process and to what extent medical decisions can be based on information provided by a digital twin of the patient (Popa et al., 2021). Such queries are indicative of the risks involved in the institutional changes brought about by the application of HDT.

Socio-Ethical Characteristics

As with any significant technological innovation, the new possibilities it introduces on a technical level have the potential to greatly influence the current landscape and affect all the parties involved. Whether this effect is positive or negative, however, depends on the implementation of the new means in each scenario. In the case of HDT, the technical advancements discussed in the previous section can generate tangible benefits for all stakeholders of healthcare system. However, a potential misuse or abusive of this technology could pose considerable threats for patients and doctors alike. Both of these outcomes should therefore be taken intro consideration when reviewing the ethical impact of HDT systems.

Opportunities and Benefits. The implementation of HDT in healthcare presents significant socioethical benefits that are closely intertwined with the technical advancements facilitated by this technology. Personalized treatment methods made possible by HDT offer the potential for more effective treatments and improved patient outcomes, while enhanced efficiency in healthcare processes can contribute to cost reduction, thereby increasing accessibility to a broader population (Popa et al., 2021). Furthermore, HDT empower patients by enabling them to play a more active role in their healthcare decisions, fostering informed choices. These potential benefits, both at an individual and societal level, encompass overall enhancements in the quality of healthcare, with some proponents even positing it as the foremost priority in the field (Popa et al., 2021).

Barriers and Dangers. The acquisition, integration, and utilization of extensive datasets, deemed indispensable for the establishment and operation of digital twins as discussed earlier, give rise to noteworthy ethical and societal dilemmas. Among these challenges, the foremost pertains to the preservation of patient privacy and the safeguarding personal data.

The academic literature highlights concerns regarding the implications of private organizations, such as healthcare institutions or insurance companies, having access to detailed representations of individuals' characteristics, including biological, genetic, physical, and lifestyle information (Lewis et al., 2019). The issue of privacy infringement is a significant concern, although some argue that the benefits of technological innovation outweigh the costs, advocating for a balance between privacy and technical advancement (e.g. Popa et al., 2021; Bruynseels et al., 2018). The risks of security breaches and data loss are heightened with the advent of digital twins, potentially amplifying rather than resolving such hazards (e.g. Huang et al., 2022; Martin et al., 2017). Breaches of privacy only serve as triggers for a more fundamental issue where individuals' own data can limit their freedom. The phenomenon, referred to as "big data discrimination" in

the literature, arises due to the overwhelming volume of information generated, making it nearly impossible for individuals to determine their own privacy boundaries (Popa et al., 2021).

On a larger, societal scale, while proponents argue that the proper implementation of HDT can help alleviate inequalities (e.g. Popa et al., 2021; Bruynseels et al., 2018), a multitude of concerns have been raised, making it highly probable for this technology to have a detrimental impact on reducing various forms of discrimination. The implementation of HDT technology carries the potential to contribute to inequality through multiple channels. Firstly, its limited accessibility and inadequate health insurance coverage can widen existing socio-economic disparities (Popa et al., 2021). Secondly, the concentration of its implementation in affluent North-Western countries, characterized by advanced research and development capabilities and intellectual property rights, may further exacerbate inequality (Popa et al., 2021). Lastly, the presence of inherent biases within the healthcare system, often tailored to the needs of the white male population, can be perpetuated through the design of digital twins, thereby resulting in disparities (e.g. Huang et al., 2022; Obermeyer et al., 2019).

In addition to the practical social implications associated with digital twin technology, the potential applications of digital twins give rise to ethical and, to some extent, philosophical inquiries, such as discerning the distinctions between therapy, preventive care, and enhancement (Giubilini & Sanyal, 2015). This differentiation is prompted by the necessity for heightened moral consideration when the fundamental objective of medical interventions is modified (Bruynseels et al., 2018). Neglecting to adequately address this concern also entails the risk of human enhancement leading to societal divisions and the emergence of distinct classes of individuals, thereby disrupting established democratic institutions (e.g. de Sio et al., 2016; van den hoven et al., 2011). Consequently, it becomes evident that despite their abstract nature, addressing these questions holds pivotal importance in preserving social structures and effectively navigating the transformative changes that this technology may bring about on a broader societal scale.

Method

In complex research fields, such as healthcare, it is important to consider the changing interrelationships within the system rather than focusing solely on isolated parts (Cristancho, 2016). Expert interviews is a qualitative research method that allows for in-depth insights from individuals with a high amount of knowledge and experience in the area of interest (Bogner et al., 2009). These experts often work at the forefront of their respective fields and are familiar with novel trends and potential future developments (Bogner et al., 2009). Expert interviews can, therefore, be particularly suited for research upon a complex and novel problem according to Bogner et al. (2009), making this method appropriate for researching the adoption of HDT in healthcare as the current informational landscape remains limited, preventing a comprehensive overview. In such cases, expert interviews offer a valuable avenue for gathering insights from specialists spanning diverse domains, enabling the aggregation of information into a cohesive framework. Within the scope of this thesis, experts will be interviewed in order to shed light on the current and future state of HDT development and implementation from a technical, legal, and ethical perspective. By doing so, the goal is to gain a holistic understanding of the trade-offs and interdependencies of challenges and opportunities within and between these perspectives.

However, as the interviews progressed and we were able to extract definitive insights from the grouped information, we came to the realisation that simply classifying these facets into opportunities and challenges within the three distinct domains doesn't fully encapsulate their intricate nature. For instance, within the ambit of technical opportunities, there are factors (1) inherently signifying the benefits of HDT, like advancements in treatment methodologies. Conversely, (2) external factors present opportunities for the integration of HDT, as exemplified by the digitization of medical practices. Similar principles extend to challenges, encompassing (3) adverse aspects or dangers emerging from the implementation of HDT, such as potential threats to privacy. Moreover, (4) barriers obstructing the adoption of HDT exist, including considerations related to legislation. In the subsequent sections of this study, we will dissect the opportunities and challenges intrinsic to each of the three domains: technological, legal, and ethical as follows:

- Opportunities → (1) benefits, (2) opportunities
- Challenges → (3) dangers, (4) barriers

Recruitment and Participants

In the first step, the target group of experts for the study was determined. It was decided to include researchers, practitioners, and industry professionals who have expertise in human digital twins in healthcare or digital health in general. Experts were selected based on the following inclusion criteria:

- Minimum of 1 year of experience in the field of HDT in healthcare or related fields of digital health
- Demonstrated expertise through relevant publications, conference presentations, or professional affiliations
- Diversity in professional backgrounds and perspectives to capture a comprehensive range of insights on technical, ethical, and legal aspects

Table 1 illustrates the information participants provided about themselves, their institution, and their tasks and experiences in the field of healthcare.

ID	Gender	Institution	Profession	Experience	Group
1	Female	Law office; healthcare-investments fonds	Lawyer specialized in M&A transactions and data privacy concerns; management of startup healthcare investments fonds	12 years	Legal
2	Male	Software development based on artificial intelligence for medical diagnostics	Quality management/project management with a focus on medical device registration	2 years	Legal/ Ethical
3	Male	University in the field of medical informatics	Professor of medical data science with a focus on pattern recognition, machine learning, sensor data analysis	6 years	Technical
4	Male	Software development for clinical decision support systems	Executive Director & Head of Technology	4 years	Technical
5	Female	Internal medicine department of a university clinic	Physician in the internal medicine department of a university clinic, participating in a project for personalized medicine	3 years	Technical / Ethical
6	Male	Investment company with focus on life science and medical technology investments, private equity	Investment manager with focus on life science & services, diagnostics & tools and digital health	10 years	Technical / Legal
7	Male	Multinational software development company with focus on virtual human modeling	Senior director for the development of advanced, multiscale computational models of human organs and body systems for use in translational medicine; responsible for life science incubator of startup healthcare companies	40 years	Technical / Legal

Table 1. Participant Information

Purposeful sampling was used to identify potential experts who met the inclusion criteria. Initial participants were selected on the basis of their expertise in the field. The sample size was determined based on the principle of data saturation, ensuring that a sufficient number of interviews were conducted to capture different perspectives and come as close to theoretical saturation with regard to the research question at hand as possible within the framework of this paper. A total of 7 experts were interviewed, which was deemed adequate for achieving a comprehensive picture from different perspectives. Experts were contacted via email or through professional networks, conferences, and research institutions. The purpose of the study, its scope, and the voluntary nature of participation were clearly explained. Additionally, experts were informed about the confidentiality and anonymity of their responses. It was further emphasized that they had the right to withdraw at any time.

Design

The interview guideline development was based on the research question at hand in order to investigate key challenges, opportunities, and trade-offs regarding ethical, technological, and legal aspects of HDT. Thus, a preliminary literature review was conducted to inform the development of the interview guide. Relevant research questions and areas of inquiry were identified to ensure a semi-structured and focused interview process (Myers, 2013). Based on the research objectives, the following broad interview topics were identified:

- Potential use cases for digital healthcare/HDT in healthcare
- Risks and chances of HDT implementation in healthcare
- Identifying drivers and inhibitors for HDT adoption in healthcare
- Exploring use cases and experiences with HDT in healthcare, as well as potential reasons for its usage or non-usage
- Assessing the current maturity of HDT in organizations
- Exploring aspects like relative advantage, compatibility, complexity, observability, and trialability of HDT adoption in healthcare

All topics relate to ethical, technical, and legal aspects, with the focus shifting depending on the expertise of the interviewees. A set of open-ended interview questions was developed for each broad topic. The questions were designed to elicit detailed responses from the experts, encouraging them to share their knowledge, experiences, and perspectives. Semi-structured interviews were conducted with the participating experts.

Data Collection

The interviews were conducted through video conferencing platforms and audio-recorded with the experts' consent. Each interview lasted approximately 45 to 60 minutes in total, allowing sufficient time for in-depth discussions while respecting the experts' time constraints. Two members of the research team were present during the interview session. A moderator facilitated the session as well as the discussion and a co-moderator or note-taker was present to document non-verbal cues, follow-up questions, and any additional contextual information that may be relevant to the analysis. Probing and follow-up questions from a deep-dive question set were used when necessary to clarify or deepen the experts' responses. Throughout the data collection process, the interviews were continuously reviewed and assessed in order to use gained insights in subsequent interviews. The session began with an introduction of the research team, the purpose of the study, and the establishment of ground rules for participation. The participants were then asked to introduce themselves and their institution and briefly explain the responsibilities associated with their job. Subsequently, the experts were encouraged to actively respond to the questions, engage in the discussions, and express their viewpoints. Interviewees were invited to provide feedback and recommendations on the study as a whole. This encompassed their impressions of the research design, methodology, and overall execution. Additionally, experts were encouraged to share their thoughts on the study's potential implications and applications in practical settings.

Data Analysis

A total of seven experts with different backgrounds and professional affiliations took part in the interviews. Six of the interviews were conducted in German and one in English. However, the transcripts were translated from German to English within this paper in order to facilitate citations.

The audio recordings of the interviews were transcribed verbatim to capture the content and nuances of the experts' responses using openAI's "whispers" speech recognition system with the large language model (OpenAI, 2023). Additionally, the transcripts were cross-checked with the interviewer's notes to ensure their accuracy and completeness. The anonymized transcripts are available upon request from the authors.

Subsequently, a qualitative thematic analysis approach was employed to identify patterns, themes, and categories within the interview data. The transcripts were coded using a combination of open, axial, and deductive coding.

Open Coding

Myers (2013) proposes open coding as the first stage of data analysis in grounded theory. With open coding, all the data is carefully examined and analyzed line by line. The goal of this process is to break down the data into smaller fragments and identify the underlying concepts that represent important occurrences or incidents related to the phenomenon being studied. Similar segments of text are identified, and they are then labeled and grouped together, creating what is referred to as "codes." This coding process involves a thorough comparison of each incident, event, quote, and instance that has been collected during the data collection phase, with a focus on identifying both similarities and differences among them (Strauss & Corbin, 1998). As an example of a coded textual segment, we highlighted "And then, as a matter of principle, European data protection law stipulates a so-called prohibition with a reservation of consent. This means that the General Data Protection Regulation first says that I am not allowed to process any personal data unless I have a justification." (Transcript 1, paragraph 10) as "legal challenge" in a first step.

Axial Coding

Subsequently, axial coding was applied to create subcodes for already existing codes. Axial coding refers to "the process of relating categories to their subcategories, termed "axial" because coding occurs around the axis of a category, linking categories at the level of properties and dimensions" (Strauss & Corbin, 1990). With this coding technique, the codes created in open coding were analyzed and transformed into new codes that display conditions, contexts, and consequences. This enables the investigation of relationships between concepts and categories that have been coded in the open coding process (Strauss & Corbin, 1990).

We used axial coding in order to gain a better understanding of the technical, legal, and ethical characteristics of the implementation of HDT in healthcare. For example for the textual segment displayed above, we applied the code "data" as a condition for the code "legal challenge".

Axial Codes for Technical Characteristics

- **benefits** understand interactions, methods, CDSS, treatment
- **opportunities** collaboration, algo learning
- **dangers** explainability
- **barriers** infrastructure, user adaption, unclear development, incentives, data availability

Axial Codes for Legal Characteristics

- **benefits** privacy
- **opportunities** regulatory, data provision
- **dangers** n.a.
- **barriers** guidelines, complexity, bureaucracy, data, service provider, regulatorics

Axial Codes for Ethical Characteristics

- **benefits** pre-testing, explainability

- **opportunities** n.a.
- **dangers** no predefinition, misclassification, human substitute, data misuse
- **barriers** n.a.

Within the technical domain, a total of 12 axial codes were established. These comprised two axial codes denoting opportunities, four indicating benefits, one highlighting dangers, and five underscoring barriers. In the legal sphere, nine axial codes were created, encompassing two axial codes related to opportunities, one emphasizing benefits, and six spotlighting barriers. In the context of ethical considerations, a total of six axial codes were formulated, with two axial codes associated with opportunities and four aligned with potential dangers. This amounts to 27 axial codes that hold significance within the purview of our research, spanning the domains of ethical, technical, and legal considerations.

Deductive Coding

Having the transcripts coded with open and axial coding, we implemented deductive coding. In the process known as deductive coding, researchers utilize and apply pre-existing concepts to the data analysis (Linneberg & Korsgaard, 2019). In certain cases, there might already be theories or prior research available on a particular phenomenon, but they could be incomplete or require more detailed explanations. The primary objective of the deductive approach is to either validate or expand a theoretical framework or theory conceptually. To begin this process, researchers first identify key concepts and use them as initial coding categories (Linneberg & Korsgaard, 2019). Then, they establish operational definitions for each coding category by drawing upon relevant theories or prior work. Subsequently, all text passages are coded using the predefined codes (Linneberg & Korsgaard, 2019). This systematic approach ensures a focused and structured analysis of the data based on existing knowledge and concepts (Hsieh & Shannon, 2005). Since HDT is a novel phenomenon, concepts found in the literature may be derived from fundamental research. Validation and expansion of these concepts through deductive coding may help to further develop them (Linneberg & Korsgaard, 2019). Due to the novelty of HDT, we focused on a few selected concepts.

From the literature, two key concepts of advantages and two key concepts of disadvantages were synthesized. The codes with their respective concept and explanation are provided in the following.

- personalization: HDT enables more personalized and precise healthcare
- prior assessment: HDT can help to enhance treatment quality by enabling physicians to evaluate patients' HDT before initiating treatment or therapy
- regulation: Regulation that complicate the implementation of HDT
- explainability: Potential danger of AI and algorithms as a black box, complicating explainability of outputs produced by HDT

The validation and extension of these concepts from deductive coding is provided in Table 3 in the discussion section. The codes "regulation" and "explainability" already exist as deductive codes and as axial codes. This is because these two concepts are prevalent in the literature as well as in our interview results.

Positive and Negative Implications of HDT in Healthcare

Table 2 provides an overview of the insights generated through the interviews by summarizing the input received from the interviewed expert into single points and categorizing these based on the features of HDT they are associated with. For each of the relevant fields associated with HDT implementation (technical, legal, ethical), the four main topics introduced in the methodology section were discussed. These topics include opportunities enabling the implementation of HDT systems (coded green), benefits a potential application of this technology could entail (coded blue), barriers hindering the realization of HDT-powered tools (coded orange), and dangers that arise from potential misuse of HDT.

The identified insights are further categorized based on the characteristics of HDT technology from which they originate. Thus, most of the mentioned opportunities, benefits, barriers, or dangers affecting the current and future state of HDT implementation are caused either by the primary use of HDT in medical applications, the dependency of developed models on the available data, as well as the advanced nature of this technology.

The following sections will dive deeper into each examined field and present the generated insights in more detail, including specific quotes from the interview transcripts. Reflection upon the generated results and presentation of the principal findings is then found in the discussion section.

Characteristics	**Technical**	**Legal**	**Ethical**
Medical application	Collaboration	Updated regulation	Improved patient outcomes
	CDSS	Patient data anonymization	More efficient work
	Increased accuracy	Medical product certification	Patient transparency
	Spherical approach	Bureaucracy & costs	Misclassification
		Regulatory restrictions	Doctor substitution
Based on data	Data availability	Data provision	Data misuse
		Data management	Coded bias
Advanced technologies	Optimization of existing technologies	High complexity	
	Black box AI	Third party services	
	User knowledge Infrastructure		
Legend			
Opportunity	Benefit	Barrier	Danger

Table 2. Overview of Identified Topics

Technical Aspects

Technical aspects involve the opportunities enabling the implementation of HDT in healthcare such as enhanced collaboration amongst professionals from diverse backgrounds and the development of advanced learning algorithms, along with main barriers hindering the implementation, for example, the missing infrastructure, insufficient data availability, and users that need to adapt to the new technology. Furthermore, the benefits and dangers that would come with the implementation mentioned by the interviewees are listed.

“Like a complex jet, we decompose the body into its different systems. So cardiologists study the heart, neurologists study the brain, then you go inside the heart, and then you have electrophysiologists who study the electrical signal, and people study the tissue mechanics and hemodynamics. So we break down these complex systems, and so as patients, we see all these different specialists that try to put it all together to understand what’s happening in our human body. And the doctors don’t have many tools to be able to do

that. So I thought, well, if we can give engineers the ability to take their specialties and bring it all together to understand how the plane would function, why can't we do the same for the human body? So give the doctors the benefit of bringing together all of their specialists and bring it together into a virtual twin." (Transcript 7, Paragraph 6)

Benefits

The benefits of the implementation of HDT in healthcare are various, but our interviewees mainly focused on three of them: enhancing Clinical Decision Support Systems (CDSS), enabling better treatment, and being able to better understand interactions of various systems within the human body (Transcripts 2, 4, 5, 6, 7).

"Again, if you think about it from an engineering perspective, the body's like a system. The blood comes out of the heart. So the output of the blood is the input to the lungs. The lung performs its function. The output is oxygenated blood, goes back to the heart. So that's an input to the other part of the heart. The output then goes to the body and circulates." (Transcript 7, Paragraph 17)

Clinical Decision Support System (CDSS). Using HDT in healthcare has the potential to create objective patient information through measurements as opposed to subjective descriptions provided by the patients during the anamnesis process. Neutral and objective information gathering in the medical field is made possible. This digitized information then can be used for preventive and therapeutic interventions (Transcript 6, Paragraphs 9, 15).

Furthermore, HDT in healthcare has the potential to avoid overdiagnosis by considering age-related changes in reference values, and having personalized reference values for the elderly or across a person's lifetime (Transcript 5, Paragraphs 11-12, 26). With these reference values, diagnostics can be supported as well as better planning of the actual treatment (Transcript 5, Paragraph 36). Additionally, digital twins and AI in the medical sector have the potential to reduce oversight errors and improve patient safety (Transcript 5, Paragraph 40).

Moreover, a potential role of HDT could be in addressing challenges like the shortage of medical professionals and the increasing complexity of medical knowledge, enhancing healthcare by facilitating earlier disease detection, prevention, and personalized treatment (Transcript 4, Paragraph 26). Subsequently highlighted is the combination of existing medical knowledge with AI enhancements, to aid time-crunched decisions in busy environments like emergency departments. Nevertheless, the aim is not to replace medical professionals but to provide them with supportive tools for improved decision-making, supporting both diagnostic and therapeutic settings (Transcript 4, Paragraph 38).

A potential integration of real-time patient data, acquired through sensors, into virtual heart models could enable the training of artificial intelligence models on these HDT to interpret realworld patient data (Transcript 7, Paragraph 20-21).

Furthermore, software in medical diagnostics can provide supplementary information rather than definitive conclusions, as software-based support offers an additional layer of information to medical practitioners (Transcript 2, Paragraph 25).

Treatment. Not only would HDT enhance the CDSS, but also the treatment process itself. Surgical fields, particularly areas like cardiology, could greatly benefit from pre-surgical planning using digital twins. Visualizing and planning surgical procedures in advance can be an advantage, enhancing decision-making and organization, especially in fast-paced environments such as emergency departments. Also emphasized is that the goal of such technology is not to replace doctors but to support them effectively (Transcript 5, Paragraph 37-38).

As mentioned before, the use of existing medical knowledge combined with AI enhancements to create understandable and transparent models can be advisable. Moreover, in time-sensitive situations, such as emergency departments, where many patients require attention and decisions could be overlooked, this approach might be indispensable in the future. Emphasized is again the supportive nature of these tools, aiming to assist medical professionals in making informed decisions (Transcript 4, Paragraph 38).

Already nowadays, digital twins are employed for pre-surgical planning in heart surgeries, using virtual surgeries and models to determine the optimal device size and placement in the patient's heart. According

to the interviewee, their approach has reportedly demonstrated 100 % accuracy in device sizing, potentially reducing the need for corrective procedures and purportedly saving around 25 % of operating room time. It is suggested that such successful cases will encourage the medical industry to recognize the economic and medical benefits of investing in comprehensive planning (Transcript 7, Paragraph 25).

Highlighted is also the practical significance of digital twins in complex surgeries, particularly those involving newborn children. As surgeons have very limited time to make critical decisions during newborn surgeries, virtual twins of patients can be used to explore different surgical techniques in advance, allowing surgeons to pre-plan and make well-informed decisions within their limited time frame. In the USA, prominent children's hospitals already have adopted this approach as standard practice (Transcript 7, Paragraph 19).

Understanding Interactions. Modelling various organs and systems within the human body and creating a comprehensive model spanning physical and biochemical aspects across scales can help understand organ interactions during health crises like COVID-19, where an isolated view of the lungs would potentially have led to overlooking interrelationships. "So we can have the heart working with the lung, working with the kidney. So we know how to connect" (Transcript 7, Paragraphs 17). Reducing the complexity of how organs interact with each other could be made possible by creating the organs as black boxes, that just display their function without the need to actually model the organ (Transcript 7, Paragraphs 17-19).

Opportunities

The respondents primarily emphasized two key opportunities that hold the potential to drive the advancement of HDT within the healthcare sector: collaboration among healthcare specialists, engineers and data scientists and the development and implementation of various algorithmic learning approaches.

Collaboration. "[…] it was perceived […] that something like the heart was way too complex to represent on a computer. And I think it was partly because they assumed that someone would have to know it all and program it into the computer. And my approach was to convince people who knew the individual parts to work together." (Transcript 7, Paragraph 10)

Interviewees underscore the essence of effective teamwork within multidisciplinary scientific undertakings. Collaborative process is exemplified as specialized experts amalgamate their proficiencies to create a comprehensive virtual representation of the human heart. This collective approach yields a result surpassing individual contributions, enabling the construction of a fullscale computerized human heart model which in future can be elaborated to a whole HDT. Such collaborative strides offer an expansive platform for investigating diseases and potential remedies (Transcript 7, Paragraph 6).

Furthermore, it is important to dismantle disciplinary barriers for seamless collaborative accomplishments. The notion that it is not possible to encapsulate the complexity of a heart within a computer is not possible is refuted. Instead of relying on a sole individual's grasp, the strategy shifts towards experts in distinct facets collaborating synergistically. This harmonious collaboration encourages knowledge amalgamation, leading to a more comprehensive portrayal of the virtual heart (Transcript 7, Paragraph 10).

Additionally, the intricacies posed by cross-disciplinary collaboration and the strategies harnessed to navigate them are highlighted. The significance of creating a shared language to bridge the gap between different fields, like engineering and medicine, is accentuated. A shared parlance, primarily facilitated through visualization, emerges as a pivotal instrument for comprehending and communicating amidst experts with disparate terminologies. This accentuates visualization's role in bridging disparities and fostering mutual comprehension in multidisciplinary projects (Transcript 7, Paragraph 12).

