

Algorithmic Accountability

Transparency, Agency and Literacy in the Age of the Algorithm

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Dipl.-Ing. Florian Cech, BSc.

Registration Number 0325023

to the Faculty of Informatics

at the TU Wien

Advisor: Associate Prof. Dipl.-Ing. Dr.techn. Hilda Tellioglu

The dissertation has been reviewed by:

Christian Sandvig

Ben Wagner

Vienna, 7th November, 2022



Florian Cech

Algorithmic Accountability

Transparenz, Agency und Literacy im Zeitalter der Algorithmen

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

eingereicht von

Dipl.-Ing. Florian Cech, BSc.

Matrikelnummer 0325023

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Associate Prof. Dipl.-Ing. Dr.techn. Hilda Tellioglu

Diese Dissertation haben begutachtet:

Christian Sandvig

Ben Wagner

Wien, 7. November 2022



Florian Cech

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Florian Cech, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 7. November 2022



Florian Cech

Acknowledgements

“If I have seen further, it is by standing on the shoulders of Giants”

Sir Isaac Newton (1642-1727)

No human endeavour—scientific or otherwise—has ever been achieved truly solitarily, and this dissertation certainly exemplifies this notion particularly well. For their support and contributions, I owe a debt of gratitude to my advisor, my colleagues, my friends and family, to whom I would like to express my heartfelt thanks.

First, I would like to thank my advisor Hilda Tellioglu for her continuous support and encouragement in this endeavour, for her fierce and unwavering fight to keep me employed during this time, for her willingness to engage with the broad-ranging and interdisciplinary themes and concepts involved in this dissertation, and her trust in giving me the academic freedom to pursue these topics. Her methodological suggestions and advice are testament to her knowledge and expertise as a researcher, and her understanding of the importance of exploring the vast landscape of algorithms not just from a technological, but a social perspective as well, are testament to her foresight as a scientist. Without her help and support, this dissertation would simply not have seen the light of day.

To the members of my proficiency and submission committee, I am thankful for their valuable feedback, their time and effort, and their epistemological curiosity in engaging with this inter- and trans-disciplinary dissertation.

To my colleagues, Fabian Fischer and Gabriel Grill, I would like to express my gratitude for their enthusiastic engagement with the design and implementation of the Critical Algorithm Studies seminar, and their tireless pursuit of the AMAS case study on so many different levels. I am sincerely grateful to Doris Allhutter and Astrid Mager for their willingness to collaborate on this case study as well, for their tireless fight against bureaucratic and political hurdles standing in the way of gaining access to the case study, but particularly for their patience and gentle, yet enlightening, introductions to Social Science and STS concepts previously foreign to me.

To my superiors at WIENFLUSS, Michael Stenitzer and Sascha Nemecek, I owe gratitude for their immediate and unquestioning support in undertaking the EnerCoach case study, and the affordances they have allowed me in order to pursue my studies and research. Furthermore, I would like to thank all participants in this case study; particularly, I would

like to express my gratitude to the late Herbert Mösch for his insightful contributions and patient recounting of the decades-long history of the EnerCoach program.

To my friends and family, who shared in both the burdens and joys of my pursuit of this work for so long, I cannot adequately express my gratitude. Both of my parents and my aunt and uncle, whose own work within and outside of the academic context I have always admired, instilled in me the fundamental belief in the value of science and research. My grandmother deserves gratitude for her stern, yet loving insistence on pursuing the highest forms of education, which always kept me going throughout the ups and downs of my studies and career. Without them, I would not have been able to complete this work. I am also thankful to all my friends, who endured both my rambling explanations of my work, and brought me back to reality at the same time.

Finally, but without question, foremost, to my wonderful wife, colleague, and partner in crime and science, Hayley Green: no words will ever be enough to express my gratitude. Without you, none of these efforts would have been possible, or indeed worthwhile. I will be eternally grateful for your expertise as a sociologist, your scientific curiosity, your incredible patience, and the many, many hours of spirited and insightful discussions we have shared on all aspects of this work. Most importantly, however, I will always be humbled and in awe of your kindness as a human being, towards me and everyone else around you. You inspire me to push forward and become a better scientist—and indeed, a better person—every day.

Of all the giant's shoulders I find myself standing upon, yours are, by far, the greatest.