Besides, the significance of varied viewpoints in assessing potential risks and usability predicaments linked to software development is underscored. Through an assemblage of a varied panel, encompassing developers, usability engineers, and risk assessors, a thorough evaluation of the product's potential challenges is attainable. This approach aligns with fundamental collaborative risk management and quality assurance principles, wherein multiple outlooks ensure a comprehensive appraisal, thwarting single-person biases (Transcript 2, Paragraph 27).

Algorithmic Learning. Advanced research in pattern recognition and its applications can support the implementation of HDT in healthcare. Contemporary research on pattern recognition includes the

utilization of advanced techniques for extrapolating meaningful insights from human data, notably for medical diagnoses and personalized health interventions.

There is an ongoing drive to enhance the interpretability of intricate deep learning models, addressing a pivotal concern in the field. A comprehensive tripartite research strategy consists of the synthesis of augmented training data, the transference of pre-trained models across divergent domains, and a deep dive into the influence of input parameters (Transcript 3, Paragraph 9). Commencing with conventional pattern recognition methods before resorting to the complexity of deep learning is advocated, with an underpinning emphasis on lucid procedural steps. This strategy seeks to illuminate the often intricate nuances of deep learning practices. A parameter-driven rationale to arbitrate between classical methodologies and deep learning paradigms, guided by the volume of annotated training data—a cornerstone factor in this determination is supported as well (Transcript 3, Paragraphs 11,13). Furthermore, the potential of deep learning to generate new insights even when used for confirmatory purposes is mentioned (Transcript 3, Paragraph 15).

A potential for generating bias through improper data augmentation does exist, but methods to augment data like probabilistic methods and Generative Adversarial Networks (GANs) aim to prevent the introduction of new biases during data augmentation (Transcript 3, Paragraph 19).

Dangers

While it becomes clear that the use of HDT in healthcare has several benefits, there remains a significant danger that was mentioned in three of the interviews that were led, consisting of the explainability of the data outputs.

While a potential benefit of AI and HDT technology is handling larger datasets and producing improved outcomes, there exist concerns about the difficulty in comprehending and verifying the suggestions or outputs generated by AI systems, especially when dealing with complex models like human replicas. The importance of finding a balance between AI-driven capabilities and human understanding in future applications of HDT is emphasized, using for example Knowledge Graphs in achieving this equilibrium (Transcript 4, Paragraph 35).

Also stressed is the significance of verifying machine outputs due to their unwavering confidence in their results. Computers can generate outputs, also referred to as hallucinations, with high confidence, even if those outputs are not accurate or realistic. That´s why the continued need for human oversight and intervention is strongly emphasized (Transcript 7, Paragraph 23).

Despite their advancements and careful assessments, there are certain limitations with regard to testing and evaluating AI systems. Using those systems, for example in a HDT environment, there is an immense number of possibilities in high-dimensional spaces that cannot be exhaustively tested. Inexplicable deep learning models might function seamlessly for years but unexpectedly cause a catastrophic event (Transcript 3, Paragraph 27).

“And at some point, a car that drives autonomously, based on an inexplicable, i.e. not entirely understandable deep learning model, works fantastically for years and at some point it goes bang [and drives into a children´s group]. So somehow it will get a constellation of data where no one will understand why a wrong classification was made. Maybe ten children were dressed in such a unique way that it was the only possibility that this miss-classification would occur.” (Transcript 3, Paragraph 27, translated)

A challenge consists of understanding why such an incident occurred when dealing with complex models and rare combinations of parameters. A comparison to highly powerful yet rare disasters is drawn, where the severity necessitates thorough contemplation (Transcript 3, Paragraph 27).

Barriers

There exist various categories of barriers that possess the capacity to impede the pace of implementation and application of HDT in healthcare. These include challenges arising from inadequate or nonexistent infrastructure, restricted availability of systemic data, uncertainties associated with user adaptation, ambiguity surrounding future developmental pathways, and the absence of effective incentivization

mechanisms. These impediments are currently obstructing progress and possess the capability to continue doing so in the future.

Infrastructure. One of the barriers that seems to be hindering the implementation of HDT in healthcare significantly is the missing infrastructure, consisting of guidelines, protocols and hardware.

"I think it's time to set up a concept that takes into account all the medical stations you go through in your life [...]. Building a lifelong data collection from that is still far in the future, I think, in terms of infrastructure, data structure." (Transcript 5, Paragraph 28)

There is a variability in the timeline for integrating research results into clinical practice for HDT, as the speed of implementation depends on individual circumstances and the specific study involved. Definitive guidelines or protocols for swiftly incorporating viable research outcomes into clinical settings are virtually non-existent. Until now, factors such as the publication location, audience size, and motivation of relevant professional societies influence the adoption rate of new information, not supported by a provided infrastructure or guidline (Transcript 5, Paragraph 33-34).

Digitization and digital technologies in the medical sector are still not state of the art, with practitioners expressing concerns related to reliance on internet connectivity and potential disruptions to medical practices. With no common practices concerning data security patient data is still vulnerable to cyberattacks. The importance of ensuring the secure operation of digital systems, especially in healthcare environments, is emphasized (Transcript 4, Paragraph 47). Pointed out was also that the current digital infrastructure in medical settings, such as doctor's offices and hospitals, is primarily geared towards 2D imaging and record-keeping. The need for substantial changes to accommodate the computational complexities of 3D data in medical practices is highlighted, offering a potential business value of incorporating digital technology into medicine, but practical and financial challenges still impede widespread adoption (Transcript 7, Paragraph 25).

The future potential of creating a lifelong data collection from various medical stations that individuals pass through in their lives is envisioned by another interviewee, starting from childhood and encompassing visits to pediatricians, general practitioners, and intermittent hospital stays. But current infrastructural and data structure challenges restrict achieving this goal (Transcript 5, Paragraph 28).

The complexities of data digitization are mentioned once again by another interviewee, referring to the necessity of accessible digital source data and the requirement for sufficient IT expertise and resources on the client side for successful integration (Transcript 4, Paragraph 12).

The importance of designing user interfaces and implementing measures to ensure the reliability and safety of HDT systems is brought up, advocating for comprehensive training for users, thorough documentation, and quality management to address potential issues and risks in the system's operation (Transcript 2, Paragraph 27).

Data Availability. Another significant barrier for the adoption of HDT in healthcare is the availability of data.

"Ultimately, which would already help if one no longer sent PDFs and faxes, but recorded the data in a structured way." (Transcript 4, Paragraph 40)

It was already stated before that creating a comprehensive lifelong data collection from various medical stations a patient interacts with throughout his or her life is a very complex task (Transcript 5, Paragraph 28). A prerequisite for this collection would be the availability of source data in digital form and the need for sufficient IT expertise and resources (Transcript 4, Paragraph 12).

Also challenging is the standardization within the context of using medical data for technologies like HDT. The interpretation of terms and standards can be quite complex, with current examples like medical coding systems (ICD-10) that can have discrepancies between their intended meaning and their actual use. Particularly within healthcare facilities, there is an ongoing optimization which is frequently geared towards enhancing billing procedures to facilitate increased invoicing. However, such an approach can potentially engender complications when employed as input for machine learning algorithms as its optimal output is not always in favor of the patient´s health. In the context of healthcare data and machine learning, the mitigation of biases and the issue of inadequate sample sizes emerge as significant challenges, warranting careful consideration (Transcript 4, Paragraph 22).

Highlighted is also the need for structured data capture and integration of various medical devices and measurements into the healthcare system. A challenge is to connect different systems and data sources to provide physicians with a holistic view of a patient's health data, emphasizing on the importance of streamlined integration and structured data for effective utilization (Transcript 4, Paragraph 40).

Also mentioned is the status of digitization efforts in Germany's healthcare system, with open resistance to digital tools and outdated practices like using fax machines. Significant advancements in digitization and supporting software within the healthcare sector is needed (Transcript 2, Paragraph 29). Centralizing medical data is indispensable to avoid losing potential and to prevent avoidable health issues. A centralized repository of medical data that provides a comprehensive and cumulative view of a patient's health history could enhance the quality of care and medical decisions significantly (Transcript 2, Paragraph 38).

User Adaption. Another barrier in the implementation process is the concept of adaptation risks associated with the introduction of new technologies, such as HDT. Users, particularly those who are less tech-savvy or busy, might not adopt the technology due to various factors. Potential challenges could involve convincing users, such as doctors, to integrate new technologies into their practices, which could lead to wasted time and resources (Transcript 6, Paragraphs 9, 19).

Addressing user experience issues from the perspective of the end users, including physicians who may not be well-versed in IT solutions, is a important step in the implementation process. To ensure that less tech-savvy users can effectively understand and use the technology is of paramount importance, as the goal is to achieve meaningful and inclusive adoption across different user profiles (Transcript 4, Paragraph 47).

Unclear Development. Highlighted as another barrier is also the necessity of assessing potential disruptions caused by emerging technologies in the healthcare sector, with a need to analyze how various technologies, including AI and hardware solutions, could potentially impact existing decisions. It is important to maintain a balance between technological advancements and existing practices (Transcript 6, Paragraph 26).

Incentives. Adopting and implementing new technologies like HDT always comes with financial risks that need to be addressed. Highlighted is the need for funding to support the time and resources required for the technology's successful integration and operation. These financial considerations are inherent in the venture cycle of technological innovations (Transcript 6, Paragraph 19).

The challenges of fostering collaboration and alignment among different manufacturers in the medical technology market are also discussed, acknowledging that individual manufacturers often prioritize their own products and profits, which can make it difficult to establish a shared incentive to work together for the benefit of the entire market (Transcript 2, Paragraph 38).

Legal Aspects

Using the same approach as for the technical aspects, different areas of the legal aspects are presented. Legal Aspects mostly focus on regulation and data provision as both opportunities enabling and barriers hindering the implementation of HDT in healthcare.

"In general, it is very important to create a legal framework, especially in such areas. Especially with regard to AI models. You can see how strongly, how quickly things like ChatGPT are developing, becoming established, and then also being used, even for things where they should perhaps rather not be used. And of course, the medical sector is a very important field, where it is also a matter of achieving less or a high level of reliability and quality." (Transcript 3, Paragraph 23, translated)

Benefits

An operational model adopted for a global project in HDT is discussed, which involves collaboration of diverse stakeholders. An overarching concern revolves around upholding data confidentiality and security within this collaborative framework. The model entails the establishment of a central data hub, functioning as an intermediary for data exchange. This approach avoids direct sharing of sensitive patient data among participants. Instead, the emphasis is on gathering physiological and diagnostic insights, with a conscious effort to detach data from individual patient identities. This operational approach aligns with established security protocols present in healthcare institutions thus, maintaining data integrity. Importantly, the

technological infrastructure orchestrating this process operates in an agnostic manner (Transcript 7, Paragraph 14).

Opportunities

Highlighted are opportunities in implementing Germany's Digital Health Supply Act (Digitales Versorgungsgesetz), which aims at accelerating healthcare innovation (Transcript 6, Paragraph 21). The importance of legal frameworks for medical AI, as it is often implemented in HDT, is emphasized for ensuring reliability and accountability (Transcript 4, Paragraph 36). Besides, complexities of law in medical studies are stressed, including considerations of consent and ethics boards (Transcript 3, Paragraph 23). Additionally, pragmatic solutions are proposed, like having researchers temporarily affiliate with medical institutions to tackle legal data processing hurdles (Transcript 3, Paragraph 23).

Barriers

"This General Data Protection Regulation has taken over quite essential basic ideas from our former [German] Federal Data Protection Act. And I think we should ask ourselves as the European Union, to what extent is this still realistic? Because if you make laws that, on the one hand, are not applicable across the board and, on the other hand, lead to us, as the European Union, uniformly cutting off our water, so to speak, so that we ourselves can no longer participate in competition, then I think that's very difficult." (Transcript 1, Paragraph 14, translated)

Guidelines. With regard to data sharing, the challenges associated with distributing medical data across different organizations within Germany and even more so across international borders are pointed out. Negotiations with various legal and ethical bodies, especially in the case of cross-border collaboration, prove to be an obstacle (Transcript 5, Paragraphs 13-14).

Remote participation in studies and projects involving HDT is acknowledged to be a difficult task. This difficulty arises from unclear regulatory frameworks and the initial lack of clarity in legal guidelines for remote studies (Transcript 5, Paragraph 20).

The impact of Germany's healthcare regulation on digitalization is also underlined. Paradoxically, this regulation hinders standardized digital infrastructure growth, unlike regions with government mandates for open standards, fostering progressive digital health strategies (Transcript 4, Paragraph 30).

Complexity. One of the biggest challenges is the inherent complexity of HDT technology. It involves technical complexities that can be difficult to understand, and the availability of approved software is limited (Transcript 2, Paragraph 16). Even individuals well-versed in technology within companies might struggle to explain the inner workings of AI which is used by HDT technology (Transcript 1, Paragraph 14). This complexity extends beyond technology and is reflected also as a legal challenge. AI systems are so intricate that legal experts find it challenging to fully understand them, leading to concerns about data protection (Transcript 1, Paragraph 12).

Another aspect integral to this multifaceted landscape is the interplay between startup enterprises and legal consciousness. Startups are known for their innovation, but they often overlook legal considerations. This aspiration for innovation might expose them to legal risks that could arise later when the HDT product is deployed (Transcript 1, Paragraph 32). Thus, it's crucial to introduce legal awareness early in the development process. Furthermore, legal implications also extend to the concept of explainability. The level of explanation provided by an AI system influences the choice of methods to comply with legal requirements (Transcript 3, Paragraph 13).

Bureaucracy. Another challenge for the implementation of HDT technology consists in obtaining certification for medical devices. The regulatory landscape, limited notified bodies, and complex evaluation processes pose significant hurdles, particularly for startups (Transcript 2, Paragraph 15 17). These challenges are magnified for software-based products due to their unique considerations and limited approved options (Transcript 2, Paragraph 15 16).

Startups, constrained by resources and experience, find it difficult to navigate regulatory expectations, leading to uncertainty and compliance challenges (Transcript 2, Paragraph 17). Financially, compliance can

be costly, with conformity assessment fees ranging from 25,000 to 40,000 euros, coupled with management costs (Transcript 2, Paragraph 21).

The international landscape brings further disparities, as the EU's Medical Device Regulation contrasts with regions like the USA, where software-based medical products face less stringent regulation (Transcript 2, Paragraph 21). Slow regulatory changes add another layer of complexity. Norms and common specifications undergo a time-consuming evaluation process, contributing to delays in certification, which can pose a serious threat, especially in fast-developing environments such as HDT (Transcript 2, Paragraph 23).

Data. Challenges in data handling emerge from a lack of familiarity with complex data privacy regulation, with peculiar considerations like religious affiliations (Transcript 4, Paragraph 18).

Changes in data privacy regulation, notably the General Data Protection Regulation, impact data utilization practices, raising the complexity of obtaining patient consent for research (Transcript 2, Paragraph 31). Balancing the need for data with legal constraints, the use of pseudonymized and anonymized data, and varying interpretations of legitimate interest form ongoing challenges (Transcript 1, Paragraphs 10, 14, 16).

Variations in data privacy enforcement across different countries are problematic, as they could be leading to disparities in applying regulations (Transcript 1, Paragraphs 10, 14). This is an issue because projects regarding the implementation of HDT may span around different countries and involve different professions like doctors and engineers (Transcript 7, Paragraph 6). The evolution of data privacy regulation raises the question of their relevance and competitiveness on a global scale (Transcript 1, Paragraph 14).

Data needs of AI contrast with principles like data minimization, complicating the privacy landscape (Transcript 1, Paragraph 14). The distinction between pseudonymized and anonymized data presents unique complexities in legal and privacy considerations (Transcript 1, Paragraphs 16). The challenge of informed consent arises from the complexity of AI systems, making transparency difficult (Transcript 1, Paragraph 16). Legal concerns extend to research, where even anonymized data can face restrictions due to reidentification risks (Transcript 3, Paragraph 21). Resolving these issues might be crucial for the implementation of HDT since many HDT systems rely on AI (Transcript 5, Paragraph 40).

Service Provider. The challenge of maintaining data protection while utilizing cloud service providers in AI applications is discussed as well. Strategies like "Data Protection Impact Assessments", for example selecting trustworthy providers, aid compliance (Transcript 1, Paragraph 28). However, the wide presence of providers like Amazon Web Services poses challenges due to potential US authority access, testing the compatibility of cloud services with European data protection (Transcript 1, Paragraph 28). This dynamic highlights the tension between legal requirements and practical considerations.

Regulation. In addition, concerns about overregulation stifling progress are mentioned (Transcript 6, Paragraphs 11, 17). The impacts of regulation, including the "Medical Device Regulation", are highlighted by the interview partners, discussing challenges such as software classification and innovation shifts to the USA (Transcript 2, Paragraphs 14, 19). Privacy by Design principles are explained, emphasizing documentation and proactive assessment (Transcript 1, Paragraphs 16, 18).

The influence of the "General Data Protection Regulation" and its practical implementation is explored, acknowledging challenges in reconciling ambition with reality (Transcript 1, Paragraphs 14, 20). The German legislative approach is discussed as challenging to execute practically (Transcript 1, Paragraphs 22, 25). The DiGA approval process is also examined, illustrating the complexities of proving efficacy while navigating regulatory requirements (Transcript 3, Paragraphs 21, 25). DiGA refers to a health care refund system for digital heath products in Germany, in which expenses for certified digital health products are refunded by insurances.

Ethical Aspects

Ethical considerations in our study primarily concentrate on the immediate influence of HDT on patients and healthcare professionals. Our emphasis revolves around identifying benefits and dangers of the implementation.

"Instead of putting it in a mouse, we put it in a virtual human. […] as I said, with the automotive industry, they do eventually test on a real car, but they do the vast majority of it safely on the virtual car. Why wouldn't we want to do it safely on the virtual human?" (Transcript 7, Paragraph 19)

Benefits

Pre-Testing. Instead of utilizing live animal testing, a virtual human model is employed as a safer and more efficient alternative. This approach aligns with practices in the automotive industry, where extensive testing is first conducted on virtual vehicles before any real-world experiments (Transcript 7, Paragraph 19). The implementation of HDT might facilitate drug discovery and testing processes, enabling assessments of pharmaceuticals and medical devices like artificial valves and pacemakers within virtual environments (Transcript 7, Paragraph 19).

Furthermore, HDT technology can be helpful in the context of surgical procedures. In many instances, physicians working with neonatal patients face challenges due to limited precedents and tight time frames. The interviewee explains that critical decisions have to be made within a mere 10-minute window after initiating surgery on an infant. The choices made during this brief period can profoundly impact the child's future. The interviewee further mentions that a children´s hospital in the USA has already established virtual twins for patients as a standard practice, affording surgeons the opportunity to explore a plethora of potential techniques on the virtual twin prior to the actual procedure. This preparatory work ensures that when the critical 10-minute timeframe starts, physicians are well-equipped to make the most informed decisions, having completed their comprehensive groundwork. As a result, HDT is able to offer both practitioners and patients a more promising outlook regarding various treatments (Transcript 7, Paragraph 19).

Explainability. Patient involvement and understanding play a significant role in healthcare. Visualizations can assist patients in grasping complex information, potentially alleviating concerns and fears. Visual aids can simplify medical concepts thus, making them more accessible to individuals (Transcript 5, Paragraph 36).

In some cases is crucial to ensure that patients comprehend the information provided to them. Currently, there is a gap in achieving this goal, especially concerning complex medical documents like doctor's notes. Patients often struggle to understand the content, leaving them feeling isolated and not clear about their own health or consequences of certain treatments. Addressing this issue is vital as informed patients are better equipped to make decisions about their health and treatment options (Transcript 4, Paragraph 49).

Furthermore, the interviewee explains that patients find the HDT concept easier to comprehend than other technical tools. This understanding likely fosters trust and willingness among patients to share their health data. Many patients seem to be open to sharing their information when they understand its purpose and implications. This underscores the importance of medical education and improved communication in bridging the gap between patients and healthcare technologies (Transcript 7, Paragraph 28).

Dangers

Data Misuse. Consent to use patient data for research purposes has been given by many patients, but questions arise when considering the transition from research to commercialization, especially in an AI context. Legal experts need to evaluate this shift. Some companies, with the assistance of legal departments, may find ways to initially label a de facto commercialization as research. Balancing the potential benefits and risks is recognized as a significant challenge, as the widespread use of patient data could lead to misuse of technology (Transcript 6, Paragraph 15).

The importance of comprehensive risk assessment is emphasized by another interviewee, particularly in the context of product development. This includes considering possible use errors, manageable misuse of the product and other factors that might not be immediately apparent to developers or usability engineers. It is suggested that a multidisciplinary approach is crucial for risk management, with input from various perspectives. The core principle is that no isolated development should occur, and multiple people should review and consider potential outcomes (Transcript 2, Paragraph 27).

It is highlighted that despite concerns, many people tend to handle their data casually, particularly when they perceive benefits, until an unfortunate incident occurs. The scenario of personal health data being accessed by unintended parties, potentially affecting one's job prospects, is mentioned (Transcript 2, Paragraph 33).

Human Substitute. One interviewee raised the question of whether it's critical for patients to have unrestricted access to all medical information and options, suggesting that while patients should not be kept in the dark about their conditions, there's value in physicians sorting through various diagnostic and therapeutic approaches before presenting them to the patient. The goal is to provide the patient with the best possible options rather than overwhelming them with every potential choice from the start (Transcript 5, Paragraph 36).

Furthermore, emphasized was to initially restrict access to medical professionals only. Given the proliferation of health apps for patients, it might be prudent to assess how patients handle their health data before introducing more interactive HDT. A collaboration with treating physicians rather than granting patients unrestricted access is recommended (Transcript 5, Paragraph 40). Also in the context of diagnostic support software designed primarily for medical professionals, concerns about predictable misuse arise. A potential issue arises where highly efficient software might lead to diagnosing a significantly higher number of patients per day. This could disrupt established healthcare roles, like medical laboratory technicians or medical assistants, and raise concerns about the misuse of the technology for cost-cutting or efficiency rather than enhancing patient care (Transcript 2, Paragraph 27).

Miscellaneous. Regardless of the proficiency and thorough testing of HDT technology and the underlying algorithmy, it remains impossible to explore the billions of potential scenarios within a highly complex space. A remote chance of parameter combinations leading to a severe misclassification always remains, possibly resulting in a dreadful scenario (Transcript 3, Paragraph 27).

Besides, the need for clear intentions and definitions of the usage of HDT is stressed (Transcript 5, Paragraph 40).

Discussion

Our interview partners provided in-depth knowledge from different fields including software development, medical practice, finance, and law. The pool of interviewees included both researchers and people working in companies, from small startups to big corporations. This mix of perspectives helped us get a comprehensive picture of how HDT might evolve and how its different characteristics interact. However, there are several limitations to which this work is subject. The upcoming pages will encompass our principal findings, limitations, and a proposal for future research.

Principal Findings

We were able to extract various interesting findings from the conducted interviews. One of those findings was that the legal and ethical issues connected to HDT are closely tied to how the technology works and the changes it brings. For instance, the advancements enabled by deep learning within HDT models enhance the precision of diagnoses and treatments, extending their applicability across a broader spectrum. However, this also raises concerns about the regulations, guidelines, and ethics related to using large amounts of sensitive patient data to train these systems. The validation and extension of the concepts introduced in the deductive coding is provided in Table 3.

Considering the current landscape of development and research, the adoption of deep learning seems to be the predominant path toward the implementation of HDT. Despite deep learning systems being notably more advanced compared to existing artificial intelligence or statistical methods, it is noteworthy that several barriers recognized in hindering HDT applications appear to be intricately connected with this approach. Particularly within critical domains like healthcare, the absence of transparency and interpretability presents significant obstacles in the responsible creation and introduction of deep learning-based HDT applications into the market. To mitigate the concerns arising from deep learning, many of our interviewees discussed approaches to increase the explainability of developed systems. These approaches encompass a more balanced integration of artificial intelligence systems, with greater emphasis placed on data engineering and the refinement of relatively transparent analytics methodologies. In essence, although the application of deep learning and its "Black Box" approach can yield robust outcomes with reduced effort, capitalizing on the expanding pool of available data, ensuring transparency and interpretability becomes more vital, especially within the healthcare domain, to ensure alignment with ethical and legal considerations.

Concept/ Key	References	Validation/ Extension from Transcripts	# Interview Partners	Frequency
Personalization	Erol et al. (2020) Garg (2021) Haleem et al. (2023) Sahal et al. (2022) Sun et al. (2023)	HDT enables monitoring of health parameters on patient-level, thus thresholds for certain diagnoses can be tailored patient-specific. HDT is able to generate information about the specific nature of a heart of a given patient. This was tested in a trial project in cooperation with the U.S. Food and Drug Administration	3	5
Prior Assessment	Haleem et al. (2023) Okegbile et al. (2022)	HDT (virtual surgery) can help surgeons to assess nature of the heart of unusual patient types e.g., newborn babies prior to surgery. Hence, surgeons have a greater understanding of the specific heart of the patient when performing the surgery. This has become common practice at a hospital in the United States.	2	2
Regulations	Bagaria et al. (2020) Yuan Li (2019)	Regulations pose a significant barrier towards the clinical implementation of HDT. Companies in Europe experience more regulatoric pressure than companies in North America.	5	18
Explainability	Huang et al. (2022)	Deep learning algorithms are state of the art for processing big data, such as in HDT. However, more explainable algorithms like support vector machines allow for transparency, which might be crucial in critical infrastructure like healthcare. These more explainable algorithms have previously often been neglected in favor of machine learing.	2	3

Table 3. Deductive Coding

Valuable insights into the current state of HDT development and application arose from interviewing experts from the above-mentioned relevant fields who were specifically engaged with start-ups in the field. This particular group proved to be a rich source of input due to startups' propensity to operate at the cutting edge of technological advancements (Dorner et al., 2017). Within the EU, startups focused on harnessing computational technology breakthroughs confront their central challenge within the regulatory framework. Although the need for proper governance was recognized in order to protect patient interests and personal information, the current perception of excessive regulation ("over-regulation") places companies within the EU at a significant competitive disadvantage vis-`a-vis their counterparts in Asia and the USA.

A number of interviewees acknowledged that practitioners are increasingly viewing the current regulatory framework within the EU as impractical. Apart from the novelty of HDT technology, many interviewees recognized the complexity of HDT systems as one of the principal reasons hindering proper and faster regulatory response from governmental units. This is primarily due to the necessity of a more profound comprehension of relevant applications by legislators to effectively evaluate the associated risks. As HDT systems transition from abstract concepts to less comprehensible entities through the application of

artificial intelligence, software developers are encountering increasing difficulties in articulating this complexity to policymakers and legal experts, introducing more tension to the field.

The main driving force behind the advancement of HDT is the aspiration to enhance diagnostic and treatment outcomes for patients. HDT applications have been primarily designed to serve as advanced tools, providing valuable support to medical professionals and healthcare personnel. Discussed use cases include more sophisticated pre-testing, treatment, and surgery preparation, improved patient communication and transparency, and enhanced diagnostic accuracy through the use of CDSS.

It was distinctly emphasized that the envisioned implementation of HDT applications is explicitly and unequivocally intended to aid medical practitioners, rather than replace or substitute them. This approach is found to be even more important when ethical and regulatory factors are taken into account. As applications in the medical field constitute critical infrastructure, both the ethical and the legal responsibility of the final decision lies with the human expert, while the tools at his disposal only serve to support or enhance the decision-making process. This distinction also extends to the way different medical products or tools are regulated, concerning performance requirements and quality management. Given the heightened capabilities of such technologies, an essential concern raised by several interviewees pertained to the potential misuse of these applications if they were to supplant medical experts. Such an imprudent use of HDT could lead to patients conducting self-diagnoses or attempting to self-manage medical conditions, or clinics moving to the reduction of medical staff due to an expected higher patient turnover through the use of HDT applications. These outcomes bring forth substantial ethical and practical considerations regarding the public's access to HDT technology and the way the capabilities of developed applications are evaluated and communicated to relevant stakeholders. In summary, this work has contributed to the research domain of HDT by directing attention towards the context within the European Union. Through a synthesis of interdisciplinary expertise and insights drawn from practitioners and researchers alike, this study facilitates a comprehensive understanding of the subject. Additionally, our approach has enabled the validation of particular concepts frequently encountered in existing literature, achieved through the application of deductive coding techniques. Our results state that HDT can significantly improve quality of diagnosis and treatments. On the other side, regulations pose a significant barrier towards HDT implementation. For ensuring explainability and transparency, HDT technology has to rely on white box methods such as support vector machines, instead of relying on black box methods such as deep learning. Our findings from deductive coding are presented in the following table.

Limitations

While expert interviews, as a qualitative research method, offer valuable insights into participants' perspectives and experiences, it is essential to acknowledge their inherent limitations. Firstly, the findings derived from these interviews are specific to the individuals involved and cannot be generalized to a larger population. This constraint is primarily due to the modest sample size and the non-random selection of participants, which consequently restrict the broader applicability of the results. Moreover, participants' statements are subjective, and similarly, the coding and processing of the discussion results are also influenced by subjectivity. As a result, the hierarchies and groupings of topics presented in the analysis should be considered as suggestive rather than definitive. Another perspective or approach might lead to different interpretations and categorizations of the data. Thus, it is important to acknowledge that qualitative research inherently involves some level of interpretation and subjectivity, and researchers should exercise caution in drawing conclusions based solely on one perspective. Furthermore, due to the limited number of experts interviewed, it is not feasible to claim data saturation, as further interviews may provide new insights and perspectives, especially with regard to the ethical implications of HDT. In addition, HDT's research field is heavily influenced by regulatory and technical advancements, therefore the identified drivers or barriers are not conclusive and may change over time, e.g., as new regulatory frameworks are established or new technologies emerge.

Future Research

Firstly, conducting additional expert interviews and involving more stakeholders would be a valuable endeavor to enhance the quantity and quality of the findings. By including voices that were not represented in the current expert interviews, a broader and more comprehensive understanding of the subject can be achieved. Especially with regard to the ethical aspects of HDT, further insights could be helpful for deeper

understanding, as it was difficult to get a diversity of insights in this area with the experts interviewed in the present study. It also seems useful to regularly reassess and examine the current state of development in order to be able to take into account new aspects emerging as a result of omnipresent advancements in technology and the regulatory landscape. Furthermore, it might be worthwhile to specifically investigate the development of HDT in the US and Asian markets, as the significantly different legal frameworks in these countries could give rise to completely different challenges and opportunities compared to those in the EU.

Regarding the future development of HDT systems, a spherical approach and consideration of this technology were found to be essential when assessing the potential implementation and practicality of such applications. This lies in the high degree of interconnection between the examined fields, which was analyzed in the discussion section. Particularly when considering marketable applications of this technology, the regulatory implications of any decisions made regarding the technology during the development process need to be considered. It therefore follows that future research in the technical potential of HDT should partly reflect on the regulatory landscape associated with the implemented technologies. This should allow for realistic research streams with feasible practical applications.

Conclusion

HDT hold great potentials e.g. in terms of better and more personalized diagnoses and treatments, however, due to the numerous intertwined technical, legal and ethical considerations, practical application is still in its early stages. Thus, our study comprehensively evaluates the practical implementation of HDT by examining technical, legal, and ethical dimensions. We conducted interviews with experts across medical informatics, law, technology leadership, and startup investment, revealing diverse insights into HDT's potential trajectory and broader implications. At its core, the purpose of HDT seems to be to aid medical professionals, rather than supplanting them. According to experts, HDT hold the potential to deliver enhanced diagnostic precision, improved patient integration, and heightened accuracy. This projection from professionals resonates with the theoretical underpinnings of HDT's capabilities. A noteworthy discovery lies in the intimate connection between technological advancements of HDT, such as the integration of deep learning, and the ethical and legal quandaries they introduce. While deep learning significantly enhances diagnostic precision, it simultaneously triggers apprehensions regarding data privacy regulations, ethical implications, and the conscientious handling of patient information. Thus, experts advocate for strategies to enhance transparency, involving the integration of AI with transparent analytics and an emphasis on robust data engineering. Furthermore, the discussions with experts have shed light on the regulatory environment, especially within the EU. Their perception of regulations being excessively constraining seems to be putting startups within the EU in a less advantageous position compared to their counterparts in other regions. Interviewees underscored the general complexity of regulations, largely stemming from the intricate nature of HDT systems. Transitioning from rather abstract ideas to technology-driven applications necessitates regulators to apprehend the intricacies of the technologies, presenting distinct challenges. Yet, this understanding appears to be important in order to create regulatory frameworks that protect patients' rights while at the same time not impeding innovation. In conclusion, our research contributes to a broader understanding of HDT, encompassing implications beyond specific contexts. By integrating interdisciplinary insights, we offer a holistic view on the topic of HDT in clinical practice. Our findings emphasize harmonizing technological progress with ethical, legal, and practical considerations. This appears to be a critical requirement for the widespread adoption of medical HDT applications, but also an important factor in effective policy making and certification of HDT components.

References

Agbo, C. C., Mahmoud, Q. H., & Eklund, J. M. (2019). Blockchain technology in healthcare: Asystematic review. *Healthcare*, *7*(2), 56.

Alam, K. M., & El Saddik, A. (2017). C2ps: A digital twin architecture reference model for thecloud-based cyber-physical systems. *IEEE Access*, *5*, 2050–2062. **https://doi.org/10.1109/ACCESS.2017.2657006**

Babylon Health, U. (2023). Smarter care, better health. **https://www.babylonhealth.com/en-us** Bagaria, N., Laamarti, F., Badawi, H. F., Albraikan, A., Martinez Velazquez, R. A., & El Saddik,

A. (2020). Health 4.0: Digital twins for health and well-being. *Connected Health in Smart Cities*, 143–152.

Barricelli, B. R., Casiraghi, E., Gliozzo, J., Petrini, A., & Valtolina, S. (2020). Human digitaltwin for fitness management. *IEEE Access*, *8*, 26637–26664. **https://doi.org/10.1109/ACCESS.2020.2971576**

Bogner, A., Littig, B., & Menz, W. (2009). *Interviewing experts*. Springer.

Boschert, S., Heinrich, C., & Rosen, R. (2018). Next generation digital twin. *Proc. tmce*, *2018*,7–11.

Bruynseels, K., Santoni de Sio, F., & Van den Hoven, J. (2018). Digital twins in health care: Ethical implications of an emerging engineering paradigm. *Frontiers in genetics*, *9*, 31.

Caputo, F., Greco, A., Fera, M., & Macchiaroli, R. (2019). Digital twins to enhance the inte- gration of ergonomics in the workplace design. *International Journal of Industrial Er- gonomics*, *71*, 20–31. **https://doi.org/10.1016/j.ergon.2019.02.001**

Chakshu, N. K., Carson, J., Sazonov, I., & Nithiarasu, P. (2019a). A semi-active human digitaltwin model for detecting severity of carotid stenoses from head vibration—a coupled com-putational mechanics and computer vision method. *International journal for numerical methods in biomedical engineering*, *35* (5), e3180.

Chakshu, N. K., Carson, J., Sazonov, I., & Nithiarasu, P. (2019b). A semi-active human digitaltwin model for detecting severity of carotid stenoses from head vibration—a coupled computational mechanics and computer vision method [e3180 cnm.3180]. *International Journal for Numerical Methods in Biomedical Engineering*, *35* (5), e3180. **https://doi.org/https://doi.org/10.1002/cnm.3180**

Chen, J., Yi, C., Okegbile, S. D., Cai, J., et al. (2023). Networking technologies for enabling human digital twin in personalized healthcare applications: A comprehensive survey. *arXiv preprint arXiv:2301.03930*.

Chen, Y. (2017). Integrated and intelligent manufacturing: Perspectives and enablers. *Engineer-ing*, *3* (5), 588–595. **https://doi.org/10.1016/J.ENG.2017.04.009**

Constantinescu, C., Rus, R., Rusu, C.-A., & Popescu, D. (2019). Digital twins of exoskeleton-centered workplaces: Challenges and development methodology [25th International Con- ference on Production Research Manufacturing Innovation: Cyber Physical Manufactur-ing August 9-14, 2019 | Chicago, Illinois (USA)]. *Procedia Manufacturing*, *39*, 58–65. **https://doi.org/10.1016/j.promfg.2020.01.228**

Corral-Acero, J., Margara, F., Marciniak, M., Rodero, C., Loncaric, F., Feng, Y., Gilbert, A., Fernandes, J. F., Bukhari, H. A., Wajdan, A., Martinez, M. V., Santos, M. S., Shamo-hammdi, M., Luo, H., Westphal, P., Leeson, P., DiAchille, P., Gurev, V., Mayr, M.,…

Lamata, P. (2020). The 'Digital Twin' to enable the vision of precision cardiology.*European Heart Journal*, *41* (48), 4556–4564. **https://doi.org/10.1093/eurheartj/ehaa159**

Cristancho, S. (2016). Lessons on resilience: Learning to manage complexity. *Perspectives on medical education*, *5*, 133–135.

de Sio, F. S., Faber, N. S., Savulescu, J., & Vincent, N. A. (2016). 27Why Less Praise for En- hanced Performance?: Moving Beyond Responsibility-Shifting, Authenticity, and Cheat-ing Toward a Nature-of-Activities Approach. In *Cognitive Enhancement: Ethical and Policy Implications in International Perspectives*. Oxford University Press. **https://doi.org/10.1093/acprof:oso/9780199396818.003.0003**

Dorner, M., Fryges, H., & Schopen, K. (2017). Wages in high-tech start-ups – do academic spin-offs pay a wage premium? *Research Policy*, *46* (1), 1–18. **https://doi.org/10.1016/j.respol.2016.09.002**

Drummond, D., & Coulet, A. (2022). Technical, ethical, legal, and societal challenges with digitaltwin systems for the management of chronic diseases in children and young people. *J Med Internet Res*, *24*(10), e39698. **https://doi.org/10.2196/39698**

Erol, T., Mendi, A. F., & Doğan, D. (2020). The digital twin revolution in healthcare. *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–7.

Ferrara, P., & Albano, L. (2020). Covid-19 and healthcare systems: What should we do next? *Public Health*, *185*, 1.

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, andRecommendations. *Minds and Machines*, *28* (4), 689–707. **https:// doi . org/ 10 . 1007 / s11023-018-9482-5**

Garg, H. (2021). Digital twin technology: Revolutionaryto improve personalized healthcare: **https://doi.org/10.52152/spr/2021.105**. *Science Progress and Research (SPR)*, *1* (1), 32–34.

Ghazal, T. M., Hasan, M. K., Alshurideh, M. T., Alzoubi, H. M., Ahmad, M., Akbar, S. S., Al Kurdi, B., & Akour, I. A. (2021). Iot for smart cities: Machine learning approachesin smart healthcare—a review. *Future Internet*, *13*(8), 218.

Giubilini, A., & Sanyal, S. (2015). The ethics of human enhancement. *Philosophy Compass*,*10*(4), 233–243. **https://doi.org/10.1111/phc3.12208**

Greco, A., Caterino, M., Fera, M., & Gerbino, S. (2020). Digital twin for monitoring ergonomics during manufacturing production. *Applied Sciences*, *10* (21). **https://doi.org/10.3390/ app10217758**

Grieves, M. (2015). Digital twin: Manufacturing excellence through virtual factory replication, michael w. *GRIEVES, LLC, Cocoa Beach, Florida, USA*.

Haleem, A., Javaid, M., Singh, R. P., & Suman, R. (2023). Exploring the revolution in healthcare systems through the applications of digital twin technology. *Biomedical Technology*, *4*, 28–38.

Harris, B. (2020). How 'digital twins' are harnessing iot to advance precision medicine'. *Health- care IT News*.

Hassani, H., Huang, X., & MacFeely, S. (2022). Impactful digital twin in the healthcare revolu-tion. *Big Data and Cognitive Computing*, *6*(3), 83.

He, R., Chen, G., Dong, C., Sun, S., & Shen, X. (2019). Data-driven digital twin technology foroptimized control in process systems. *ISA Transactions*, *95*, 221–234. **https://doi.org/10.1016/j.isatra.2019.05.011**

Hirschvogel, M., Jagschies, L., Maier, A., Wildhirt, S. M., & Gee, M. W. (2019). An in silico twinfor epicardial augmentation of the failing heart [e3233 cnm.3233]. *International Journal for Numerical Methods in Biomedical Engineering*, *35* (10), e3233. **https://doi.org/https://doi.org/10.1002/cnm.3233**

Hsieh, H.-F., & Shannon, S. (2005). Three approaches to qualitative content analysis. *Qualitativehealth research*, *15*, 7–8**. https://doi.org/10.1177/1049732305276687**

Huang, P.-h., Kim, K.-h., & Schermer, M. (2022). Ethical issues of digital twins for personalized health care service: Preliminary mapping study. *Journal of Medical Internet Research*,*24*(1), e33081.

Iyengar, K., Mabrouk, A., Jain, V. K., Venkatesan, A., & Vaishya, R. (2020). Learning oppor- tunities from covid-19 and future effects on health care system. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(5), 943–946.

Jaensch, F., Csiszar, A., Scheifele, C., & Verl, A. (2018). Digital twins of manufacturing systems as a base for machine learning. *2018 25th International conference on mechatronics andmachine vision in practice (M2VIP)*, 1–6.

Karlsson, M., Iversen, T., & Øien, H. (2018). Aging and health care costs.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (n.d.). Digital twin in manufactur-ing: A categorical literature review and classification. ifac-papersonline. 51, 1016–1022 (2018).

Lewis, M., Alexander, T., & Blais, W. H. C. (2019). A reference architecture for human behaviour representations.

Lin, Y., Chen, L., Ali, A., Nugent, C., Ian, C., Li, R., Gao, D., Wang, H., Wang, Y., & Ning, H.(2022). Human digital twin: A survey. *arXiv preprint arXiv:2212.05937*.

Liu, Y., Zhang, L., Yang, Y., Zhou, L., Ren, L., Wang, F., Liu, R., Pang, Z., & Deen, M. J.(2019a). A novel cloud-based framework for the elderly healthcare services using digitaltwin. *IEEE Access*, *7*, 49088–49101. **https://doi.org/10.1109/ACCESS.2019.2909828**

Liu, Y., Zhang, L., Yang, Y., Zhou, L., Ren, L., Wang, F., Liu, R., Pang, Z., & Deen, M. J. (2019b). A novel cloud-based framework for the elderly healthcare services using digitaltwin. *IEEE access*, *7*, 49088–49101.

Liu, Z., Meyendorf, N., & Mrad, N. (2018). The role of data fusion in predictive maintenanceusing digital twin. *AIP conference proceedings*, *1949*(1), 020023.

Lutze, R. (2020). Digital twin based software design in ehealth - a new development approach for health / medical software products. *2020 IEEE International Conference on Engi- neering, Technology and Innovation (ICE/ITMC)*, 1–9. **https://doi.org/10.1109/ICE/ITMC49519.2020.9198546**

Ma, X., Tao, F., Zhang, M., Wang, T., & Zuo, Y. (2019). Digital twin enhanced human-machineinteraction in product lifecycle [11th CIRP Conference on Industrial Product-Service Systems]. *Procedia CIRP*, *83*, 789–793. **https://doi.org/https://doi.org/10.1016/j.procir.2019.04.330**

Martin, G., Martin, P., Hankin, C., Darzi, A., & Kinross, J. (2017). Cybersecurity and healthcare:How safe are we? *Bmj*, *358*.

Mendi, A. F., Erol, T., & Doğan, D. (2021). Digital twin in the military field. ***IEEE Internet Computing***, *26*(5), 33–40.

Miller, M. E., & Spatz, E. (2022). A unified view of a human digital twin. *Human-IntelligentSystems Integration*, *4*(1-2), 23–33.

Myers, M. D. (2013). *Qualitative research in business & management*. SAGE.

Negri, E., Fumagalli, L., & Macchi, M. (2017). A review of the roles of digital twin in cps-based production systems. *Procedia manufacturing*, *11*, 939–948.

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

Okegbile, S. D., Cai, J., Yi, C., & Niyato, D. (2022). Human digital twin for personalized health- care: Vision, architecture and future directions. *IEEE Network*.

Onan Demirel, H., Irshad, L., Ahmed, S., & Tumer, I. Y. (2021). Digital Twin-Driven Human-Centered Design Frameworks for Meeting Sustainability Objectives [031012]. *Journal of Computing and Information Science in Engineering*, *21* (3). **https://doi.org/10.1115/1. 4050684**

OpenAI, L. (2023). Whisper.

Pidgeon, N., & Rogers-Hayden, T. (2007). Opening up nanotechnology dialogue with the publics: Risk communication or 'upstream engagement'? *Health, Risk & Society*, *9* (2), 191–210. **https://doi.org/10.1080/13698570701306906**

Pool, A. (2021). *Digital twins in rail freight-the foundations of a future innovation* (Master's thesis). University of Twente.

Popa, E. O., van Hilten, M., Oosterkamp, E., & Bogaardt, M.-J. (2021). The use of digital twins in healthcare: Socio-ethical benefits and socio-ethical risks. *Life sciences, society and policy*, *17*(1), 1–25.

Prainsack, B. (2017). Personalized medicine. In *Personalized medicine*. New York University Press.

Rich, E., Lewis, S., Lupton, D., Miah, A., & Piwek, L. (2020). Digital health generation? youngpeople's use of "healthy lifestyle" technologies. *The Digital Health generation: the im- pact of 'healthy lifestyle'technologies on young people's learning, identities and health practices-funded by Wellcome Trust*.

Rosen, R., Von Wichert, G., Lo, G., & Bettenhausen, K. D. (2015). About the importance of autonomy and digital twins for the future of manufacturing. *Ifac-Papersonline*, *48* (3), 567–572.

Roy, C. M., Bollman, E. B., Carson, L. M., Northrop, A. J., Jackson, E. F., & Moresky, R. T. (2021). Assessing the indirect effects of covid-19 on healthcare delivery, utilization andhealth outcomes: A scoping review. *European Journal of Public Health*, *31* (3), 634–640.

Ruckenstein, M., & Schüll, N. D. (2017). The datafication of health. *Annual review of anthro- pology*, *46*, 261–278.

Sahal, R., Alsamhi, S. H., & Brown, K. N. (2022). Personal digital twin: A close look intothe present and a step towards the future of personalised healthcare industry. *Sensors*,*22* (15), 5918.

Sanders, H. (2017). Garbage in, garbage out: How purportedly great ml models can be screwedup by bad data. **https://api.semanticscholar.org/CorpusID:52830669**

Schluse, M., Priggemeyer, M., Atorf, L., & Rossmann, J. (2018). Experimentable digital twins—streamlining simulation-based systems engineering for industry 4.0. *IEEE Transactions on industrial informatics*, *14* (4), 1722–1731.

Sharotry, A., Jimenez, J., Wierschem, D., Méndez Mediavilla, F., Koldenhoven, R., Valles, D., Koutitas, G., & Aslan, S. (2020). A digital twin framework for real-time analysis and feedback of repetitive work in the manual material handling industry, 2637–2648. **https://doi.org/10.1109/WSC48552.2020.9384043**

Sharotry, A., Jimenez, J. A., Wierschem, D., Mediavilla, F. A. M., Koldenhoven, R. M., Valles, D., Koutitas, G., & Aslan, S. (2020). A digital twin framework for real-time analysis and feedback of repetitive work in the manual material handling industry. *2020 Winter Simulation Conference (WSC)*, 2637–2648.

Shengli, W. (2021). Is human digital twin possible? *Computer Methods and Programs in Biomedicine Update*, *1*, 100014.

Siemens Healthineers, P. /. S. (2023). Taking a look at digital twin technology: A new frontierin personalised healthcare. **https://ai.myesr.org/articles/taking-a-look-at-digital-twin-technology-a-new-frontier-in-personalised-healthcare/**

Skjott 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**

Stark, R., Fresemann, C., & Lindow, K. (2019). Development and operation of digital twins fortechnical systems and services. *CIRP Annals*, *68* (1), 129–132. **https://doi.org/https://doi.org/10.1016/j.cirp.2019.04.024**

Strauss, A., & Corbin, J. (1998). Basics of qualitative research techniques, 101–106.

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications, Inc.