Kurzfassung

Die fortschreitende digitale Transformation der Gesellschaft läutet ein - von manchen Beobachter:innen provokativ titulierte - "Zeitalter der Algorithmen" ein: eine Zeit, in der menschliches Handeln immer öfter von Algorithmen und algorithmischen Systemen mediiert, unterstützt, geregelt, bestimmt, strukturiert oder sogar ersetzt wird. Die Verheißungen dieser Technologien sind ebenso hochtrabend wie tiefgründig: nie dagewesene Effizienz und Geschwindigkeit, detaillierte Einblicke in komplexe Probleme durch die Analyse unfaßbar großer Datensätze, sowie mehr Fairness und Objektivität durch automatisierte Entscheidungsprozesse um Vorurteile und Diskriminierung durch menschlicher Entscheidungsträger zu reduzieren - dies sind nur einige der Vorteile, die diese Technologien mit sich bringen sollen.

Gleichzeitig führt die wachsende Zahl kritischer Problemstellen, die sich aus dem Einsatz algorithmischer Systeme ergeben, zu Forderungen nach einer besseren Accountability (Rechenschaftspflicht) und Transparenz algorithmischer Technologien. Im selben Maße, in dem wir diesen Systemen Macht geben und gleichzeitig unsere eigene Verantwortung delegieren, müssen wir jedoch auch neue Wege finden, diese Systeme und diejenigen, die sie einsetzen, für die Auswirkungen ihres Handelns zur Rechenschaft zu ziehen. Nur so können wir den verschiedenen Herausforderungen, die der Einsatz dieser oft komplexen, undurchsichtigen "Black-Box" Systeme mit sich bringt, bewältigen. Während in den Fachdiskursen der verschiedenen betroffenen akademischen Disziplinen - von Human-Computer Interaction (HCI) und Computer-Supported Cooperative Work (CSCW) in der Informatik bis hin zu Science and Technology Studies (STS), Politikwissenschaft und Governance Studies - weitgehend Einigkeit über diese Notwendigkeit herrscht, fehlt es noch immer an kohärenten Definitionen für algorithmische Accountability, algorithmische Transparenz, und einer holistischen Auseinandersetzung mit Fragen zum allgemeinen Verständnis Betroffener über algorithmische Systeme im Sinne einer algorithmischen "Literacy", sowie sinnvoller menschlicher Handlungsfähigkeit im Kontext dieser Systeme. Ebenso fehlt es an konkreten Anleitungen oder Frameworks, die algorithmische Accountability in verschiedenen Anwendungskontexten und für unterschiedliche algorithmische Technologien, einschließlich Artificial Intelligence (AI) oder Automated Decision-Making (ADM), unterstützen können.

Der zentrale Fokus dieser Dissertation zu algorithmischer Accountability und Transparenz, eingebettet im Kontext des interdisziplinären Forschungsfelds Critical Algorithm Studies

(CAS), liegt somit auf der Frage, wie wir uns diesen Herausforderungen stellen und diesen Problemstellungen begegnen können. Aufbauend auf den theoretischen Grundlagen und unterschiedlichen Definitionen der Termini “Algorithmus” und “Algorithmisches System”, den themenverwandten Problemstellungen zu *Bias*, *Diskriminierung* und *Transparenz*, sowie dem aktuellen Forschungsstand zu *öffentlicher Rechenschaftspflicht* (“Public Accountability”), bediene ich mich Bovens’ oft zitierter Definition für Accountability, übertrage dieses Konzept auf algorithmische Systeme und erarbeite im Zuge dessen eine neue, oft übergangene, Perspektive auf algorithmische Rechenschaft in Form von *prozeduraler Mikro-Accountability*. Diese theoretischen Überlegungen münden schließlich in der Analyse zweier konkreter Fallstudien algorithmischer Systeme: (1) das EnerCoach Energiebuchhaltungssystem, und (2) das Arbeitsmarkt-Assistenz-System (AMAS), das Profiling-System des österreichischen Arbeitsmarktservice (AMS).

Durch eine kontextuell situierte, (auto-)ethnographische Studie des *EnerCoach* Energiebuchhaltungssystemes, unter Zuhilfenahme qualitativer Interviews, Code-Reviews und anderer unterstützenden Datenquellen, werden die zentralen Herausforderungen in Bezug auf *systemweite Transparenz* und *ex-post Erklärbarkeit* herausgearbeitet. Die Ergebnisse dieser Analyse führen zu einer konkreten Intervention mittels Methoden des partizipativen Designs, in welcher Stakeholder des Systems im Zuge eines Co-Design-Prozesses konkrete Maßnahmen zur Verbesserung der vorher identifizierten Problemstellungen erarbeiten. Den Abschluß dieser Fallstudie bildet die Evaluierung der neu gestalteten Maßnahmen und eine Diskussion der Erkenntnisse, die aus der Analyse gewonnen werden konnten.