Sun, T., He, X., & Li, Z. (2023). Digital twin in healthcare: Recent updates and challenges. *Digital Health*, *9*, 20552076221149651. **doi:10.1177/20552076221149651**

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, *3* (1), 17. **https://doi.org/10.1038/s41746-020-0221-y**

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Mar-kets*, *31* (2), 447–464. **https://doi.org/10.1007/s12525-020-00441-4**

United Nations, D. o. E., & Social Affairs, P. D. (2019). World population ageing 2019 highlights.https :// www . un . org / en / development/ desa / population / publications / pdf / ageing / WorldPopulationAgeing2019-Highlights.pdf

van den hoven, J., Lokhorst, g.-j., & Poel, I. (2011). Engineering and the problem of moral overload. *Science and engineering ethics*, *18*, 143–55. **https://doi.org/10.1007/s11948-011-9277-z**

van der Valk, H., Hunker, J., Rabe, M., & Otto, B. (2020). Digital twins in simulative applica-tions: A taxonomy. *2020 Winter Simulation Conference (WSC)*, 2695–2706.

Wilsdon, J., & Willis, R. (2004). *See-through science: Why public engagement needs to move upstream*. **https://doi.org/10.13140/RG.2.1.3844.3681**

World Health Organization, W. (2021). Global health expenditure database. **https://apps.who.int/nha/database/ViewData/Indicators/en**

World Health Organization, W. (2022a). Ageing and health. **https://www.who.int/news-room/fact-sheets/detail/ageing-and-health**

World Health Organization, W. (2022b). Global expenditure on health: Rising to the pandemic'schallenges. **https://apps.who.int/iris/rest/bitstreams/1484345/retrieve**

Yeganeh, H. (2019). An analysis of emerging trends and transformations in global healthcare. *International Journal of Health Governance*.

Yuan, B., & Li, J. (2019). The policy effect of the general data protection regulation (gdpr) on the digital public health sector in the european union: An empirical investigation. *International journal of environmental research and public health*, *16* (6), 1070.

Privacy-Related Behaviour Change When Using Smart Home Technologies in Different Social Contexts

Critical Information Infrastructures, Winter Term 23/24

Elena Fantino
Master Student
Karlsruhe Institute of Technology
uyqel@student.kit.edu

Alwin Faßbender
Master Student
Karlsruhe Institute of Technology
ufudh@student.kit.edu

Linda Günder
Master Student
Karlsruhe Institute of Technology
uxuqj@student.kit.edu

Shanice Steinecke
Master Student
Karlsruhe Institute of Technology
unucp@student.kit.edu

Abstract

Background: *The concept of smart homes has evolved with advancements in the field of the Internet of Things (IoT) and ubiquitous computing. While the technology is now installed in many living spaces, privacy issues posed by smart home devices have not been fully explored in recent years.*

Objective: *This research paper aims to analyse user behaviour in response to privacy implications in smart homes.*

Methods: *Qualitative, semi-structured interviews are conducted to collect data for our research. Thematic analysis was used to identify and analyse patterns within the data.*

Results: *It is discovered that individuals are conscious of privacy risks in their personal spaces and adjust their behaviour to maintain a sense of security. In contrast, concerns in public settings are less significant. Additionally, the social environment in which participants find themselves contributes to different privacy behaviours especially regarding conversation topics and privacy protection measures.*

Conclusion: *Our findings highlight that it is important for developers and policymakers to consider diverse privacy expectations in different social contexts. We suggest a move towards more transparent and user-centric approaches in the design and regulation of smart homes.*

Keywords: smart home, privacy behaviour

Introduction

"When we think of what happens between the walls of our homes, we think of it as a trusted, private place. In reality, we find that smart devices in our homes are piercing that veil of trust and privacy — in ways that allow nearly any company to learn what devices are in your home, to know when you are home, and learn where your home is." - David Choffnes, Executive Director of the Cybersecurity and Privacy Institute at Northeastern University. (2023)

In today's interconnected world, smart home technologies have become widespread, providing unparalleled convenience and control over our living environments. However, this rapid technological evolution also raises concerns, particularly regarding privacy and security. A report by the Internet Society revealed that almost 63% of consumers are concerned about their smart devices being hacked, which could lead to potential data breaches (Internet Society, 2019). Furthermore, Kaspersky's comprehensive study found that in 2020 alone, smart home devices experienced over 1.5 billion attacks, highlighting the tangible threats to users' personal data (Kaspersky, 2020).

Therefore, it is important to balance the benefits of innovation with the need for privacy protection. The handling of sensitive data by these systems raises concerns, especially regarding potential data breaches and unauthorized access. Additionally, smart homes are considered to be Critical Information Infrastructures (CIIs) as they are not just passive repositories of information but active accumulators and disseminators of data. Thus, they fulfil the characteristics and functions of CIIs. Their interconnected nature and reliance on digital communication channels make them vulnerable to various security threats. Safeguarding privacy within these systems is challenging due to their critical status. Any breach or misuse of data can have far-reaching consequences. The dual nature of smart homes, as facilitators of modern convenience and potential breaches of privacy, sets the stage for a complex discussion on the balance between technological advancement and user privacy.

This research paper aims to analyse user behaviour in response to privacy implications posed by smart homes. It seeks to explore the strategies employed by individuals to navigate the privacy-convenience trade-off and how users adjust their interactions with smart home devices amidst growing privacy concerns in different social contexts. According to APA Dictionary of Psychology (2024), social contexts defines “the specific circumstance or general environment that serves as a social framework for individual or interpersonal behaviour. This context frequently influences, at least to some degree, the actions and feelings that occur within it.” (APA Dictionary of Psychology, 2024). The investigation of this paper centres around the following research question.

RQ: *“How do people change their behaviour when using smart home technologies in different social contexts in terms of privacy?”*

The study utilizes a comprehensive research approach to investigate this question. It analyses privacy behaviour in different social contexts, studies behavioural adaptations, explores sharing patterns, and identifies drivers of behavioural change.

The study is structured as follows. Chapter Background offers background information, including a comprehensive overview of smart home technologies, their classification as Critical Information Infrastructures (CIIs), and the privacy challenges they pose. Chapter Methodology outlines the methodology for collecting and analysing data to understand user privacy behaviour and adjustments in smart home contexts. Chapter Results presents the study's empirical findings on privacy behaviour analysis, behavioural adjustments, and data sharing patterns among smart home users. Chapter Discussion provides a comprehensive discussion of the principal findings, implications for stakeholders, potential directions for future research, limitations of the study, and a synthesis of how users navigate the trade-offs between convenience and privacy in smart home technology usage. This coherent structure ensures a thorough examination of the relationship between smart home technologies and privacy. It aims to make a significant contribution to the discussion on privacy and user-centred technology design.

Background

Privacy

Privacy has been researched and called a fundamental right as early as 1890 (Warren & Brandeis, 1890). An early definition of privacy was "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others." (Westin, 1967, p. 7). The definition was well-placed in a time in which most privacy invasions were of physical nature (Cate, 2000), but in the digital age, actions once considered private or confined to a small circle now generate data trails that reveal our preferences, characteristics, beliefs, and intentions (Acquisti et al., 2015). The theory on privacy therefore had to be expanded to incorporate the new challenges offered by the inter-connectedness of the digital age. For example, Westins work has expanded to attend to privacy on the

internet (Kumaraguru & Cranor, 2005). More modern theories argue that privacy expectations and norms vary by context, suggesting that what constitutes an invasion of privacy in one setting may be considered acceptable in another. This notion was relevant to this study in the context of smart homes, where the boundaries between public and private spaces are constantly negotiated by technology, and where data collection devices are installed directly in a person's home.

Smart Homes

The concept of smart homes has evolved with advancements on the Internet of Things (IoT) and ubiquitous computing. Ubiquitous computing, introduced by "The Computer for the 21st Century" (Weiser, 1999), is a vision that has significantly influenced the development of smart home technologies. Weiser's work (1999) anticipates a world where computers are seamlessly integrated into everyday objects and environments, making technology invisible yet omnipresent, a core principle in smart home design.

Smart homes refer to living environments that are equipped with interconnected devices and systems that provide automated control over various aspects of the home environment, such as lighting, temperature, and entertainment (Chan et al., 2009). These systems are designed to enhance convenience, efficiency, medical care, and security (Chan et al., 2009; Marikyan et al., 2019).

Privacy in the Context of Smart Homes

The matter of privacy has been identified as a major deficiency of smart homes (Arabo et al., 2012; Bugeja et al., 2016). Previous research has shown that data gathered from smart home devices can be used to identify and track the behaviour of the inhabitants (Apthorpe et al., 2017; Molina-Markham et al., 2010). Similarly, other technical and design challenges of smart homes must be met (Arabo et al., 2012; Bugeja et al., 2016). Previous research has also unveiled privacy concerns as an adoption barrier for smart homes, with people that are more concerned for their privacy are less likely to own smart home devices (McCreary et al., 2016) or to use them (Guhr et al., 2020).

While numerous technical solutions have been proposed to address these issues (Alami et al.; Chakravorty et al., 2013; Panwar et al., 2019; Yao et al., 2019) some argue that understanding user behaviours is essential for creating privacy mechanisms that are truly usable (Jacobsson & Davidsson, 2015). This motivates further sociological research to uncover user interactions with smart home devices.

Privacy behaviours in general have been shown to be dependent on the context (Acquisti et al., 2015), and privacy awareness in smart homes is dependent on the context as well (McCreary et al., 2016). Potential social issues arising from multi-user smart homes have been uncovered by previous research (Zeng et al., 2018). The research explored in this work includes those social issues and the dependency of privacy behaviours on context by investigating the social contexts of privacy behaviours in smart homes.

Zheng et al. (2018) conducted interviews with smart home device owners, exploring their views on privacy risks associated with smart homes. They analysed the measures people take to safeguard their privacy against entities outside the home that develop, oversee, monitor, or govern IoT devices and their data. They uncovered that a users' acceptance of data collection by external entities is based on the perceived benefits received, suggesting a transactional view of privacy where advantages can mitigate privacy concerns. Furthermore, their research revealed a dichotomy in user trust: while users generally place trust in smart home device manufacturers to safeguard their privacy, they seldom take steps to verify the implementation of such protections. Additionally, their study highlighted a significant lack of awareness among users regarding the privacy risks posed by inference algorithms, especially those analysing data from non-audio/visual IoT devices. This finding confirms the results of prior research, which found lapses in users' understanding of IoT threat models, meaning that users partly were unaware of the privacy risks smart homes pose (Zeng et al., 2018).

Privacy Behaviours

To explore user privacy behaviours within the digital realm, it's essential to understand the strategies individuals employ to manage their personal information. The categorization of privacy behaviours by Dehling and Sunyaev (2023) offers a structured framework to examine how individuals interact with information processing systems. It includes the seven behaviour categories: Privacy practice assessment,

disclosure, concealment, information deletion, information flow management, multiparty privacy protection, and privacy violation response. Understanding these behaviours provides insight into the ways individuals handle their privacy in digital environments, and in this case, how privacy in smart home environments is navigated. This categorization therefore serves us as a foundation for analysing privacy behaviours.

Privacy behaviours are influenced by factors such as uncertainty, context dependence, and the malleability and influence of external factors (Acquisti et al., 2015). Particularly, context dependence plays a pivotal role in this study, as it is examined how privacy behaviours change across different social contexts. The actual privacy behaviours have also been shown to differ from the intentions held by people (Acquisti et al., 2015; Norberg et al., 2007). Furthermore, in a survey of 569 individuals it was found that the intended use of collected data is more relevant to a person's privacy behaviour than the sensitivity of the data itself (Martin & Nissenbaum, 2015). These themes of privacy behaviour are considered in the process of this research.

The "chilling effect", a form of self-censorship induced by the perception of surveillance, is a phenomenon arising from dataveillance, the extensive and often opaque collection and analysis of digital data, leading individuals to modify their communication behaviour to avoid perceived negative consequences (Büchi et al., 2022). Uncertainty about data collection practices further reinforces these chilling effects (Solove, 2007) and will also be discussed in the context of this study.

Methodology

Qualitative Interviews

Qualitative, semi-structured interviews to collect data for this research were conducted. Qualitative research seeks to analyse the significance of experiences, aiming to understand not only the actions people take but also the reasons behind these actions (Douglas et al., 2009). Semi-structured interviews combine a pre-determined set of open-ended questions with the flexibility to explore themes or responses in more detail. This approach allowed us to address different levels of knowledge and experience of diverse participants. It permits the customization of questions for each person, considering their distinct backgrounds and perspectives. With this approach the goal is to gather diverse data that helps to better understand the research topic.

Sample Characterization

It has been reached out to participants within our personal networks, aiming to include individuals of diverse genders, ages, and levels of digital literacy. It was particularly important for us to interview both smart home owners and those who do not use smart home technology. During the recruitment the word "privacy" was not mentioned to avoid bias for the first sections of our interview regarding the smart home technologies. The goal was to ensure a broad representation of the population, believing that the varied experiences and backgrounds would provide diverse perspectives on the study's subject matter. The socio-demographic distribution of the 20 interview partners can be seen in Table 1.

Procedure

The study procedure started with the recruitment of the participants as described above including contacting potential participants and scheduling interview appointments.

The second step included the development of an interview guideline, outlining the main topics and questions to be addressed. This ensured consistency across all interviews while allowing for the exploration of individual experiences and perceptions. Prior to starting the individual interviews, participants were assured of the anonymity and confidentiality of their responses.

The interview guideline was divided into four main sections. The first part served as an introduction, which collected background information of the participants as shown in Table 1. Following the introduction was a section exploring the smart home usage of each interviewee. Then, general privacy understanding, and behaviours of the participants are discussed. The next section focused on examining how privacy concerns might influence their behaviour within smart homes, including whether this behaviour changed in the

presence of others. For participants without smart home technology, specific scenarios were presented to understand their potential concerns and behavioural intentions in hypothetical smart home situations.

Gender	Female: 8 Male: 12
Age range	19 – 64
Digital literacy self-assessment (scale: 1 – 10)	Mean: 7.15 Median: 7 Min: 2 Max: 10
Smart home (device) owner	Yes: 12 No: 8

Table 1. Participants' Socio-Demographic Information

In a third step the semi-structured interviews were conducted one-to-one, partly virtual, and partly in person. The interviews were conducted in English as well as German. It's important to mention that the word "privacy" was not translated but always used in English to ensure the same understanding of the term to all participants. Depending on the interviewees knowledge and willingness to speak, each interview took about 30 minutes.

Analysis Method

To analyse the semi-structured interviews, we employed thematic analysis as the most suitable approach, outlined by (Braun & Clarke, 2006). "Thematic analysis is a method for identifying, analysing, and reporting patterns (themes) within data" (Braun & Clarke, 2006, p. 6) "A theme captures something important about the data in relation to the research question and represents some level of patterned response or meaning within the data set" (Braun & Clarke, 2006, p. 10). Themes can be identified in two ways: in an inductive or „bottom up" way, or in a theoretical or „top down" way. The inductive approach is data-driven as the identified themes are greatly linked to the data. This approach provides a rich description of the overall data. The theoretical approach is analyst-driven as researcher's search for themes that answer the research question. This approach provides a more detailed explanation of some features of the data (Braun & Clarke, 2006).

Braun and Clark (2006) describe a step-by-step guide to conduct thematic analysis. The first step is about familiarizing yourself with your data by reading through the data multiple times while taking notes. In a second step the ideas about your data should be turned into initial codes. A code points out a feature of the data that seems interesting. Next, those codes are sorted and organized into potential themes. After that the themes are reviewed and refined before they are defined and named.

This analysis was started by dividing the four interview topics among ourselves. Using an inductive approach, we familiarized ourselves with our respective interview section to identify initial codes. These initial codes were then organized into top-codes and sub-codes, allowing us to propose potential themes without yet considering the research question. In the next phase, we gathered for an analysis session to share and discuss our findings from each interview section. Here, we transitioned to a theoretical approach, collaboratively examining the codes to develop themes directly tied to the research question. This process enabled us to identify our main themes and connect the relevant codes to each theme. For our final step, we conducted a comprehensive data review to validate our main themes. This thorough verification process ensured the robustness and reliability of our findings.

Results

The main themes that emerged from the interviews are the following: First, the physical environment in which participants find themselves influences their privacy-related behaviour. Second, the social environment, consisting of other people being present when (hypothetically) using smart home devices, plays a major role. In Figure 1 the main findings in relation to these two themes are visually presented.

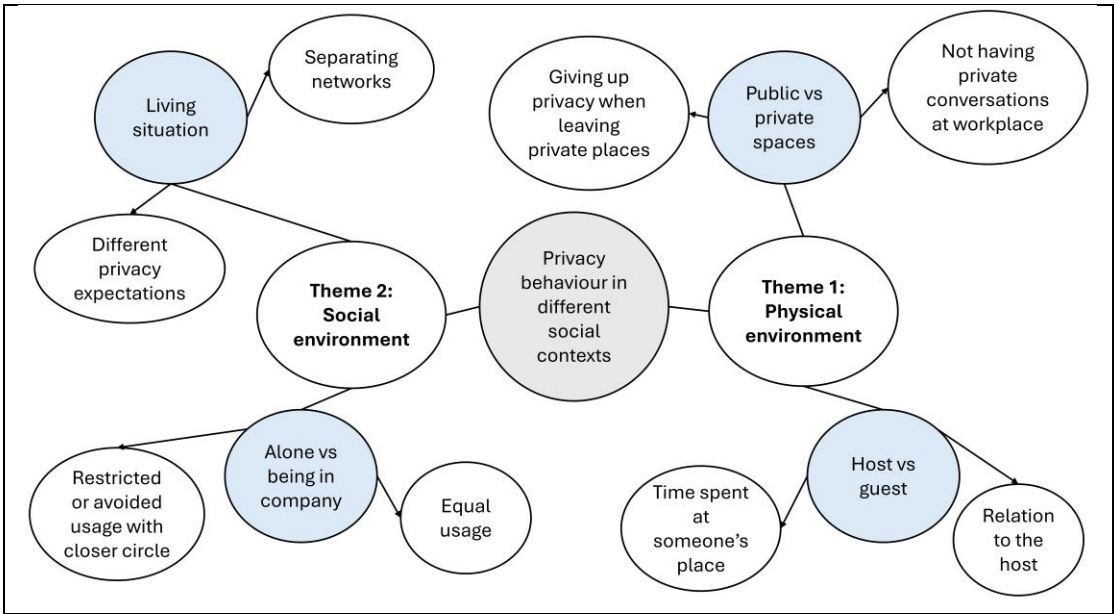

Figure 1. Identified Themes of Thematic Analysis

Theme 1: Physical Environments

One of the most prominent differences in participants' behaviour could be recognised in relation to their physical location when having interactions with smart home devices. Some interviewees in fact stated that they would not be comfortable in using smart devices at home but that they would not be bothered by having them at their workplace.

> *"[I'm protecting my privacy] not at work, more privately, and on the internet in general, so it wouldn't actually bother me having smart home devices in the office."*
>
> * * *
>
> *"Yeah, so I would say in general, my digital life. [Privacy] doesn't concern me too much when I'm working even though I'm also quite conscious about that. Like for instance, on my work laptop, if I would hand it in afterwards, it is quite easily possible to track down which websites I've been watching and which data was on my laptop […]."*

This was also justified as to be related to the fact that they tend to have no private conversations at work and are, therefore, not disclosing any information that would make them feel uncomfortable when knowing a voice-controlled device is potentially listening.

Another young person, not owning a smart home device, also claimed that he would behave differently when knowing there is such a device in a private place, but would not change his behaviour in public spaces, such as university, in case of smart devices' presence:

> *"I don't think I would behave any differently in public spaces. At university, you feel more under surveillance anyway. For example, when I talk to friends in the cafeteria, I can always assume that the people at the next table are, listening to me. At home, on the other hand, I feel safe."*

To the question why there was no behavioural difference in presence of smart devices in public spaces he answered that by leaving the private sphere to enter the public sphere he already assumes giving up his privacy.

Another curious behavioural contradiction was identified in the context of having people over at one's place. Besides some interviewees coherently claiming they would turn off possible active devices when being hosts and expecting this behaviour also when being guests at someone's place:

> *"Definitely [would I turn off active voice-controlled smart devices]. I would also be in favour of deactivating Alexa when I go somewhere else and there's an Alexa device there while I'm there. What everyone does personally at home is up to them. But as long as I'm there, I'd like it to be deactivated."*

Some other participants, in fact, stated that they would be grateful to be informed about the presence of smart home devices as guests at a friend's place but did at the same time realise that they do not inform their guests about active smart home devices when being hosts. One interviewee, who wishes to be informed at someone else's place, commented on his own devices:

> *"[...] I can see when someone has walked through the living room or the kitchen, and can therefore see, for example, that someone was out and about in the house at 4 o'clock in the morning. That's not relevant for me and I don't look at things like that, but basically there is a risk and the visitors and people living here aren't aware of it."*

Another one explained that he would feel uncomfortable if he realized that there are active smart devices at a friend's place he is not informed about as he does not know what data protection measures have been taken, but at the same time stated:

> *"Maybe I would inform my friends about it and ask if they are okay with it, but practically I don't think.... I mean in the scenario that I would have this Alexa I don't think I would ask everybody for his consent of the hypothetical recording of their voice, but yeah maybe that's the right thing to do."*

And added in relation to deactivating the device:

> *"I don't think so, I mean if I trusted enough to record me all the time I also... maybe if my friends see it and say "hey please turn it off" I would turn it off, of course."*

Another participant, self-assessing to highly value privacy, admitted that he would appreciate but not ask for switching off potential active devices, but that this is also because he never spends a lot of time at someone's place, thus, presuming not to share sensitive data:

> *"And since I usually don't assume that people have the best knowledge and, different perspectives on the entire topic, I am usually not mad about it, but I appreciate it when someone has the awareness and tells me about it. [...] But I don't usually go around and ask them if they do have a device. [...] But I also think that that's something to do with the fact that I'm not at their place for regular and super long times."*

At the same time, earlier in the interview this participant not owning smart devices mentioned that he would not feel comfortable conversating with friends in presence of a voice-controlled smart device out of fear that sensitive dialogues could potentially be recorded.

Another distinction was made by another participant regarding the relation someone has with the host:

> *"I think it would bother me if the person I'm visiting is in my close social circle, because I am likelier to talk about sensitive information with people I trust, so if close friends with someone I would expect them to reveal such information to me. If I do not know the host of the house that much I think it's ok as I will probably not be having privacy concerning conversations."*

This example is also closely related to the social environment and, thus, leads to the second theme.

Theme 2: Social Environment

The second theme which resulted in the analysis is the social environment. It includes sharing one's physical surroundings with people and interacting with them in different ways.

Some people in this case, for example, feel more uncomfortable when having conversations with friends and family rather than with colleagues or strangers in presence of voice-controlled smart home devices:

> *"So, I guess in the main thing for Alexa would be if I'm having a private conversation for example and to have these conversations which in theory are just something I would talk about with maybe a family member or a friend to have these uploaded somewhere in a way that would harm anyone or take an advantage of for example. I guess that would be a main concern in terms of privacy at least."*

At the same time, some participants do not see any difference in who is in their proximity when interacting with smart devices and try to get the best out of it in each specific situation, regardless of whether they are alone, with friends or with their family:

> *"I use it in pretty much every situation you can think of. When I'm by myself, it's super handy for things like setting reminders or controlling the lights without having to get up. When the family's around, we often use it to play music or check the weather before heading out. And when friends are over, it's great for setting the vibe, like dimming the lights and picking a playlist. It's like this little assistant that's always there, ready to help out, no matter who I'm with or what I'm doing."*

Another aspect influencing some behaviours is the living situation. As to this one participant explained that the reason for not owning smart home devices is the different expectations flatmates and friends might have regarding their own privacy:

> *"Well, the living situation is part of it, I would say. That's actually kind of a major part because I feel as though I would have to do a lot of changes to my network because I want to create a safe network and while I'm living with other people, I feel like that might be the biggest weakness to the network, other people that use the network."*
>
> * * *
>
> *"I would probably not expect that maybe if someone is in my network and has my Wi-Fi - they usually don't have malicious intent to do something bad and still they could have malware on it. So, I usually, divide the network into guests and my own private network. So, they are separated. So, it's always about, the intent of people."*

Furthermore, another interviewee, owning several smart home devices, also mentioned the possible weaknesses introduced by external visitors entering the network and plans to accordingly adapt his behaviour after having noticed how easy it sometimes is to access sensitive information such as a wi-fi password:

> *"Yes, I do have concerns. When I see how easy it is to infiltrate someone else's WLAN, I do worry. I have to say that very clearly. Some time ago, I noticed how someone was able to read my Wi-Fi password in plain text from their iPhone and give it to someone else. That was OK for me at the time because of the group of people involved. However, I haven't implemented this yet, but I will change my behaviour accordingly. I already have two Wi-Fi networks anyway, so I'm going to separate them completely. Only me and my partner will be in one and all guests will be excluded."*

On the other hand, some participants not owning voice-controlled smart home devices claimed they are worried about the possible data collected about them when entering a friend's place with an active Alexa:

> *"I mean, I could have privacy concerns if I had an Alexa at home but I'm not having that, I'm rather concerned about my privacy when I'm at other like a friend's place for instance. They are having their smart home devices on where data is gathered about me. And usually I do not have the power to decide or I do not feel comfortable telling a friend of mine that they should shut up their Alexa for instance, or even that they should shut up their phone because that's the one thing that is recording us basically all the time and I don't know if you're in the position to ask others to do that. So, this is a reason why I would say that my privacy might be limited, is limited, when others do not have that high privacy concerns. I'm having and they are therefore allowing much more data to be gathered about them, which also results in more data that is gathered about me."*

* * *

> *"[...] I would be bothered by a smart home device if it was active and I wasn't told about it. I think there's also an awareness about the fact that, the way I see privacy, I'm fairly strict about it, but not everyone feels the same way about it. [...] And so, I would say, I appreciate it if someone told me, but I don't expect it from them the same way."*

This last example is also closely related to the first theme, in which contradictory expectations and actions of guests and hosts were presented. However, it also shows how the interviewees adapt to certain situations in relation to their own (hypothetical) smart home and how they perceive being in someone else's smart environment.

Generally, the two presented themes are strongly interconnected, as social and physical environment often depend on each other. Nevertheless, a distinction must be made to better recognise concrete patterns and eventually distinguish between different influential factors.

Discussion

Principal Findings

The key findings of this study demonstrate how individuals adapt their privacy behaviours in response to smart home technologies, with a significant emphasis on the impact of different social contexts. The thematic analysis emphasized the significant influence of both physical and social environments, which intertwine and mutually impact each other.

From the interviews it can be derived that participants are very aware of their privacy within their personal space, often perceiving a high threat to their privacy in their homes when thinking of the usage of smart home devices. Participants expressed a strong desire for safety and privacy in their private spaces, where they are more likely to engage in personal and sensitive conversations. The presence of voice-controlled devices, particularly in hypothetical scenarios, often made interviewees uncomfortable. To mitigate these concerns, some participants stated to turn off devices or avoid certain topics of conversation to restore a sense of security.

In contrast, in public spaces such as workplaces or universities, participants generally did not feel obliged to change their behaviour when surrounded by smart devices, as they assume that privacy in such environments is not guaranteed by default and also do not expect it. While some were conscious of potential privacy issues in public places, their concerns were not as noticeable as those in their homes. Some participants were indifferent to the presence of smart home devices in their workplace, indicating that the existence of such devices does not significantly impact privacy perceptions in public environments. Therefore, it highlights a nuanced understanding and acceptance of privacy norms based on the location.

When visiting friends' or family members' homes, participants varied in their awareness of smart home devices. Some were certain of their presence, while others were not. Interestingly, participants appeared to have higher privacy concerns compared to public spaces on the basis that most of their private conversations are conducted in a private scenario. In addition, concerns were raised about the presence of voice-controlled devices in others' homes, although few usually take action or are mostly unsure how to address it. This hesitation may come from a reluctance to infringe upon others' privacy in their own homes, further strengthening the observation that privacy perceptions are influenced by the location.

Building on that, this research indicates a distinction in privacy expectations based on whether one is a host or a guest. Some participants expressed a desire to be informed about the presence of smart home devices when visiting others, mentioning discomfort without knowing which devices were active and questioning the homeowner's data protection measures. Interestingly, when smart home device owners were asked if they inform their guests about these devices, participants often do not. This reveals a significant contradiction, since participants value being informed as guests, but do not provide this information as hosts unless asked.

Within the social environment a difference between two groups of participants was recognised. Some stated to feel more uncomfortable having conversations with friends and family rather than with colleagues or strangers in the presence of voice-controlled smart home devices. At the same time, there were others who

felt equally comfortable in any company while using these devices, whether they were alone, with friends, family, or other visitors.

Participants' behaviours are also influenced by their living situations. Individuals in shared apartments, particularly students, mentioned refraining from owning smart home devices due to potential differing perspectives on privacy within the household and a desire to maintain a sense of safety.

The study suggests that the presence of smart devices in various social settings, such as family gatherings versus solitary environments, influences users' privacy management strategies, reflecting a restrained balance between privacy and convenience. The study revealed that some users tend to change their interactions with smart home devices depending on the presence of others, indicating a previously unexplored social aspect to privacy concerns. The results of this study add to the list of research indicating a context-dependence of privacy behaviours (Acquisti et al., 2015).

Moreover, the "chilling effect" on personal expression in environments with active smart devices was evident. There was a conscious or even unconscious way of adapting behaviour by limiting the expressed thoughts and facts out of fear of possible surveillance by smart home devices. The impact of this effect differed significantly depending on the social setting, indicating that users' privacy behaviours are contextually sensitive. Most participants explained that they tend to avoid certain conversational topics with friends or family members, mainly regarding personal, medical, or financial data, when aware of the presence of a voice-controlled smart device. A few interviewees also mentioned uncertainty in relation to when such a device is active, as well as what kind of data is collected and where it is stored.

Furthermore, trust in technology providers and perceived control over data are identified as critical aspects in shaping users' acceptance and use of smart home devices. The trade-off between information disclosure and the perceived benefits of a technology observed by previous research (Martin & Nissenbaum, 2015; Zheng et al., 2018), align with the participants' statements of how the advantages of smart homes can sometimes overshadow privacy considerations when trust in the smart system is given. The interviewees weigh privacy risks of smart homes against perceived benefits.

The trust in the smart home providers was an important factor for the interviewees as proposed by Zheng et al. (2018), but they did not go great lengths to confirm their beliefs in detail. Brand image and location of the smart home company seem to contribute the most to this trust level, but not many concrete steps were taken to ensure that the smart home companies of their choice handled privacy protection in the way they believed.

While previous studies have revealed that smart home users are unaware of the privacy risks smart homes pose (Zheng et al., 2018), we noticed that some interviewees were completely indifferent to these risks. This might indicate "privacy fatigue", a phenomenon resulting from the complexity of controlling the online data of oneself, which might contribute more to the privacy behavior of people than even privacy concerns do (Choi et al., 2018). On the other hand, many participants appeared to be very concerned about the potential privacy violations occurring by the great disclosure of data when using smart home devices. Aligned with already cited studies (Guhr et al., 2020; McCreary et al., 2016) this was one of the main reasons mentioned for not owning or using such a device and was described as an intentional privacy protection measure.

Implications for Practice

The study provides valuable insights into the aspects that influence privacy-related user behaviour in interactions with smart home devices. Therefore, technology developers, policymakers, and privacy advocates could facilitate the use of smart home devices by understanding these different privacy behaviours and, thus consequently, offering different protection possibilities for users' privacy.

Developers should prioritize security by implementing robust privacy controls, offering clear data management options (e.g., optional updates or consents), and enhancing transparency to build users' trust in smart home devices. Policymakers are encouraged to enact regulations that safeguard consumers' privacy, ensuring that smart home technologies comply with stringent data protection standards in different environmental settings. Additionally, there is a critical need for education initiatives where public education campaigns could use real-life scenarios to illustrate how smart devices collect and use data, enabling consumers to better understand their privacy rights and possible violations of them.

Limitations

This study was conducted through 20 interviews. However, the sample size and diversity of the participants may limit the scope of the conclusions. Additionally, the lack of quantitative validation restricts the findings to investigative insights rather than confirmatory statements. Therefore, the outcomes of this research may have limited generalizability. It is also possible that the biases of the interviewees affected their self-reports of behaviour. As has been shown by Moore (2017), privacy behaviour is different across cultures. The validity of these results is therefore limited to the cultural region of central Europe.

Future Research

As presented in the results, social contexts influence privacy perceptions in smart home settings. Therefore, we recommend further research in this direction. Longitudinal studies in communal environments, such as the Energy Smart Home Lab at KIT, can provide a better understanding of user behaviour over time. Furthermore, it is important to evaluate the balance between the benefits provided by smart devices and the need of protecting privacy. This requires a critical assessment of the surveillance capabilities inherent in these technologies. Regarding the limitations of the study, further research should examine the impact of cultural differences on privacy perceptions in the smart home context and expand the participant base to use quantitative measures for stronger validation.

Conclusion

In conclusion, the study contributes to existing research by highlighting the nuanced ways individuals adapt their privacy behaviours within the context of smart home technologies, particularly emphasizing the influence of social environments. It is found that participants are acutely aware of privacy threats within their personal spaces, adjusting behaviours to maintain a sense of security, whereas concerns in public settings are less pronounced. At the same time, privacy in relation to their closer circle of people is valued higher and is tried to be protected and preserved as much as possible by a more cautious behaviour. This underscores the need for developers and policymakers to consider varied privacy expectations across different social contexts, suggesting a shift towards more transparent and user-centric approaches in smart home design and regulation.

References

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514. **https://doi.org/10.1126/science.aaa1465**

Alami, A., Benhlima, L., & Bah, S (2015). An overview of privacy preserving techniques in smart home Wireless Sensor Networks. In *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)* (pp. 1–4). **https://doi.org/10.1109/SITA.2015.7358409**

APA Dictionary of Psychology. (2024, January 25). Social context. In *APA Dictionary of Psychology*. **https://dictionary.apa.org/social-context**

Apthorpe, N., Reisman, D., & Feamster, N. (2017, May 18). *A smart home is no castle: Privacy vulnerabilities of encrypted IoT traffic*. **http://arxiv.org/pdf/1705.06805v1**

Arabo, A., Brown, I., & El-Moussa, F. (2012). Privacy in the age of mobility and smart devices in smart homes. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (pp. 819-826). IEEE. **https://doi.org/10.1109/SocialCom-PASSAT.2012.108**

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. **https://doi.org/10.1191/1478088706qp063oa**

Büchi, M., Festic, N., & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society*, *9*(1), 205395172110653. **https://doi.org/10.1177/20539517211065368**

Bugeja, J., Jacobsson, A., & Davidsson, P. (2016). On privacy and security challenges in smart connected homes. In *2016 European Intelligence and Security Informatics Conference (EISIC)* (pp. 172-175). IEEE. **https://doi.org/10.1109/EISIC.2016.044**

Chakravorty, A., Wlodarczyk, T., & Rong, C. (2013). *Privacy preserving data analytics for smart homes*. In *2013 IEEE Security and Privacy Workshops* (pp. 23-27). IEEE. **https://doi.org/10.1109/SPW.2013.22**

Chan, M., Campo, E., Estève, D., & Fourniols, J.-Y. (2009). Smart homes - current features and future perspectives. *Maturitas*, *64*(2), 90–97. **https://doi.org/10.1016/j.maturitas.2009.07.014**

Choi, H., Park, J., & Jung, Y. (2018). The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior*, *81*, 42–51. **https://doi.org/10.1016/j.chb.2017.12.001**

Dehling, T., & Sunyaev, A. (2023). A design theory for transparency of information privacy practices. *Information Systems Research*, 0(0). **https://doi.org/10.1287/isre.2019.0239**

Douglas, H. A., Hamilton, R. J., & Grubs, R. E. (2009). The effect of BRCA gene testing on family relationships: A thematic analysis of qualitative interviews. *Journal of Genetic Counseling*, *18*(5), 418–435. **https://doi.org/10.1007/s10897-009-9232-1**

Guhr, N., Werth, O., Blacha, P. P. H., & Breitner, M. H. (2020). Privacy concerns in the smart home context. *SN Applied Sciences*, *2*, 247. **https://doi.org/10.1007/s42452-020-2025-8**

IMDEA Networks Institute (2023, October 26). New research reveals alarming privacy and security threats in smart homes. *Tech Xplore*. **https://techxplore.com/news/2023-10-reveals-alarming-privacy-threats-smart.html**

Internet Society. (2019, May 16). Concerns over privacy and security contribute to consumer distrust in connected devices. *Internet Society*. **https://www.internetsociety.org/news/press-releases/2019/concerns-over-privacy-and-security-contribute-to-consumer-distrust-in-connected-devices/**

Jacobsson, A., & Davidsson, P. (Eds.) (2015). Towards a model of privacy and security for smart homes. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)* (pp. 727-732). **doi:10.1109/WF-IoT.2015.7389144**

Kaspersky (2020). Kaspersky Security Bulletin 2020. Statistics. Kaspersky. **https://go.kaspersky.com/rs/802-IJN-240/images/KSB_statistics_2020_en.pdf**

Kumaraguru, P., & Cranor, L. F. (2005). Privacy indexes: A survey of Westin's studies. Carnegie Mellon University. **https://doi.org/10.1184/R1/6625406.V1**

Marikyan, D., Papagiannidis, S., & Alamanos, E. (2019). A systematic review of the smart home literature: A user perspective. Technological Forecasting and Social Change, 138, 139–154. **https://doi.org/10.1016/j.techfore.2018.08.015**

Martin, K. E., & Nissenbaum, H. (2015). Measuring privacy: Using context to expose confounding variables. Columbia Science & Technology Law Review, 18, 176-218. **http://dx.doi.org/10.2139/ssrn.2709584**

McCreary, F., Zafiroglu, A., & Patterson, H. (2016). The contextual complexity of privacy in smart homes and smart buildings. In F. H. Nah & C. H. Tan (Eds.), HCI in Business, Government, and Organizations: Information Systems. HCIBGO 2016 (Vol. 9752, Chapter 7). Springer. **https://doi.org/10.1007/978-3-319-39399-5_7**

Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E., & Irwin, D. (2010). Private memoirs of a smart meter. In Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building (BuildSys '10) (pp. 61–66). Association for Computing Machinery. **https://doi.org/10.1145/1878431.1878446**

Moore, B. (2017). Privacy: Studies in Social and Cultural History (3rd ed.). Routledge. **https://doi.org/10.4324/9781315172071**

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. Journal of Consumer Affairs, 41, 100–126. **https://doi.org/10.1111/j.1745-6606.2006.00070.x**

Panwar, N., Sharma, S., Mehrotra, S., Krzywiecki, Ł., & Venkatasubramanian, N. (2019, April 10). Smart Home Survey on Security and Privacy. **http://arxiv.org/pdf/1904.05476v2**

Solove, D. J. (2007). The First Amendment as Criminal Procedure. New York University Law Review, Article 82, 112. **https://scholarship.law.gwu.edu/faculty_publications/946**

Warren, S. D., & Brandeis, L. D. (1890). The Right to Privacy. Harvard Law Review, 4(5), 193–220. **https://doi.org/10.2307/1321160**

Weiser, M. (1999). The computer for the 21st century. SIGMOBILE Mobile Computing and Communications Review, 3(3), 3–11. **https://doi.org/10.1145/329124.329126**

Westin, A. F. (1967). Privacy and Freedom. Washington & Lee Law Review, 25, Article 166. Available at: **https://scholarlycommons.law.wlu.edu/wlulr/vol25/iss1/20**

Yao, Y., Basdeo, J. R., Kaushik, S., & Wang, Y. (2019). Defending my castle: A co-design study of privacy mechanisms for smart homes. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19) (Paper 198, pp. 1–12). Association for Computing Machinery. **https://doi.org/10.1145/3290605.3300428**

Zeng, E., Mare, S., & Roesner, F. (2017). End user security and privacy concerns with smart homes. In Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017) (pp. 65-80). USENIX Association. Retrieved from **https://www.usenix.org/conference/soups2017/technical-sessions/presentation/zeng**

Zheng, S., Apthorpe, N., Chetty, M., & Feamster, N. (2018). User perceptions of smart home IoT privacy. Proceedings of the ACM on Human-Computer Interaction, 2 (CSCW), Article 200. **https://doi.org/10.1145/3274469**

Large Language Models: Fragmented Market or the Winner Takes it All?

Trustworthy Emerging Technologies, Winter Term 23/24

David W. König
Master Student
Karlsruhe Institute of Technology
david.koenig@student.kit.edu

Julian Faber
Master Student
Karlsruhe Institute of Technology
julian.faber@student.kit.edu

Jingyi Xie
Master Student
Karlsruhe Institute of Technology
jingyi.xie@student.kit.edu

Daniel Loder
Master Student
Karlsruhe Institute of Technology
daniel.loder@student.kit.edu

Abstract

Background: *Since late 2022, Large Language Models (LLMs) from major players like OpenAI, Anthropic, Google, and Meta have advanced significantly, prompting reflection on AI's societal impact. Concerns about monopolistic trends in tech, exemplified by companies like Microsoft, Alphabet, and Meta underscore the need to scrutinize market dynamics. The integration of Generative AI (GenAI) into workflows raises questions about consolidation, with OpenAI's platformization efforts indicating a potential trend toward monopolization.*

Objective: *This research paper aims to analyze the market structure of the GenAI landscape, focusing on Large Language Model Foundation Providers (LLMFPs) and Large Language Model Layer Providers (LLMLPs). Despite recognizing risks, research on consolidation trends in the LLM market is limited, motivating this study to analyze current dynamics and assess potential monopolistic or oligopolistic outcomes.*

Methods: *The methodology employed in this research paper involves a qualitative approach utilizing interviews with industry experts. The final sample size consisted of eight interviewees who are active as investors, consultants or entrepreneurs in the field of GenAI. Data analysis revolved around a robust coding framework, with selective coding as the primary approach. Seven key concepts of market structure served as initial codes, and 156 text passages from the transcripts were assigned to these codes.*

Results: *The research findings suggest that the LLMFPs face high barriers to market entry due to resource-intensive requirements for training models and potential legal challenges. Established players, like OpenAI and Anthropic, dominate the market, with proprietary models likely to lead to further consolidation. While some argue that open-source models foster competition, the trend indicates a move towards consolidation with limited new entrants. Regulatory bodies, like the Federal Cartel Offices, could play a crucial role in shaping this trajectory. Conversely, LLMLPs have lower barriers to entry, facilitated by APIs from LLMFPs. However, they face the risk of obsolescence as LLMFPs integrate similar functionality or discontinue APIs. Despite the regulatory landscape, LLMLPs face less legal challenges as they mostly work and train their refined models with available customer data.*

Conclusion: *This research sheds light on the competitive dynamics of the GenAI market. LLMFPs face formidable entry barriers, leading to market concentration among a few dominant players, while LLMLPs benefit from lower barriers but must contend with risks of dependence on LLMFPs. Our findings emphasize the importance of regulatory caution to foster innovation and fair competition in this evolving landscape.*

Keywords: Large Language Model, LLM, monopoly, competitive landscape, market fragmentation

Introduction

Large Language Models (LLMs) have made strong advances since the end of 2022 (Douglas, 2023). The current main LLM Foundation Providers (LLMFPs) include OpenAI (GPT-4), Anthropic (Claude), Google (Gemini), Meta (Llama 2) (Kosma, 2023; Lutkevich, 2023; Meta, n.d.). In particular, the introduction of the OpenAI natural language processing (NLP) tool, ChatGPT, has penetrated the public consciousness (Morris, 2023). ChatGPT answers all types of written questions, maintains human-like conversations, and provides code in various programming languages. The freemium business model of OpenAI enables many people to integrate Generative Artificial Intelligence (GenAI) into their daily workflows, enabling them to personally experience its features through the free use of ChatGPT. GenAI made society reflect on the potential of the AI revolution, and the negative and positive implications that could follow (Morris, 2023).

However, concentrated markets and the formation of monopolies, particularly in the technology sector, have already been observed in the past (Birch & Bronson, 2022). Monopolistic platforms pose a risk because they can leverage their market dominance to manipulate prices or undermine competition and innovation (CMA, 2020). Examples of powerful technology companies include Microsoft (a major shareholder of OpenAI), Alphabet, and Meta, which have already been confronted with governmental actions to regulate their market power in the past (Kollnig & Li, 2023). At the same time, these firms play a pivotal role in the field of LLMs. For this reason, the cross-section of market dynamics and the advancement of GenAI raises several questions about future developments and implications.

This apprehension is relevant to the LLM market as well, where dominant LLM providers may accumulate power (Verdegem, 2024) by integrating more solutions into their platforms (Kamps, 2023), thereby potentially marginalizing smaller competitors (Kollnig & Li, 2023) to reinforce their market position. The first steps towards platformization can already be observed, as OpenAI recently presented its new GPTs, which enable the integration and creation of plugins (OpenAI, 2023). As mentioned above, this can lead to monopolization, which incorporates the ability to implement constraints on smaller entities and clients, jeopardizing future innovation and competition (Podszun, 2017). Furthermore, if only a few LLMFPs deploy the infrastructure, many companies could become highly dependent on them (Kamps, 2023), which further strengthens the market position of these few providers.

Although this potentially has a major impact on current and future market trends, it is not clear whether a few LLM providers will dominate the market or how this dynamic will influence competition. Moreover, the lack of clarity about the structure, interdependencies, and imbalances of this market and the potential emergence of market-dominating players adds urgency to this exploration. The example of ChatGPT already shows its capability to undermine and outperform established search engines, thereby disrupting this market (Cheng & Liu, 2023). However, even if risks and enabling factors of monopolies on the LLM market have been identified (Cheng & Liu, 2023) and potential market structures have been defined (Geertsema et al., 2023), scientific research on dynamics and consolidation trends in the LLM market has not yet been carried out. This development is becoming an increasingly urgent question during the ongoing AI revolution (Morris, 2023) and will have a major impact on society, politics, established companies, startups, and investors.

Therefore, the objective of our study is to analyze the current market dynamics within the LLM market, specifically assessing whether a trend toward market consolidation exists or has already occurred, potentially leading to monopolistic or oligopolistic dynamics. In doing so, we aim to answer the following research question in this seminar paper:

Research Question: *What is the impact of market dominance of main players in the field of GenAI, particularly in LLMs, on the competitive dynamics of the industry?*

To answer this, our first step is to analyze the market structure within the GenAI landscape, with a focus on LLMFPs and LLM Layer Providers (LLMLPs). Subsequently, we explore the criteria for a consolidated market. These criteria are then mapped to the dynamics of the LLM market to derive a statement about the risk of market consolidation. By conducting such a market analysis, we created a comprehensive framework of dynamics and consolidation trends for the LLM market. This framework reveals dependencies, assesses the potential for market displacement of smaller entities, and serves as a foundation for decision-makers in politics or economics and for future research in this area.

In the remainder of this seminar paper, we first elaborate on the theoretical foundations of the LLM market, followed by an explanation of our methodology. Subsequently, we present the results of our market research and interviews with domain experts. Our paper is concluded by our discussion section, in which we address the principal findings, implications, future research, and limitations of our contribution to the current body of knowledge.

Theoretical Background

Large Language Model

From a technical perspective, an LLM is a statistical illustration of a language, which indicates the probability of a specific sequence (such as a word or phrase) appearing in that language (Luitse & Denkena, 2021). LLMs are a subset of GenAI and artificial intelligence systems that focus on generating human-like text, responding to inquiries, or completing other language-related tasks (Floridi & Chiriatti, 2020; Kasneci et al., 2023; Sejnowski, 2023; Thirunavukarasu et al., 2023). Trained on extensive text data, they can perform new tasks based on natural language instructions (Scao et al., 2023). Some advanced LLMs are also multimodal, meaning that they can understand, generalize, and operate across different types of input, such as image, audio, or video next to a text input (OpenAI, n.d.; Pichai & Hassabis, 2024).

Currently, many of the advanced LLMs are adapted versions of the transformer model architecture proposed by Google researchers in 2017 (Radford et al., 2018; Scao et al., 2023; Vaswani et al., 2023). Based on the proposed transformer architecture, OpenAI developed in 2018 the first generative pre-trained transformer model, and therefore the name GPT (Radford et al., 2018). Since then, more advanced LLMs based on modified versions of the originally proposed transformer architecture were released, for example, OpenAI with GPT-4 (OpenAI, n.d.) or Google itself with Gemini 1.5 (Pichai & Hassabis, 2024). Presently, existing LLMFPs can be categorized into two primary types (Kosma, 2023). The first category consists of proprietary or closed-source LLMFPs, such as OpenAI with its GPT-4 model or Anthropic with Claude. In contrast, the second category comprises open-source LLMFPs, such as Meta's Llama 2 (Meta, n.d.), Google's Gemma open models (Banks & Warkentin, 2024; Schmid et al., 2024) and BLOOM (Scao et al., 2023).