Für die Fallstudie des AMAS präsentiere ich die Ergebnisse eines kollaborativen Forschungsprojektes auf Basis einer qualitativen Dokumentenanalyse von mehr als 134 internen und öffentlich verfügbaren Schriftstücken mit Bezug zu diesem System. Nach einer sozio-technischen Beschreibung des Systems, der betroffenen Stakeholder und der organisatorischen Einbettung des Systems in Form einer ‘dichten Beschreibung’ nach Geertz präsentiere ich einen Überblick über die kritischen Problemstellungen zu Bias und Diskriminierung im System AMAS. Den Kern der Fallstudie macht der Fokus auf die Defizite des Systems in Bezug auf *systemweite Transparenz* und *ex-post Erklärbarkeit* aus, sowie die Anknüpfungspunkte und Auswirkungen dieser Defizite zu algorithmischer Accountability.

Durch die Synthese der so gewonnenen Erkenntnisse in Form einer komparativen Fallstudie entwickle ich schlußendlich das Algorithmic Accountability Agency Framework (A³ framework) als analytische Herangehensweise zur Strukturierung und Evaluation algorithmischer Accountability-Prozesse mit Hilfe einer Reihe von Leitfragen. Aufbauend auf Bandura’s *Sozio-Kognitiver Theorie* zu *emergenter interaktiver Handlungsfähigkeit* (“emergent interactive agency”) eröffnet das Framework eine neue Sichtweise auf Mikro- und Makro-Accountability-Prozesse mit einem speziellen Fokus auf die Ausprägungen und Einschränkungen menschlicher Handlungsfähigkeit in diesem Kontext. Mittels einer exemplarischen Anwendung des Frameworks auf die beiden Fallstudien illustriere ich die breite Anwendbarkeit und den Nutzen des Frameworks für die Evaluation und Bewertung algorithmischer Accountability-Prozesse, und demonstriere, welches Potential die

Anwendung des Frameworks für einen kritischen Diskurs zu - und die Gestaltung von - konkreten Maßnahmen zur Verbesserung von Accountability und Transparenz birgt.

Ziel dieser Dissertation ist es, sowohl theoretische, konzeptionelle als auch höchst praktische Beiträge zu dem noch jungen Forschungsfeld **CAS** zu leisten. Die Diskussion der fundamentalen theoretischen Grundlagen und die Konzeptionalisierung von algorithmischer Accountability als *wicked problem* soll zukünftigen Forschungsvorhaben eine belastbare Grundlage bieten, um die kritischen Problemstellungen algorithmischer Systeme aus einer holistischen, inter-disziplinären Sicht zu betrachten. Die Fallstudien belegen den Nutzen einer detaillierten qualitativen und quantitativen Analyse algorithmischer Systeme als komplexe sozio-technische Assemblagen, und können als konkrete Best-Practice Beispiele für die Anwendung partizipativer Methoden herangezogen werden. Zu guter Letzt stellt das **A³ framework** ein direkt anwendbares und praktisches Tool zur Evaluation und Analyse algorithmischer Accountability-Prozesse dar, und eröffnet neue Einblicke in die komplexen Verflechtungen zwischen algorithmischer Accountability, Transparenz, und menschlicher Handlungsfähigkeit in diesem “Zeitalter der Algorithmen”.

Abstract

The continuing digital transformation of society has given rise to what some scholars provocatively call *The Age of the Algorithm*: a time in which human endeavours are increasingly mediated, supported, regulated, determined, structured and even replaced by algorithms and algorithmic systems. The promises of algorithmic technologies are as lofty as they are profound: unprecedented efficiency and speed, intricate insights into complex problems with hitherto insurmountably large data sets, and improved fairness and objectivity of previously human decision-making processes burdened with personal bias and discrimination are just some of the benefits promised to us. At the same time, a growing number of critical issues arising from the use of algorithmic systems have led to calls for improved accountability and transparency of algorithmic technologies. As we delegate power to technology, we also must find new ways of holding those employing these systems to account for their conduct in order to face the various challenges presented by complex, opaque and black-boxed socio-technical assemblages. While the academic communities of the various related disciplines—from [Human-Computer Interaction \(HCI\)](#), [Computer-Supported Cooperative Work \(CSCW\)](#) in computer science, to [Science and Technology Studies \(STS\)](#), political science and governance studies—agree, by and large, on the importance of accountability, a coherent and agreed-upon definition of algorithmic accountability, transparency and the related issues of algorithmic literacy and meaningful human agency has yet to emerge. Likewise, concrete guidelines and frameworks to support algorithmic accountability across various application contexts and a wide range of technologies including, but not limited to, [Artificial Intelligence \(AI\)](#) and [Automated Decision-Making \(ADM\)](#), are still scarce.