In the LLM market, in addition to LLMFPs, many companies offer solutions powered by one or more LLMs of the LLMFPs. These types of companies are defined in this seminar paper as LLMLPs and are visualized in Figure 1. Well-known examples include OpenAI with ChatGPT (Plus) (OpenAI, n.d.), Microsoft, which powers its Bing search engine and is even planning to integrate LLMs into more offerings, such as Word, Duolingo with its new offering Duolingo Max, and Expedia, which is utilizing it for its new vacation planning AI assistance (Marr, 2023).

Market Dynamic Criteria

In economic landscapes, market structures play a pivotal role in shaping the behavior of firms and influencing predictions (Onozaki & Yanagita, 2003). Two key paradigms within this framework are monopolies and oligopolies. Monopoly arises when a single entity dominates the entire supply of a particular product or service, granting it exclusive control over pricing and output (Varian, 2010). When assessing the market for LLMs, it tends to be described as a natural monopoly reminiscent of software markets. Like software, the development of LLMs involves substantial costs, while their distribution to consumers is relatively straightforward (Narechania, 2021). Natural monopolies are based on economies of scale, where the advantages of scale, such as decreased average costs per customer, extend throughout the entire relevant market (Narechania, 2021). In such cases, the most efficient way to serve the market may be through a single monopolistic entity. In markets characterized by natural monopolies, optimal productive

efficiency is achieved by continuously adding new users to the existing provider, making the introduction of a competitor impractical. This remains true irrespective of the number of competitors initially present in the market. Natural monopoly markets may, at times, have multiple participants, resulting in an initially inefficient allocation of market inputs.

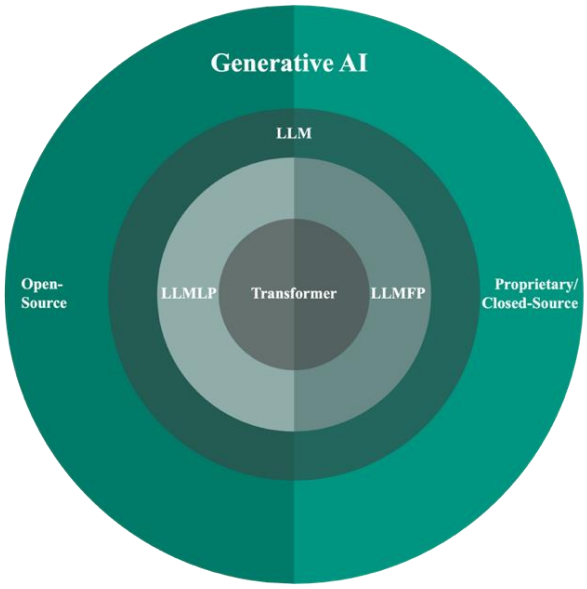

Figure 1. Simplified Visualization of GenAI, LLM, and Transformer

On the other hand, oligopolies involve several competitors on the market, but not so many that their interconnections are negligible (Friedman, 1982). This results in a market structure characterized by a small number of companies, each possessing the ability to considerably impact market dynamics. A possible structure for how the bias of dominant LLMs will influence the information space was proposed by Geertsema et al. (2023). There are three options: on the one hand, one model could completely dominate, which would lead to a monopolistic information space. Second, two opposing camps could emerge, or third, numerous models could form a fragmented information space (Geertsema et al., 2023). This structure of the potential information space depends on the dominance of the individual LLMs and can, therefore, be abstracted to market power, implying a potential market structure from a monopoly over two opposing camps to a fragmented market.

We have identified seven characteristics that are often used to predict market structures: market concentration (1), barriers to entry (2), potential for differentiation (3), network effects (4), interdependence (5), merger and acquisition (M&A) activities (6), and the regulatory environment (7).

Market concentration relies heavily on the number of companies providing the product and their respective industry output shares (market share) (Miller, 1967). In industries with a limited number of firms, monopoly pricing, and output can manifest itself either through a formal cartel agreement or implicit acknowledgment of "mutual interdependence". In contrast, in industries with a multitude of independent firms making individual decisions on pricing and output, competitive behavior can be observed. The likelihood of oligopolistic behavior in an industry increases with a higher concentration of output among the dominant firms. Furthermore, the market structure is influenced by the strategies employed by potential and existing competitors (Bain, 1956). Empirical research on potential market entry has primarily focused on the effects of **barriers to entry** that act as obstacles preventing new entrants from accessing the market. In the GenAI market, concerns raised by ChatGPT regarding antitrust issues could create barriers to entry. Inputting confidential information may lead to unfair competition, hindering small tech innovators (Cheng & Liu, 2023). Cheng & Liu (2023) formulate the apprehension that high entry barriers could result in a consolidation of the market by powerful technology companies and thus restrict small and medium-sized providers. Within the context of oligopolies and monopolies, the likelihood of a single entity dominating the market increases when positive and robust **network effects** are coupled with high multi-

homing costs and a lack of **differentiation opportunities** (Eisenmann et al., 2006). Introducing another dimension, Rysman (2009) suggests that providers of complementary goods can play a role in a winner-take-all outcome by offering differentiated products. This dynamic is particularly evident in the digital platform realm, where the overall strength of network effects often stems from the accumulation of data on a specific platform (Ruutu et al., 2017). Adding to these considerations is the crucial factor of **interdependence** between competitors. In instances where the level of interdependence between products is low, stability can be maintained (Kopel, 2009). **M&A activities** can considerably impact market structure through the formation of monopolies or oligopolies (Stigler, 1950). The consolidation of companies often results in increased market concentration, leading to reduced competition and potentially higher prices.

The last identified factor to assess the market structure is the **regulatory environment**, which is imposed by the state. Stringent regulations can prevent monopolies by fostering competition and facilitating market entry (Anton & Gertler, 2004). On the contrary, ambiguous regulations may favor the formation of oligopolies, as larger entities may face fewer constraints and can dominate more easily. For that reason, effective regulation is crucial to ensure healthy market competition and prevent the formation of monopolies or oligopolies.

Methodology

Literature Research

The literature review used a multifaceted approach to gather insights from a diverse range of sources. This review extends beyond traditional academic literature and incorporates gray literature, which includes news articles, market analyses, industry reports, and other nonpeer-reviewed materials. The purpose of this inclusive approach is to gain a comprehensive and nuanced understanding of the existing landscape and dynamics within the GenAI market.

To initiate the process, a search strategy was devised, combining both academic databases and specialized repositories for gray literature. This strategy involved querying databases such as PubMed, IEEE Xplore, and Google Scholar for journal articles, conference papers, and white papers related to GenAI. Simultaneously, searches were conducted on news websites, industry publications, and market analysis platforms to capture real-time developments, trends, and expert opinions. To examine the current funding landscape of LLMs, we used Crunchbase, a commercial database for information from the technology sector.

This comprehensive review formed the foundational knowledge base for the study, allowing the identification of key players, market structure, and relevant theoretical frameworks. Following the literature review, a framework was established to analyze the LLM market within the context of monopoly and oligopoly. Drawing on established market structures and relevant key factors, the framework aimed to conceptualize the interplay of dominant market forces, strategic interactions, and the impact on overall market behavior. This theoretical foundation served as a roadmap for the subsequent analysis of the GenAI market.

Expert Interviews

Research Design

To obtain insights directly from industry experts, a qualitative research approach was used by conducting interviews based on predefined guidelines. These guidelines were developed in alignment with the research questions and the theoretical framework, drawing upon the methods outlined by DeMarrais and Lapan (2004) to ensure that interviews produce information pertinent to the study objectives. By engaging with industry professionals, the objective is to capture nuanced perspectives, strategic insights, and industry-specific knowledge that may not be readily available through desk research alone. Interviews allow for the collection of qualitative data, which is essential for understanding the context, trends, and nuances within the technology and LLM market (Rosenthal, 2016). This type of information is often difficult to capture using quantitative methods alone, especially in young emerging markets.

Interview Partner Acquisition and Interview Conduction

To gather comprehensive information for this study, a multifaceted approach was employed to identify suitable interview partners. Initially, personal connections played a crucial role, taking advantage of existing networks to establish contacts with individuals who possess relevant expertise in the field. Specifically, the focus was on individuals actively involved in startups that specialize in LLMs. In addition, experts who serve as consultants or investors within the LLM sector were sought, to capture diverse perspectives and insights from professionals. Furthermore, outreach efforts were extended to key personnel within prominent companies. This method proved effective in ensuring a diverse range of perspectives from industry professionals. The recruitment strategy also included a proactive engagement with potential interviewees through professional platforms, such as LinkedIn. Cold outreach through this platform allowed the identification and connection with individuals who exhibited expertise in the subject. The final sample size consisted of a carefully selected group of eight interviewees consisting of companies and governments, consultants, and entrepreneurs in the field of GenAI to investors (see Table 1). The semi-structured interview format was chosen to provide a framework that allowed flexibility in exploring diverse perspectives while ensuring that key themes were consistently addressed across all interviews (Myers & Newman, 2007). Within the scope of this, we have prepared an interview guideline (see Appendix A).

Inter-view	Age	Gender	Business segment	Job Title	Perspective	Work Experience
1	54	m	Venture Capital (VC)	Managing Director	Investor	13 years
2	-	f	Consulting firm	Senior Associate	AI Consultant	3 years
3	-	m	LLM Startup	Founder and CEO	Entrepreneur and Computer Scientist	7 years
4	30	m	Consulting firm	Associate	AI Consultant and Computer Scientist	4 years
5	30	m	Consulting firm	Senior Associate	AI Consultant	2 years
6	41	m	Corporate VC	Investment Director	Investor and Computer Scientist	2.5 years
7	27	m	LLM Startup	Founder and CEO	Entrepreneur and Computer Scientist	1.5 years
8	53	f	AI Advisory	Strategic Advisor	AI Expert	25 years

Table 1. Demographic Information on Interview Partners

Each interview was designed to last around 30 minutes and averaged 32.2 minutes. To preserve the accuracy and integrity of the data, all interviews were recorded during video calls. This method not only facilitated transcription and subsequent analysis but also eliminated a major source of interviewer bias, enriching the depth of the qualitative data collected (Bucher et al., 1956).

Data Analysis Procedures

To enable analysis, the post-interview phase was initiated through the transcription of the collected audio data. This crucial step involved the transformation of qualitative spoken content into a written format, preserving the richness of participant responses (Bailey, 2008). Recognizing the importance of precision and consistency in capturing participant responses, the transcription procedure followed established guidelines for a comprehensive representation of the interview content.

In this study, data analysis revolved around a robust coding framework, primarily employing *selective coding* as the cornerstone methodology (Hsieh & Shannon, 2005). In the process of selective coding, codes

were predetermined before the interviews were analyzed. The reuse and application of established concepts was used to validate or expend upon a theoretical framework (Abraham et al., 2013). The seven key concepts of market structure, that were introduced in the Market Dynamic Criteria chapter, served as initial codes of the analysis. A total of 156 text passages from the transcripts were assigned to the seven codes. For example, the text passage "In any case, expertise is also a barrier to entry for small and medium-sized companies. Microsoft, Google and co. have no problem with this. They have plenty of people." was assigned to the code "Barriers to Entry". Table 2 provides a detailed overview of the codes used and the assigned text passages.

Codes	**Number of Mapped Passages**
1. Market Concentration	31
2. Barriers to Entry	30
3. Differentiation	33
4. Interdependence	22
5. Network Effects	21
6. Recent M&A Activities	7
7. Regulatory Environment	12

Table 2. Overview of Codes and Mapped Text Passages

The goal was to extract meaningful insights from the conducted interviews with a diverse array of stakeholders, including LLM service providers, consultants, investors, and analysts. The foundation of our analytical approach was to apply the seven factors, that indicate an oligopolistic market, to check if they also apply to the GenAI market.

To enrich the analysis, a *constant comparison method* was used to examine and contrast data across various participant groups (Wiesche et al., 2017). Through the comparison of data and concepts in all interviews, distinctions, similarities, and characteristics were examined to ensure the reliability of the findings. Given the different professional backgrounds of the interviewees, this comparative analysis revealed potential divergences and convergences in their perceptions of the GenAI market. Constant comparison facilitated the identification of patterns and trends, which enabled a comprehensive understanding of how different stakeholders conceptualize the dynamics of the market structure (Strauss & Corbin, 2003).

Results

Overview of the LLMFP Market

The strong growth of the LLM market can be explained by technical advances in hardware such as GPUs and TPUs, more sophisticated AI algorithms, data availability (Grand View Reserach, 2024), and the accessibility of application programming interfaces (Markets and Markets, 2023). While hardware manufacturers and service providers also play a pivotal role in this market (IOT Analytics, 2023) we will not consider them further in this analysis, as we focus only on LLMFPs. For our research on their funding, we focus solely on companies founded since 2015. However, in our research, we excluded companies such as Google or Meta, as they are publicly traded on the stock market. Our market research indicates that OpenAI raised USD 11.3 billion in equity (Crunchbase, n.d.). Therefore, they represent more than 53% of the funds collected from companies in this segment, illustrated in Figure 2. They are followed by Anthropic with USD 7.6 billion, and the first company in this list with headquarters in Europe, Aleph Alpha, with USD 643 million raised. The remaining companies, Mistral AI (USD 530 million), Cohere (USD 435 million), Adept AI (USD 415 million) and Hugging Face (USD 395 million), account for around USD 1.8 billion or 8.3% of the segment's funding.

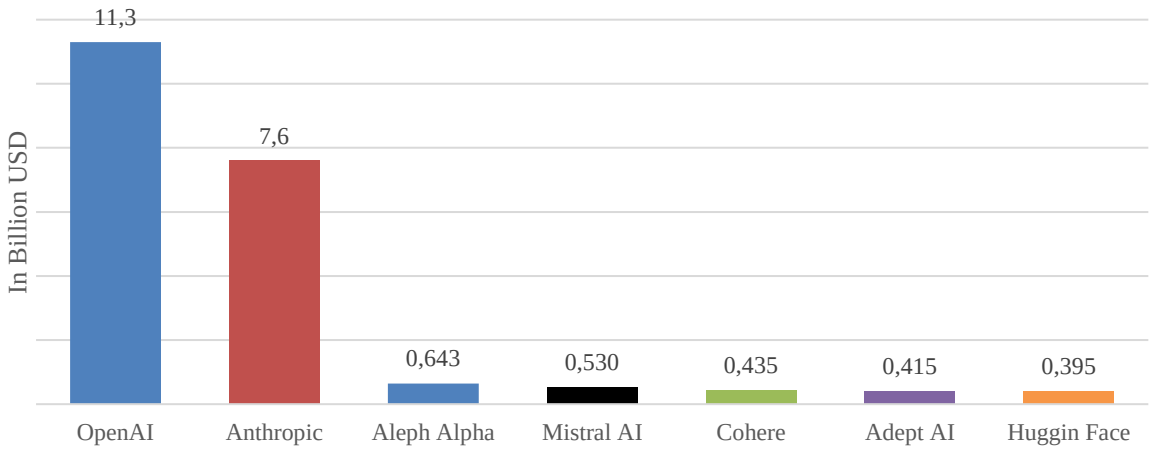

Figure 2. Funding in the Segment of LLMFPs Based on Data from Crunchbase (n.d.)

Comparable results emerge when examining the market shares of the GenAI foundation models and their platforms, which is also headed by OpenAI with a market share of 39% (IOT Analytics, 2023). OpenAI continuously improves its models (GPT-3.5, GPT-4, GPT-4 Turbo) and is often represented in the upper ranks of independent model assessments and rankings.

Microsoft is second on the list directly behind OpenAI with a total market share of 30% and is simultaneously a shareholder of OpenAI. Its Azure AI platform integrates an extended version of ChatGPT with additional functions and improved data protection and is used by over 20,000 paying customers. In addition to that, Microsoft offers its customers the usage of other LLMs, such as Llama 2, and integrates LLMs into various existing products and services such as Microsoft Azure or Office 365. AWS (with a market share of 8%) offers access to various models such as AI21 Labs, Anthropic, and Cohere, all of which have a market share of around 2% and extend their offer with development tools for customers. AWS was able to gain a large market share through its existing cloud customers and is characterized in general by independence and flexibility.

The fourth-largest market player, Google, with a market share of 7%, already had its cloud platform called Vertex AI, which covers machine learning and was extended in 2023 by the multimodal Gemini model. In addition, there are other providers such as Anthropic, AI21labs, Cohere, Aleph Alpha, Hugging Face, Alibaba Cloud, IBM, Baidu, and others, which add up to a total of 16% (IOT Analytics, 2023). The described market shares are summarized and presented in Figure B 1 in Appendix B.

The estimated market sizes of LLMs vary between different sources. The market analysis report from Valuates Reports and Grand View Research explicitly analyzes the LLM market and calculates market sizes of USD 12.74 billion (Valuates Reports, 2023) and USD 4.34 billion in 2023 (Grand View Reserach, 2024). The compound annual growth rates (CAGRs) are estimated to be 21.4% for the former and 35.9% for the latter. These input values lead to different forecasts. We illustrate the development of the market size in Figure 3 and derive the lower and upper bounds from the different estimations.

Furthermore, the terms LLMs and GenAI are often used synonymously, while different industry sectors and technology areas are included in the market size calculations. Overall, there are considerably more sources that focus on the GenAI market size.

However, the current market size calculations and CAGRs fluctuate strongly, resulting in fundamentally different forecasts. Table C 1, in the Appendix C, illustrates these strong fluctuations but also emphasizes that there is a consensus about strong market growth in the upcoming years. A future trend analysis outlines that the focus remains continuously on fine-tuning and integrating additional modalities such as images and videos into the models until 2025 (Markets and Markets, 2023). In the medium term (starting 2026), the market is expected to specialize further towards different industries and niches, while users are expected to control and personalize models increasingly. In the long term, towards the end of the decade, AI appears

to become increasingly autonomous, and ethical concerns are moving even more into the focus (Markets and Markets, 2023).

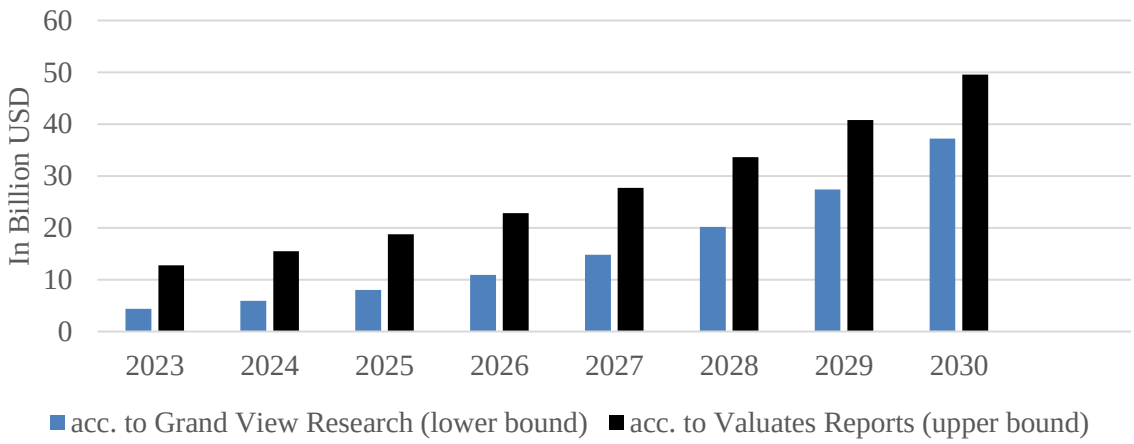

Figure 3. Development of the LLM Market Size with Upper and Lower Bounds Based on Valuates Reports (2023) and Grand View Research (2024)

Dynamics in LLM Markets

Market Concentration

Currently, there are only three to four dominant LLMFPs on the market [I1-5, I7]. However, this is unlikely to remain the case (in contrast, for example, to the social media market) [I1, I5, I8] due to influences from legal considerations and the democratization of model trainings [I5]. There also exists the assumption that ultimately no more than three foundation models will prevail, and no new models will emerge as there are no further hypotheses or new insights [I7]. This is supported by the fact that it is very challenging to build a new foundation model from scratch [I6]. Evidence of this assumption is OpenAI, which is already dominating the entire segment in terms of revenue. In contrast to only a few LLMFPs, there exist numerous LLMLPs [I1, I5], and there appears to be a trend toward even more personalized LLMs [I7, I8].

When discussing market concentration, the dominance of a few main players, such as OpenAI and Google, is perceived as a potential threat to other competitors [I2, I3]. The substantial market influence and financial capacity of these companies, especially in model development, could lead to even faster progress and strengthen their dominant position [I3]. However, current dominance fosters competition and motivates various companies to improve or maintain their current market share [I2, I4]. It is also important to look at open-source LLMFPs while observing market concentration. One of the biggest players is Meta, but there are also powerful models such as Mistral from Europe or Falcon from Saudi Arabia [I4]. Open-source models offer the opportunity for others to improve existing models and eventually build better ones. Historically, some foundation LLMs started as accessible-to-everyone and then transitioned to proprietary models due to monetary incentives [I8]. This also appears to be the case for OpenAI, and why Elon Musk has a lawsuit against them (Nidumolu et al., 2024; Novet, 2024). On the contrary, there is also the opinion that if the models were accessible to everyone, this would probably result in a denser market concentration [I8].

Barriers to Entry

Regarding barriers to entry, it is evident that challenges may arise in relation to data access, model training, and computational power required [I2, I4, I8]. Therefore, to overcome these high barriers and enter the market as a new LLMFP, a substantial amount of funding is necessary [I1, I2, I4, I7]. As the expertise of the foundation model is scarce, this presents a further challenge for smaller businesses [I4]. Not all data scientists have the necessary skills, since it is a niche problem in NLP. Larger companies like Microsoft and

Google have more resources to overcome this barrier. Engineering talent capable of developing foundation models is scarce and therefore represents a limiting factor. [I4, I6].
LLMFPs require talent with deep research expertise, while LLMLPs focus more on software engineering talent to build their product and implement the API connection [I7]. In addition to that, many companies had to completely rethink their products. Startups, on the other hand, can act quickly and offer a direct value add. Before 2022, deep research expertise was very valuable and is still demanded by LLMFP, but a reset took place, as programming skills and software engineering have become much more important due to the directly accessible APIs of LLMs.

However, on the other hand, existing LLMs are allowing more people with no engineering or math background to enter the LLM market [I8]. Consequently, they then increase the basic set of competitors. Despite that, new versions of OpenAI's models have already pushed some startups out of the market [I5, I6]. This particularly affected LLMLPs that have only built a very flat technology layer atop the foundation model. A further competitive advantage of larger companies is the available resources, which allows them to deal with legal disputes with other companies [I5, I6]. An example would be the lawsuit filed by The New York Times against OpenAI [I1, I5, I6].

In contrast to the other experts, an entrepreneur highlighted from personal experience that the current period is favorable for startups due to easy access to resources such as computing power, while it is still possible to obtain initial investment [I3]. However, for startups, there is a risk that innovative concepts may be adopted by larger corporations, complicating the competitive landscape [I2-I5, I7]. Therefore, the most substantial challenge lies in ensuring sustainable introductions in an ever-changing environment, where volatility, agility, and rapid pace of technological advancement make it difficult to anticipate future developments [I2, I5]. Despite the abundance of startups, their success rate is modest, with only approximately 5% achieving viability [I3].

Differentiation

Currently, LLMFPs seem to offer similar products [I1]. However, performance, which is often the main deciding factor for customers like LLMLPs, seems to be alternating between the major LLMFPs [I1, I3, I5, I7]. The performance of LLMFPs can theoretically be measured in intelligence per dollar [I7]. Furthermore, practical experience suggests that better-performing foundation models are often the cheaper ones, and that any increase in price is expected to bring a noticeable improvement in performance.

In addition to performance, it is crucial to adopt a distinctive approach in the establishment of new businesses, particularly in the context of the ongoing technological revolution [I3]. This "extra unique spin" is considered essential to gain a foothold in a market characterized by formidable competition from established players.

However, new ideas, such as local LLM deployment on end devices, could disrupt the market [I5]. As an example, Apple is working on those local LLMs which run on end devices without the need of an internet connection or a cloud infrastructure.

Furthermore, there is great potential for differentiation in niche markets rather than the introduction of foundation models or features [I2-6]. Models can be specialized for different industries, regions, or regulatory requirements, as specific use cases will continue to demand distinct products [I2, I8]. Small businesses should focus on a specific niche, build expertise in that area, and expand gradually, rather than trying to conquer the entire market at once [I2, I3]. The greatest challenge for companies, especially small ones, is to navigate various options, drawing parallels to the early Internet era [I4]. Similarly, just as the initial emergence of LLMFPs and LLMLPs transformed the offerings of existing businesses, companies must be prepared to adapt their offerings to further technological advances, as these will continue to emerge at a fast pace [I8].

Targeting specific verticals (e.g., supply chain) can be a great advantage, which can be further accelerated by integrating LLMs into existing software solutions for these niches to boost performance [I6]. LLM can be considered a tool for computer scientists to handle unstructured data very well.

Consequently, there are not many entirely new ideas; instead, there are various improvements that improve the functionality of existing concepts more effectively than before [I7]. In addition, many established

companies had to completely rethink their software products and user interfaces. In contrast, startups have the advantage of agility, which allows them to act quickly and deliver value rapidly.

Investor experience from portfolio companies indicates that there are already startups that have been able to win against established players, precisely due to a modern user interface [I6]. As the software iteration cycle accelerates, the lifetime of startups is getting shorter because they are being replaced more quickly [I7]. However, LLMLPs can differentiate themselves from big players such as OpenAI through efficient fine-tuning, better data collection, and attractive pricing. On the other hand, Microsoft has a diverse product portfolio and direct access to user interfaces in which it can integrate LLMs effectively [I2, I4, I6].

Interdependence

Startups, which are LLMLPs, should keep in mind that they can build enough intellectual property (IP) on top of the LLM so that the startups themselves are valuable [I1]. Even though there are only a few LLMFPs, it is likely that they will not try to cannibalize the startups, but rather be a wholesale that monetizes the API requests of the LLMLPs.

The interdependencies in model improvements are distinguished between closed-source (e.g., GPT 3.5) and open-source (for example, Llama 2) models [I3, I4]. Closed-source models create a 100% dependence on providers like OpenAI, which affects pricing and strategy. On the contrary, open-source models with good licensing offer independence. However, it is not clear whether these models will remain accessible to all or, rather, follow the monetary incentives [I8]. LLMFPs such as OpenAI not only benefit monetarily from their APIs, but also increase the opportunity costs for everyone to develop a new foundation model instead of using the existing one [I4, I6]. If many applications use the same API, then a standard can be enforced, increasing their market power [I6].

For users, it will often not matter which LLM powers the application [I7]. A possibility could also be to develop an interface layer to provide modularity, which enables the choice of which LLM the application should connect to. Some software developers are already using automated evaluation procedures that beforehand determine which LLM is best for a particular task and then fine-tune it [I7]. Furthermore, technology companies are also developing an interface layer as a kind of middleware to modularly connect to different LLMs with applications [I5, I6]. In the competitive landscape, giants like Amazon, Google, and Microsoft engage in intense pricing competition [I4, I5]. For example, Google's Gemini Ultra is highlighted for its cost-effectiveness compared to GPT-3.5 [I4]. The continuous drive to reduce prices, as seen with OpenAI's GPT-Turbo, aims to make advanced models more accessible to businesses [I4]. If a leading LLMFP reduces its prices, others must follow to remain competitive [I5]. This can potentially cause challenges for smaller players as they may be unable to adapt, which ultimately may lead to their business failures.

Although there are only a few major LLMFPs on the market, there are still no observable price fixing or coordinated behaviors [I1, I6]. The market is still in its early stages, characterized by experimentation with various pricing strategies [I6]. In this context, NVIDIA's role is crucial. They are the only hardware supplier for this market and even invest in startups via access tokens instead of cash. Generally, there appears to be a dependency on big tech companies in the computing power market, especially cloud computing, as there is already a large market concentration (Davies, 2024). This is also shown by the strategic partnership between Microsoft and Mistral AI.

Although coordinated behaviors were not observed, there is a possible risk of relying on major entities, especially in the case of possible changes in terms of use, which affect the pricing and viability of established business processes [I1-I3]. The integration of AI into business operations often occurs through cloud platforms, where the utilization of LLMFPs infrastructure is widespread, creating dependencies as a result [I2- I4]. In addition, many technology departments in companies have been shown to not have the skills necessary to develop complex models [I3, I4]. In this context, large LLMFPs act as crucial partners for the implementation of models in companies [I2].

Network Effects

Generally, companies are encouraged to integrate LLM tools into their daily business workflows, as these tools are already used frequently by individuals in their private lives [I8]. There, this has already led to

increased efficiency and creativity, making employees question whether their companies would allow their use. This can even become a deciding factor for talents when choosing which company to join. Enterprises like Microsoft benefit from the dynamics of these network effects, as it is comparatively easy for them to integrate LLMs into their existing product user interface [I6]. In addition, they also benefit when customers engage with established products, such as those offered within the domain of cloud computing [I4]. Customers often want to analyze their data directly in the software architecture (e.g., CRM runs on Azure) by accessing the chatbot already integrated in the CRM [I6]. Drawing parallels with the dynamics observed in the cloud market, it becomes apparent that the interplay within this ecosystem serves to intensify the competitive landscape [I4, I6].

Another point is that startups that use LLMs in their business model have in mind to use multiple LLMFPs to power their solutions [I1]. The advantages of multi-homing to avoid dependencies, particularly in relation to pricing strategies and the ability to switch providers, are widely recognized [I1-I4]. However, this is only applicable if it is technically and economically reasonable. In this context, there are doubts about the feasibility of multi-homing in the current landscape of LLMs, highlighting the technical challenges due to the unique characteristics of each LLM [I2, I4]. Another observed issue is the lack of funding for startups, which prevents them from using multiple LLMs simultaneously [I1]. On the contrary, there is a tendency for companies to use multiple models simultaneously based on their specific needs, costs, and product requirements.

There are platforms for the concurrent use of LLMs, such as Amazon Bedrock, which allows the user to choose from different models, such as Titan, AI21 and Claude [I5]. Platforms like Flowise AI integrate various models for tasks such as document analysis, optimizing cost and efficiency. And Perplexity (Pro) as LLMLP lets the user choose between different models, such as GPT-4 or Claude 2 for its answer engine (Perplexity, n.d.).

LLMFPs certainly want to expand vertically to control application layers without intermediaries (LLMLP), thus securing access to more data [I7]. GPT-4 and Google's Gemini model are already natively multimodal, and future LLMs will efficiently integrate various model functionalities [I5]. Any company with the opportunity to collect more data on a large scale can become a competitor in the race for artificial general intelligence [I7].

The GPT Store will probably help in the selection and suggest certain tools from the store [I7]. Manual searches are no longer necessary, and solutions can be recommended on a customized basis. Tools may be able to position themselves better through additional payments to reach more customers, creating a win-win situation for OpenAI and tool providers. In addition to different stores, APIs will also remain important.

Recent M&A Activities

Companies strengthen their market dominance through strategic acquisitions [I3]. Acquisitions like these are not exclusive to the LLM market. They occur in various markets and sectors when they align with the interests of larger companies. Specific examples, such as Getty Images and Google Deep Mind, are mentioned to illustrate this pattern. There is a notably high frequency of large-scale takeovers and smaller acquisitions in the M&A landscape, as the M&A market holds important importance for LLM companies. However, in recent years have not appeared to be noteworthy M&A activities in the LLM market [I1]. Only small acquisitions were observed in which large players bought smaller providers to complete components of their software, but there were no large acquisitions between two main players [I6]. An explanation for this could be that the market is still in its initial states and companies usually need a few years to grow [I1]. From a perspective of a VC, this would not be the desired type of merger, as this would be rather an acquihire. However, a possible strategy for smaller LLMLPs could be to offer a lucrative niche offering and then sell to larger companies, allowing them to benefit from the customer relationships of acquired companies [I5].

In contrast, there appear to be some uncertainties about the market currently [I1]. For many VC investors, it is currently unclear whether LLMLPs can offer something new and create enough IP on top of the foundation LLMs, which is likely also applicable for all other industries [I1, I7]. On the other hand, investments in LLMFPs and subsequent acquisitions from strategic and institutional investors are likely to continue [I7]. This trend is driven by the substantial capital requirements of these companies. An example of such a strategic investment is the Series B round of Aleph Alpha, which involved Bosch.

Regulatory Environment

When discussing the regulatory environment in the context of M&A activities, the presence of various regulations is affirmed, highlighting the importance of the cartel office as a relevant authority [I3]. An illustrative example is the proposed merger between Adobe and Figma, which was ultimately abandoned due to regulatory complications.

Currently, the regulatory process of LLM companies appears to be such that the development of case law precedes the enactment of legislation [I1]. This situation is unsettling for companies with respect to the implementation of LLMs in their processes due to the potential legal risks associated with such integration [I8]. As a result, there are various approaches to the usage of these offerings among different institutions.

However, for most LLMLPs, regulation will not be a major challenge unless they operate in the medical or military sector [I7]. Otherwise, there will only be minor adjustments to the general terms and conditions of the company, while there are certainly also differences between different markets such as the US and Europe. The entire regulatory field with respect to copyrights remains "wild west", as there are no legally binding interpretations yet, but it is likely that a certain market standard will be established [I6]. For these instances, it can be assumed that the issue is not the usage itself, but rather the lack of compensation for authors whose works are used to train the LLMs [I8]. Considering this, one approach could be to designate non-critical areas as sandboxes for developing best practices. Currently, despite the absence of legal frameworks for LLMs, OpenAI is facing legal challenges, including a lawsuit from The New York Times [I5]. Large companies will most likely face litigation when training models without clear agreements. Although newer companies benefit from anticipating legal challenges, it may not necessarily accelerate their development. Overregulation risks hindering model development, yet it offers protection for smaller entities by safeguarding skills and creativity. However, it is important to avoid overregulation to maintain competitiveness in Europe [I8].

The aforementioned findings can be in summary form in the Appendix D, in Table D 1, along with a summary overview of the statements of each expert interviewed in Table D 2.

Discussion

Principal Findings

The IMF forecasts that inflation in the global economy will start to decline, and growth will continue (Gourinchas, 2024). Additionally, GenAI has the potential to increase global GDP by 7%, or almost USD 7 trillion, over a 10-year period (Goldman Sachs, 2023). Moreover, it could lift productivity growth by 1.5%-points in the same period. On the other hand, AI systems could expose 300 million full-time jobs to automation.

This shows the importance of understanding the dynamics in the GenAI market for the economy, but also illustrates the possible major impacts on society. Our work improves the understanding of the GenAI market and its competitive dynamics by specifically analyzing a subset, the LLM market, and exploring its dynamics. In this context, we successfully planned and conducted market research and expert interviews. The interviews were then analyzed and screened for characteristics that are commonly used to assess the market dynamics. Based on our research, we discovered that there are two major groups in this market, the LLMFPs and LLMLPs, with different market dynamics. Therefore, our principal findings and implications are divided into these two groups.

LLMFPs

The findings of our work in the context of LLMFPs implicate that there are high barriers to entry into the market due to the nature of this technology. Our experts named challenges, i.e. the resources required to train foundation models such as computing power, data access, and skilled personnel. Litigations with other companies represent another barrier to entering the LLMFP market. Our market research indicates that overcoming these challenges requires a considerable amount of capital, as many established companies in this domain have already undergone multiple funding rounds. For example, OpenAI secured funding of USD 11.3 billion. This could be one of the reasons why there are currently only a few dominant LLMFPs on the market. Additionally, it seems that LLMFPs are offering similar products and their model performance

is likely to converge. Furthermore, increasing the similarity in research approaches among various LLMFPs could also contribute to this convergence. An additional factor appears to be that the best performing models are often also the cost-efficient ones. Consequently, this increases the opportunity costs for other actors even further, discouraging them from relying solely on available APIs of existing LLMFPs. Current models from major LLMFPs such as OpenAI or Anthropic are proprietary, but there are also open-source models, such as Meta's Llama 2. One perspective is that it is not certain that these models will always stay open-source, and that monetary incentives are playing a key role. Therefore, the current observations point to a trend towards a consolidated LLMFP market with a few main players prevailing, and no new main players entering this market. However, there are also other opinions that open-source models improve current models and foster competition. This is supported by the indicator that our experts interviewed did not identify major M&A activities, which could be explained by the fact that the market is still in the early stages and too early for this type of transaction. However, it appears that there is a general trend that LLMFPs are dependent on larger tech companies, such as Microsoft, especially in computing power.

It is crucial to note that the Federal Cartel Offices could play an important role if the trend is developing in the direction of a complete consolidated market. However, legislators and governments currently seem to be more focused on the regulation of the deployment and usage of AI systems, rather than the market dynamics, for example, with the approved EU AI Act (European Commission, 2021; European Parliament, 2024). The implications of our results for lawmakers are that they should proceed with caution, as overregulation may put LLMFPs within the EU at a competitive disadvantage in the global market.

LLMLPs

For LLMLPs, however, it is relatively easy to enter the LLM market, as foundation models are often straightforwardly accessible by the APIs of the LLMFPs. Additionally, these APIs make it easier for competitors to enter the LLMLP market, especially for those who do not have an engineering background. With this API access and an additional layer, these models can then be customized to their needs and integrated as a feature in existing offerings or placed as new ones. As a result, there are various actors on the LLMLP market. There also appears to be a trend toward more personalized LLMs, which implicates additional market opportunities for new offerings. However, our research findings suggest that companies with existing widely used offerings, such as Microsoft, have a competitive advantage. These companies can just distribute their new offerings through this channel. Furthermore, new companies have other disadvantages. Our result indicates that LLMLPs also need to ensure that they generate enough IP atop the foundation model with their offerings. This is especially the case for LLMLPs with a flat layer; as stated, the barrier for new competitors to enter the market is relatively low. Another risk is that LLMFPs could integrate features similar to LLMLPs in their offerings or shut down the API, making the LLMLP obsolete or unable to continue. However, this appears to be unlikely for many LLMLPs offerings, as the LLMFPs probably rather will focus on optimizing their models and act as a wholesale with their APIs. Nevertheless, many LLMLPs have these risks in mind, why they plan to protect themselves with multi-homing, or are already doing so as part of their offerings. Yet, many new LLMLPs do not have the resources to do so, making them vulnerable.

The M&A activity landscape for LLMLPs is comparable to that of LLMFPs, and no notable activities have been observed. Similarly, the regulatory environment for LLMLPs mirrors that of LLMFPs. However, LLMLPs have a distinct advantage: they are responsible only for the data used in customization, not for the extensive data set required to train the foundational model. In addition to sensitive sectors such as military and health, our research suggests that LLMLPs can quickly respond to regulatory changes by adapting their terms and conditions.

Implications for Research and Practice

The research findings presented in this study carry crucial implications for both academia and various stakeholders involved in the field of GenAI, particularly within the LLM domain. Exploring LLMFPs reveals high barriers to entry and potential market consolidation, suggesting a trend towards a few dominant players. This insight has implications for policymakers, urging them to carefully balance regulatory actions, as overregulation could put LLMFPs within the EU at a competitive disadvantage in the global market. Additionally, the identification of distinct dynamics for LLMLPs, characterized by lower entry barriers but potential vulnerability, emphasizes the need for these providers to strategize IP protection and consider

multi-homing approaches. Furthermore, the study sheds light on the regulatory environment, emphasizing the unique advantage of LLMLPs in responding to regulatory changes more swiftly. These implications not only inform policymakers and industry players, but also raise awareness among researchers about the evolving landscape of GenAI. The identified trends towards market consolidation and the interplay between LLMFPs and LLMLPs serve as a foundation for future studies in understanding the complex dynamics of the LLM market. Researchers should continue to explore the economic and societal implications of GenAI and foster an ongoing dialogue with policymakers to ensure that regulatory frameworks adapt to market dynamics without stifling innovation. Furthermore, researchers should explore deeper the dynamics of the LLM markets and explore the impacts of potential regulatory frameworks.

Limitations

In the absence of an established body of literature related to the research topic, the methodology used in this study was driven by the need to gather first-hand insights and perspectives. Given the nascent nature of the chosen topic, the scarcity of reviewed literature compelled the adoption of a primary data collection method: expert interviews. In situations where the existing literature fails to provide comprehensive insights, interviews emerge as a valuable means to bridge the gap.

It is important to acknowledge certain limitations inherent in the research design, particularly with respect to the sample group. A notable weakness lies in the limited scope of the sample size of eight participants, which may constrain the generalizability of the findings to the broader industry (Alshenqeeti, 2014). The subjective nature of interview interactions raises a critical concern about potential interviewer bias. Unconscious biases stemming from preconceived notions, cultural background, or personal experience can subtly permeate the interview process, leading to biased responses from participants and compromising the integrity of the research (Alsaawi, 2014). Additionally, no LLMFP companies were interviewed as part of the study. Therefore, the results do not include the perceptions of a key stakeholder group in the market. Furthermore, the interviews were conducted from a predominantly European or western perspective. This means that the findings may not be applicable to non-Western markets, or that additional insights could be gained from these markets.

When interpreting the results, it is important to keep in mind the rapidly changing nature of the LLM market. The market is still evolving, and innovations are emerging frequently. The interview results only capture the interviewee's perspective on the current market situation.

Finally, it must be noted that the market analysis performed also has some limitations. It was not always possible to trace the exact calculations of the market sizes and forecasts. The gray literature used often consists of reports from mostly commercially motivated market research or consulting institutions, which make this information available only to paying customers and do not always publish their research approach or the basis for data collection.

Future Research

In the last decade, the number of AI publications has already doubled (AI Index Stanford, 2023), which was followed by notable advances in LLMs since 2022 (Douglas, 2023) and a fast-growing market (IOT Analytics, 2023). In addition to examining these market dynamics regarding LLM providers and their platforms, it would also be interesting to conduct further research regarding enabling hardware, as there are even stronger monopolistic tendencies observable (IOT Analytics, 2023). The interdependencies between software providers and hardware manufacturers remain a field of tension, as OpenAI has announced that it is willing to play a pivotal role in the hardware segment as well (Tong et al., 2023). The market could also be further categorized into sub-segments such as foundation models, open-source, and closed-source, following an in-depth analysis of each segment. Just as hardware is an enabler for LLMs, one could take the next step and examine for which industries the LLMs could become a mayor enabler in the upcoming years. Based on these findings, it would also be possible to derive recommendations for strategic action for various industries, corporations, small and medium-sized enterprises, and startups. The impact of LLMs on the entire software sector remains particularly interesting. A comparison with other technologies such as databases and cloud computing could also be used to identify potential risks and trends. Furthermore, the regulatory situation remains largely unresolved, so an exchange with political and legal experts would be valuable. In addition to the international market, it is also possible to examine strong

regional markets, such as the Chinese. Finally, a visualization, analysis, and categorization of all relevant international LLM players would provide a helpful overview for future research activities.

Conclusion

In conclusion, our research provides valuable insight into the competitive dynamics of the GenAI market, with a particular focus on the LLM segment. Through a comprehensive analysis of expert interviews and market research, we have identified two primary groups shaping this market: LLMFPs and LLMLPs.

For LLMFPs, our findings highlight substantial barriers to entry, including the substantial resources required for model training, such as computing power, data access, and skilled personnel. The concentration of capital among a few dominant players underscores the challenges faced by new entrants. Additionally, the convergence of model performance and the potential consolidation of the market suggest a future in which a handful of large players will dominate. However, new business applications of existing models could shift the landscape in favor of competitors. One example is the emergence of local LLMs operating on end-user devices. Furthermore, regulatory caution is warranted to avoid over-regulation that could disadvantage LLMFPs in global competition.

On the contrary, LLMLPs face relatively lower entry barriers and use the APIs provided by LLMFPs to develop tailored services. Although this accessibility promotes market diversity and innovation, LLMLPs must address risks such as dependence on LLMFPs and the potential integration of similar features by larger players. Despite these challenges, LLMLPs can adapt quickly to regulatory changes and focus on creating IP on top of foundation models to differentiate themselves in the market.

In general, our research highlights the complex interplay between market consolidation, regulatory frameworks, and technological advancements in the LLM industry. The potential emergence of a consolidated market dominated by a few large players poses a substantial threat to competition and innovation. As GenAI increasingly influences the global economy and societal dynamics, stakeholders must exercise greater vigilance to effectively navigate this evolving landscape. Understanding the implications of concentrated market power is crucial for policymakers to foster innovation while ensuring fair competition and consumer interests in the LLM market.

References

Abraham, C., Boudreau, M.-C., Junglas, I., & Watson, R. (2013). Enriching our theoretical repertoire: the role of evolutionary psychology in technology acceptance. *European Journal of Information Systems*, *22*(1), 56–75. **https://doi.org/10.1057/ejis.2011.25**

AI Index Stanford. (2023). *AI Index Report 2023 – Artificial Intelligence Index* (pp. 1–386) [Market Report]. Standford University. **https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf**

Alsaawi, A. (2014). A Critical Review of Qualitative Interviews. *European Journal of Business and Social Sciences*, *3*(4), 149–156. **https://doi.org/10.2139/ssrn.2819536**

Alshenqeeti, H. (2014). Interviewing as a Data Collection Method: A Critical Review. *English Linguistics Research*, *3*(1), 39–45. **https://doi.org/10.5430/elr.v3n1p39**

Anton, J. J., & Gertler, P. J. (2004). Regulation, Local Monopolies and Spatial Competition. *Journal of Regulatory Economics*, *25*(2), 115–141. **https://doi.org/10.1023/B:REGE.0000012286.33952.6c**

Bailey, J. (2008). First steps in qualitative data analysis: transcribing. *Family Practice*, *25*(2), 127–131. **https://doi.org/10.1093/fampra/cmn003**

Bain, J. S. (1956). *Barriers to New Competition*. Harvard University Press. **https://doi.org/10.4159/harvard.9780674188037**

Banks, J., & Warkentin, T. (2024, February 21). *Gemma: Introducing new state-of-the-art open models*. Google. **https://blog.google/technology/developers/gemma-open-models/**

Birch, K., & Bronson, K. (2022). Big Tech. *Science as Culture*, *31*(1), 1–14. **https://doi.org/10.1080/09505431.2022.2036118**

Bloomberg Intelligence. (2023). *2023 Generative AI Growth report* [Growth Report]. **https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/**

Bucher, R., Fritz, C. E., & Quarantelli, E. L. (1956). Tape Recorded Interviews in Social Research. *American Sociological Review*, *21*(3), 359–364. **https://doi.org/10.2307/2089294**

Cheng, L., & Liu, X. (2023). From principles to practices: the intertextual interaction between AI ethical and legal discourses. *International Journal of Legal Discourse*, *8*(1), 31–52. **https://doi.org/10.1515/ijld-2023-2001**

CMA. (2020). *Online platforms and digital advertising* (pp. 1–437) [Market study final report]. Competition & Markets Authority. **https://assets.publishing.service.gov.uk/media/5fa557668fa8f5788db46efc/Final_report_Digital_ALT_TEXT.pdf**

Crunchbase. (n.d.). *Crunchbase: Discover innovative companies and the people behind them*. Crunchbase. Retrieved February 28, 2024, from **https://www.crunchbase.com**

Davies, P. (2024, February 27). *Microsoft's deal with Mistral AI faces criticism and further scrutiny*. Euronews. **https://www.euronews.com/next/2024/02/27/furious-critics-question-microsofts-deal-with-mistral-ai-as-eu-set-to-look-into-it**

DeMarrais, K., & Lapan, S. D. (Eds.). (2004). *Foundations for research: methods of inquiry in education and the social sciences* (1st ed.). Lawrence Erlbaum.

Dimension Market Research. (2023). *Generative AI Market By Component (Software, Services), By Technology, By Application- Global Industry Outlook, Key Companies (Synthesia, MOSTLY AI Inc., Genie AI Ltd. and others), Trends and Forecast 2023-2032* (Market Report No. IT-117; pp. 1–267). **https://dimensionmarketresearch.com/report/generative-ai-market/overview**

Douglas, M. R. (2023). *Large Language Models*. 1–47. **https://doi.org/10.48550/arXiv.2307.05782**

Eisenmann, T., Parker, G., & Van Alstyne, M. (2006). Strategies for two-sided markets. *Harvard Business Review*, *84*, 92-101+149.

European Commission. (2021). *COM/2021/206 (Draft)*. Publications Office of the European Union. **https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021PC0206**

European Parliament. (2024, March 13). *Artificial Intelligence Act: MEPs adopt landmark law | News | European Parliament*. News European Parliament. **https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law**

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. **https://doi.org/10.1007/s11023-020-09548-1**

Fortune Business Insights. (2023). *Generative AI Market Size, Share & Trends Analysis* (pp. 1–120) [Market Report]. Fortune Business Insights. **https://www.fortunebusinessinsights.com/generative-ai-market-107837**

Friedman, J. (1982). Oligopoly theory. In *Handbook of Mathematical Economics* (Vol. 2, pp. 491–534). North-Holland Sole distributors for the U.S.A. and Canada, Elsevier North-Holland. **https://doi.org/10.1016/S1573-4382(82)02006-2**

Geertsema, P. G., Bifet, A., & Green, R. (2023). *ChatGPT and Large Language Models: What are the Implications for Policy Makers?* 1–32. **https://doi.org/10.2139/ssrn.4424048**

Goldman Sachs. (2023, April 5). *Generative AI Could Raise Global GDP by 7%*. Goldman Sachs. **https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html**

Gourinchas, P.-O. (2024, January 30). *Global Economy Approaches Soft Landing, but Risks Remain*. IMF. **https://www.imf.org/en/Blogs/Articles/2024/01/30/global-economy-approaches-soft-landing-but-risks-remain**

Grand View Research. (2023). *Generative AI Market Size, Share And Growth Report, 2030* (Market Report No. GVR-4-68040-011-4; pp. 1–100). **https://www.grandviewresearch.com/industry-analysis/generative-ai-market-report**

Grand View Reserach. (2024). *Generative AI Market Size, Share & Trends Analysis Report* (Market Report No. GVR-4-68040-186-2; pp. 1–100). **https://www.grandviewresearch.com/industry-analysis/large-language-model-llm-market-report**

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288. **https://doi.org/10.1177/1049732305276687**

IOT Analytics. (2023). *Generative AI Market Report 2023–2030* (pp. 1–154) [Market Report]. **https://iot-analytics.com/wp/wp-content/uploads/2023/12/INSIGHTS-RELEASE-The-leading-generative-AI-companies.pdf**

Kamps, H. J. (2023, November 6). Startups learn the hard way that relying on OpenAI's tech can burn them. *TechCrunch*. **https://techcrunch.com/2023/11/06/get-the-pdf-outta-here/**

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 1–9. **https://doi.org/10.1016/j.lindif.2023.102274**

Kollnig, K., & Li, Q. (2023). Exploring Antitrust and Platform Power in Generative AI. *arXiv*, 1–3. **https://doi.org/10.48550/arXiv.2306.11342**

Kopel, M. (2009). Oligopoly Dynamics. In *Handbook of Research on Complexity* (1st ed., pp. 124–168). Edward Elgar Publishing. **https://www.elgaronline.com/edcollchap/edcoll/9781845420895/9781845420895.00012.xml**

Kosma, B. (2023, December 6). Which companies are working on LLMs and ChatGPT alternatives? *Tech Monitor*. **https://techmonitor.ai/technology/companies-large-language-models-llms-chatgpt-alternatives**

Luitse, D., & Denkena, W. (2021). The great Transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, *8*(2), 1–14. **https://doi.org/10.1177/20539517211047734**

Lutkevich, B. (2023, October 3). *16 of the best large language models*. WhatIs. **https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models**

Markets and Markets. (2023). *Generative AI Market by Offering* (pp. 1–541) [Market Report]. Markets and Markets. **https://www.marketsandmarkets.com/Market-Reports/generative-ai-market-142870584.html**

Marr, B. (2023, May 30). *10 Amazing Real-World Examples Of How Companies Are Using ChatGPT In 2023*. Forbes. **https://www.forbes.com/sites/bernardmarr/2023/05/30/10-amazing-real-world-examples-of-how-companies-are-using-chatgpt-in-2023/**

Meta. (n.d.). *Llama 2*. Meta AI. Retrieved January 16, 2024, from **https://ai.meta.com/llama-project**

Miller, R. A. (1967). Marginal Concentration Ratios and Industrial Profit Rates: Some Empirical Results of Oligopoly Behavior. *Southern Economic Journal*, *34*(2), 259–267. **https://doi.org/10.2307/1055042**

Morris, M. R. (2023). Scientists' Perspectives on the Potential for Generative AI in their Fields. *arXiv*, 1–26. **https://doi.org/10.48550/arXiv.2304.01420**

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, *17*(1), 2–26. **https://doi.org/10.1016/j.infoandorg.2006.11.001**

Narechania, T. N. (2021). Machine Learning as Natural Monopoly. *107 Iowa Law Review 1543 (2022)*, 1543–1614. **https://doi.org/10.2139/ssrn.3810366**

Nidumolu, J., Soni, A., & Dang, S. (2024, March 1). Elon Musk sues OpenAI for abandoning original mission for profit. *Reuters*. **https://www.reuters.com/legal/elon-musk-sues-openai-ceo-sam-altman-breach-contract-2024-03-01/**

Novet, J. (2024, March 12). *OpenAI denies Elon Musk lawsuit claim that there ever was founding agreement*. CNBC. **https://www.cnbc.com/2024/03/11/openai-denies-musk-lawsuit-claim-that-there-was-founding-agreement.html**

Onozaki, T., & Yanagita, T. (2003). Monopoly, oligopoly and the Invisible Hand. *Chaos, Solitons & Fractals*, *18*(3), 537–547. **https://doi.org/10.1016/S0960-0779(02)00675-6**

OpenAI. (n.d.). *GPT-4*. Retrieved January 19, 2024, from **https://openai.com/gpt-4**

OpenAI. (2023, November 6). *Introducing GPTs*. Introducing GPTs. **https://openai.com/blog/introducing-gpts**

Perplexity. (n.d.). *What model does Perplexity use and what is the Perplexity model?* Retrieved March 14, 2024, from **https://www.perplexity.ai/hub/technical-faq/what-model-does-perplexity-use-and-what-is-the-perplexity-model**

Pichai, S., & Hassabis, D. (2024, February 15). *Our next-generation model: Gemini 1.5*. Google. **https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/**

Podszun, R. (2017). *Innovation, Variety & Fair Choice – New Rules for the Digital Economy: Expert Opinion for Finanzplatz München Initiative (FPMI)*. 1–54. **https://doi.org/10.2139/ssrn.3243403**

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. 1–12.

Rosenthal, M. (2016). Qualitative research methods: Why, when, and how to conduct interviews and focus groups in pharmacy research. *Currents in Pharmacy Teaching and Learning*, *8*(4), 509–516. **https://doi.org/10.1016/j.cptl.2016.03.021**

Ruutu, S., Casey, T., & Kotovirta, V. (2017). Development and competition of digital service platforms: A system dynamics approach. *Technological Forecasting and Social Change*, *117*, 119–130. **https://doi.org/10.1016/j.techfore.2016.12.011**

Rysman, M. (2009). The Economics of Two-Sided Markets. *Journal of Economic Perspectives*, *23*(3), 125–143. **https://doi.org/10.1257/jep.23.3.125**

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., … Wolf, T. (2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 1–73. **https://doi.org/10.48550/arXiv.2211.05100**

Schmid, P., Sanseviero, O., & Cuenca, P. (2024, February 21). *Welcome Gemma - Google's new open LLM*. Hugging Face. **https://huggingface.co/blog/gemma**

Sejnowski, T. J. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation*, *35*(3), 309–342. **https://doi.org/10.1162/neco_a_01563**

Statista. (2023). *Generative AI - Worldwide | Statista Market Forecast*. Statista. **https://www.statista.com/outlook/tmo/artificial-intelligence/generative-ai/worldwide**

Stigler, G. J. (1950). Monopoly and Oligopoly by Merger. *The American Economic Review*, *40*(2), 23–34. JSTOR. **http://www.jstor.org/stable/1818020**

Strauss, A. L., & Corbin, J. M. (2003). *Basics of qualitative research: techniques and procedures for developing grounded theory* (2. ed.). Sage Publ.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. **https://doi.org/10.1038/s41591-023-02448-8**

Tong, A., Cherney, M. A., Bing, C., Nellis, S., Tong, A., & Bing, C. (2023, October 6). Exclusive: ChatGPT-owner OpenAI is exploring making its own AI chips. *Reuters*. **https://www.reuters.com/technology/chatgpt-owner-openai-is-exploring-making-its-own-ai-chips-sources-2023-10-06/**

Valuates Reports. (2023). *Global Large Language Model (LLM) Market Research Report 2023* (Market Report No. QYRE-Auto-30B13652; pp. 1–96). **https://reports.valuates.com/market-reports/QYRE-Auto-30B13652/global-large-language-model-llm**

Varian, H. R. (2010). *Intermediate microeconomics: a modern approach* (8th ed). W.W. Norton & Co.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need. *arXiv*, 1–15. **https://doi.org/10.48550/arXiv.1706.03762**

Verdegem, P. (2024). Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. *AI & SOCIETY*, *39*(2), 727–737. **https://doi.org/10.1007/s00146-022-01437-8**

Wiesche, M., Jurisch, M. C., Yetton, P. W., & Krcmar, H. (2017). Grounded Theory Methodology in Information Systems Research. *MIS Quarterly*, *41*(3), 685-A9. **https://doi.org/10.25300/MISQ/2017/41.3.02**

Appendix

A. Interview Guideline

1. Introduction (5 mins)

Introduction and further explanation of the purpose of the interview. Clarifying organizational details (e.g., recording, time restrictions) and structure of the interview.

1.1 Interview objective and structure
1. Interviewer: Brief introduction to each other **2. Further explanation of the purpose of the interview:** **a.** Experts' perspective on the current market situation in the LLM field. **Clarifying organizational details:** **3. Interview Duration** a. Planned time frame 30 mins (scheduled 45 mins, incl. buffer) **4. Interview Structure** a. Introduction (Interviewee background & experience, conceptual basis for the research subject) b. Question phase on the interviewee's experience with LLM to identify certain characteristics that might be an indicator for a consolidated market that may lead to a Monopoly / Oligopoly c. Deep dive into adoption aspects mentioned (e.g. Characteristics / Framework for Oligopolies), as needed. d. Closure/ Next Steps e. This agenda is only an orientation. Feel free to ask questions anytime or interrupt me to share your view and address further aspects that might be important. **5. Interview Recording** a. The interview will be recorded. b. All data will be **anonymized** and will **not allow any conclusions** regarding individual persons or companies. **6. Study results will be provided if you wish.** *Do you have any further questions or requests for the interview?*
1.2 Bridging the gap between the Market structure of LLM and its consequences
In the contemporary landscape of GenAI, there exists the concern of established LLM providers to accumulate power, displace smaller competitors, and assimilate more solutions into their platforms. Therefore, only a few providers could deploy infrastructure for many dependent companies. Major LLM providers currently include OpenAI (GPT-4), Anthropic (Claude), Google (Gemini), and Meta

(Llama 2). However, the latter LLM is open-source. ChatGPT recently presented its new GPTs, which enable the integration and creation of plugins, cementing the transformation towards a platform provider. Conversely, platformization can lead to monopolization, which incorporates the ability to implement constraints on smaller entities and clients and jeopardizes future innovation and competition.

Typical criteria for consolidated markets that we would like to look at are market concentration, entry barriers, product differentiation, interdependencies, network effects, M&A activities, and regulation.

Your opinion will be very valuable as research input to gain insights from your job role in practice.

1.3 Personal Introduction/Who is in the room?

Start Recording

At first, the interviewee's background and experience are gathered to understand the interviewee's general attitude towards the LLM market in their job role.

- **Interviewee:** LLM provider, customer, investor, or industry expert with at least one to two years of professional experience in related fields such as Artificial Intelligence, Large Language Models, etc.

- **Demographics**
  - Job title
  - Industry and Company Size
  - Years of professional experience
  - Working experience/ experience with LLM, perception of LLM (e.g. how would you rate the importance of LLM?)
  - Location
  - Age
  - Gender

2. Question Phase (20 mins)

Please share any thoughts or experiences that come to your mind regarding the dynamics of the LLM market

a) Market Concentration

- Would you consider the market power/share of the big players opposes a threat for other competitors, and why?
- How might the consolidation of market power among a few key players influence user experience and preferences?

b) Barriers to Entry

- Consolidated markets often have high barriers to entry, making it difficult for new competitors to enter the market. What barriers do you see for new players entering the LLM market (e.g., data access, computational power, intellectual property)?
- How easy is it for firms to copy features from other competitors and implement them into their own model?

c) Differentiation

- In a consolidated market, there's often a lack of remarkable product differentiation. How do you see product differentiation playing out in the LLM market currently and in the future?
- What are your observations regarding pricing strategies of LLM providers?
- Are there niche markets, products or features where smaller companies could be successful?

d) Interdependence

- Is there strategical interdependence among current big players, meaning that the actions of one firm can strongly impact the others (e.g., pricing strategies or coordinated behavior)?
- How do you perceive the openness of LLM to complementary services (e.g. via API)?

e) Network Effects

- Are there network effects in the market that favor a winner-take-all dynamic (e.g., high switching costs)?
- Can you observe any trends toward multi-homing (using various big providers in parallel)?

f) Recent Mergers & Acquisition activities?

- Are a few companies acquiring others, leading to increased concentration, and do you see any emerging dominant players because of these activities?
- *How do ongoing M&A activities influence innovation within the GenAI or LLM sector, and does increased consolidation affect healthy competition?*

g) Regulatory Environment

- Do regulatory authorities restrict entry or influence the behavior of firms (e.g., EU AI Act)?
- *Do regulatory restrictions create major barriers for new entrants looking to participate in the GenAI or LLM market?*

Stop Recording

3. Closure (5 mins)

Conclusion and Next Steps

- Over the next month, we will conduct further stakeholder interviews and inform you about the evaluation and results if you wish.
- Are there any colleagues or contacts in your network who might also be interested in contributing to our research by providing insights during an interview?
- How do you feel about the interview? Are there any further questions or suggestions/ recommendations you would like to make regarding the interview itself or the overall study?

Thank you for your time and the valuable insights provided!

B. GenAI Market Share

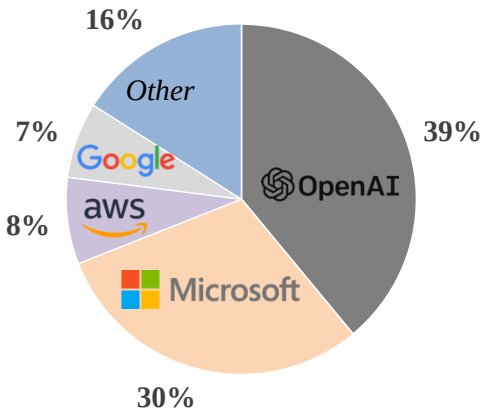

Figure B 1. Market Shares of GenAI Foundation Models and Platforms in 2023 According to IOT Analytics (2023)

C. Global market sizes of LLMs and GenAI

Source	Market	Start	End	CAGR
Valuates Reports (2023)	LLM	USD 10.5 billion (2022)	USD 40.8 billion (2029)	35.9%
Grand View Research (2024)	LLM	USD 4.35 billion (2023)	USD 37.24 billion (2030)	21.4%
Markets and Markets (2023)	GenAI	USD 11.3 billion (2023)	USD 76.8 billion (2030)	31.5%
Dimension Market Research (2023)	GenAI	USD 14.9 billion (2023)	USD 266 billion (2032)	37.8%
Bloomberg Intelligence (2023)	GenAI	USD 40 billion (2022)	USD 1.3 trillion (2023)	42%
Fortune Business Insights (2023)	GenAI	USD 29 billion (2022)	USD 668 billion (2030)	47.5%
Grand View Research (2023)	GenAI	USD 10.14 billion (2022)	USD 115.9 billion (2030)	35.6%
Statista (2023)	GenAI	USD 66.62 billion (2024)	USD 207 billion (2030)	20.8%

Table C 1. Global Market Size of LLMs and GenAI with Forecast

D. Interview Results

D.1 Summary Findings

Criteria \ Market	LLMFP	LLMLP
Market Concentration	Currently, there are only a few dominant players on the market. Open-source models as counterparts, however, it is unclear if they will always remain open-source.	There are various LLMLPs existing on the market, and there appears to be a trend for even more personalized LLMs.
Barriers to Entry	There are high barriers, such as computing power or hiring, existing, which require substantial funding to overcome.	Simplified entry into the market, as less engineering skills are required, expanding the set of competitors.
Differentiation	Similar offerings, with different performance. Innovations, such as local LLM on end devices, could disrupt the market. There is a market for niche models.	Niche markets also exist for LLMLPs. Startups have the advantage of being agile. However, the offer could be displaced by larger competitors.
Interdependence	Will likely focus on wholesale of API request rather than cannibalize startups. Therefore, they also increase the opportunity for everyone to enter the market.	The LLMLP are required to build enough IP on top of the foundation models. Using open-source models can additionally protect LLMLPs from full dependencies of LLMFPs. The choice of LLM depends mainly on performance.
Network Effects	Generally, companies are encouraged to integrate LLMs in their workflows due to the success of LLMs in the private environment. Companies that are already widely established in this environment, such as Microsoft with its Office Suite, therefore, have the advantage of placing themselves as LLMLPs. Many LLMLPs already have in mind to use multiple LLMs, and some platforms allow users to choose which models should be used.	
Recent M&A Activities	No major acquisition between two large players was observed. Probably too early in the market.	The same applies to LLMLPs. Additionally, it is unclear whether enough IP is created so that the LLMLP itself is valuable. However, one strategy of these providers could be to penetrate a lucrative niche and then sell it.
Regulatory Environment	M&A of the few LLMFPs could be rejected by authorities similar to those of other industries. Many regulations do not yet exist, leaving unclarity about legal risks. However, overregulation should be avoided to preserve competitiveness.	Legal unclarity also applies for companies that want to integrate LLMs into their workflows. However, it appears that for most LLMLPs it is relatively easy to adapt to regulatory changes.

Table D 1. Summary of Dynamics in the LLM Markets

D.2 Overview Interview Results

Interview Partner / Market Dynamic Criteria	I1: VC Investor	I2: Consultant	I3: Entrepreneur	I4: Consultant 2	I5: Consultant 3	I6: Corporate VC Investor	I7: Entrepreneur 2	I8: Advisor
Market Concentration	Only three to four major LLMFPs. However, there is a higher variety of LLMLPs on the market.	There are some big players, that threaten other competitors. Nevertheless, this will foster competition.	A few very dominant Players that may pose a threat to other competitors, especially in model development.	Microsoft and OpenAI wield outstanding influence in a market dominated by three hyperscalers. Despite this, the presence of open-source models ensures a diverse and fair market, with many players developing smaller or specialized models.	Major players for LLMs include OpenAI, Google, and AWS, while new entrants like Aleph Alpha and Mistral enter the competition. Training has become more accessible, but legal and competitive factors shape the market.	Monopolization will happen with OpenAI as the biggest player. It is too late and will be too difficult to build a new foundation model.	No more than three LLMFPs will prevail, as there are no new hypotheses or insights.	No complete consolidated market to be expected, as there will be more personalized LLMs.
Barriers to Entry	High entry barrier to entering the market as an LLMFP, but as an LLMLP, lower-threshold.	The most remarkable challenge lies in ensuring sustainable introductions in an ever-changing environment.	Low entry barriers but intense competition on the market.	High costs, a scarcity of LLM expertise, and a niche problem in NLP. Small and medium-sized companies face additional hurdles compared to tech giants.	Entering the LLM market at scale is challenging; startups should find unique niches to compete effectively. Identifying untapped functionalities becomes crucial for success.	Computing capacities are needed. Talent in terms of engineering expertise is restricted, but regulatory expertise is important to deal with legal issues. Many LLMLP with flat technology on top were erased by the new OpenAI version.	For LLMFPs, the barriers (talent and research) are too high. For LLLMLP, the focus shifted from research expertise to programming skills through APIs.	High computing power and energy may pose a challenge to entering the market. Conversely, existing LLMs enable more people to enter the market, increasing the basic set of competitors.

Differentiation	LLMFPs seem to offer similar products, the performance of which currently changes from time to time.	Potential for models to be specialized for different industries, regions, or regulatory requirements. Smaller companies may establish in specific niches	Lack of differentiation and few new concepts or features of Foundation Models. Great potential for niche markets, though.	Differentiation of products using language models like GPT imposes great challenges due to complexity, cost barriers, and rapid technological evolution, impacting smaller companies' entry.	N/A	Target niches and specific verticals and integrate LLMs into existing solutions (as a new tool to handle unstructured data). While startups can win against established companies due to modern UI, Microsoft can leverage their existing products.	Performance as a key metric for LLMs, but price is also important (intelligence per Dollar). Better models are often cheaper.	Due to the arising of LLMFPs and LLMLPs, existing offerings needed to be changed. This will continue at a fast pace, as now people without engineering or mathematic skills can develop new ideas with new LLM offerings.
Interdependence	LLMFPs will likely focus on wholesale and monetizing API requests instead of competing with layer providers. Coordinated behavior or price fixing are not observable.	Small and Large Companies face crucial reliance on key players. Risk involved if API conditions, pricing and viability of business processes change.	High interdependence between provider and customer that uses the provider platform.	Closed-source models, like GPT 3.5, create high dependence on providers, impacting pricing and strategy. Open-source models with good licensing offer independence. Strong price interdependency exists among major players.	Major players' prices force others to follow suit, potentially leading smaller entities to financial challenges or bankruptcy.	No coordinated pricing, but NVIDIA has a lot of market power as the only hardware supplier. LLMFP monetize their APIs and increase opportunity costs for competitors.	For users, the LLM is not relevant and with interface layers, it could be possible to swap them in the future.	Started as accessible-to-everyone, however, today standing is not clear anymore.

Network effects	Startups have in mind to use multiple LLMFPs. However, they often do not have the fund to do so.	There are doubts about the feasibility of multi-homing since there are technical challenges involved due to the unique characteristics of each LLM.	Trend towards multi-homing to prevent reliance on a single company.	For streamlined integration, companies in a single-cloud setup, like Microsoft, find it easier to use corresponding models. Multi-cloud scenarios pose integration challenges despite the possibility of utilizing diverse providers.	Diverse LLMs can be utilized on platforms like Amazon Bedrock and Flowise AI, offering flexibility based on needs and applications. GPT-4 and Gemini are natively multimodal. Upcoming platforms will efficiently integrate diverse model functionalities.	Existing product portfolio (especially at Microsoft) can lead to network effects, as users often want to analyze their data in the software and use already integrated Chatbots (ecosystem).	LLMFPs want to expand vertically to gather more data, while the GPT store will become a monetized selection platform for tools.	Due to the accessibility and usage of LLMs in the private context, people experience a reduction in workload, increased efficiency, and creativity. Because of this, companies may be forced to also allow usage in the corporate environment to stay attractive for talents.
Recent M&A activities	Only investments, no recent M&A activities observed.	N/A	High frequency of large and small acquisitions.	N/A	Emphasizing the value of customer data, larger companies may acquire smaller firms for established clientele access, not just emulate them.	Small M&A activities to complete certain software components, but no large acquisitions between big players.	Normal industry regarding investments and M&A, while LLMFP will receive more funding from strategic and institutional investors.	N/A

Regulatory Environment	Jurisdiction preceding the enactment of legislation. Currently, users may be liable for the usage of the results of the LLMs if they are hallucinated.	N/A	Various regulations, highlighting the relevance of the Cartel Office for M&A activities.	N/A	Legal challenges for OpenAI arise as they trained on New York Times articles without clear legal basis. The pending EU AI Act brings uncertainty, leading larger firms to face potential lawsuits. Overregulation may slow model development, but offers protection to smaller entities like artists.	The regulatory field remains "Wild West" as there are no legally binding interpretations. It is likely that a certain market standard will be established.	For most LLMLPs, regulation will not be a major challenge (only in the medical or military sector), but certain differences between regions may arise.	Identify data source; compensate third-parties. Clarify, don't add rules. Avoid Europe's overregulation for competitiveness. Test in safe areas for best practices. Permission in some areas, e.g., universities, restricted. Address companies' legal worries under existing regulations.

Table D 2. Overview Interview Results