To address these issues, I present this dissertation on algorithmic accountability and transparency situated within the emerging, inter-disciplinary field of [Critical Algorithm Studies \(CAS\)](#). Building on the theoretical foundations of the term *algorithm* from various perspectives, the related issues of *bias*, *discrimination* and *transparency*, and prior work on *public accountability*, I then appropriate and adapt Bovens' widely used definition of accountability for algorithmic systems, and introduce the notion of procedural *micro-accountability* as an important and often overlooked perspective. To apply these theoretical considerations, I subsequently present two case studies: (1) the EnerCoach energy accounting system, and (2) the [Arbeitsmarkt-Assistenz-System \(AMAS\)](#), an unemployment profiling system used by the [Public Employment Service Austria \(AMS\)](#).

Through a situated algorithmic (auto-)ethnography of the EnerCoach system based on qualitative interviews, code reviews and other auxiliary data sources, I identify crucial challenges related to *system-level transparency* and *ex-post explainability*. Following this analysis and employing an interventionist approach founded in participatory design methodologies, I describe how stakeholders co-designed concrete measures to address the previously identified issues and evaluate both the use of participatory approaches and the success of these measures.

For **AMAS**, I present the results of a collaborative research project founded in a qualitative document analysis of more than 134 internal and public documents of the **AMAS** system. After describing the system's socio-technical configuration, its stakeholders and organisational embedding within the **AMS** to arrive at a *thick description*, I summarize the critical issues of *bias* and *discrimination* as manifested by the system. The core of this case study analysis focuses on its lack of *system-level transparency* and *ex-post explainability* and the relation of these issues to algorithmic accountability.

By synthesizing the insights gained in the two case studies in the form of a comparative case study, I introduce the **Algorithmic Accountability Agency Framework (A³ framework)** as an *analytic lens* to structure accountability processes through set of *guiding questions*. Building on Bandura's *Social Cognitive Theory* of *emergent interactive agency*, the framework models both micro- and macro-accountability processes through the lens of human agency. In applying the framework to the two case studies, I then showcase its ability as a widely applicable toolset for both evaluation and assessment of algorithmic accountability processes, as well as its potential to support a critical discourse that encourages the ideation of concrete socio-technical measures to improve these processes.

With this dissertation I aim to contribute to the nascent field of critical algorithm studies in both a theoretical and conceptual, as well as a very practical manner. The summary of theoretical foundations leading to the conceptualization of algorithmic accountability as a *wicked problem* is meant to support future efforts in addressing this critical issue from a broad and inter-disciplinary perspective. The case studies showcase the value of in-depth, qualitative and quantitative analyses of algorithmic systems as complex socio-technical assemblages, and provide concrete best-practice examples for the use of participatory design methodologies to involve all stakeholders in addressing these critical issues. Finally, the **A³ framework** is a directly applicable, practical tool to evaluate algorithmic accountability processes, and further explicates the complicated relationship between algorithmic accountability, transparency, algorithmic literacy and human agency in this 'Age of the Algorithm'.

Associated publications

This dissertation is both founded on and expands upon a series of prior publications on the topics of *algorithmic transparency*, *literacy*, and *accountability*. While these prior publications present pertinent contributions to the research questions addressed in this dissertation, they are not directly integrated into this work, as would be the case for a cumulative doctoral dissertation. Instead, this dissertation is presented in the form of a *monography*, as it both transcends and expands on the content of these prior publications. To integrate the results of this prior work, the key points are replicated in the requisite chapters, albeit embedded in a much more detailed and comprehensive description of the subject matters.

The previous publications can be roughly categorized into four groups, pertaining to either (1) the EnerCoach case study described in Chapter 4, (2) the AMAS case study described in Chapter 5, (3) the comparative case study and A³ framework framework presented in Chapter 6, and (4) auxiliary publications that are tangentially relevant throughout the entire dissertation.

Disclosure of Authorship and Contribution

All publications for which I am not listed as the first author are the result of intensive and equal research efforts, collaboration and co-writing. Where possible, a declaration of equal contribution to this effect was included in these publications. For the publications on the AMAS case study system in particular, I led the research and writing efforts for the requisite sections most pertinent for this dissertation (i.e., *algorithmic transparency* and *accountability*).

With the exception of the German final report on the AMAS research project, all publications underwent a rigorous, double-blind peer review process; for the German research report, we were supported by a *scientific advisory board* of domain experts who provided feedback and non-blind peer review for the publication. The members of this advisory board included, in alphabetical order:

- Univ.-Prof. Dr. Bettina Berendt, Professor for Internet and Society at the TU Berlin and Director of the Weizenbaum Institute for a Networked Society

- Univ.-Prof. Dr. Iris Eisenberger, Professor Public Law and European Economic Law at the University of Graz
- ao.Univ.-Prof. Dr. Johanna Hofbauer, Associate Professor at the Institute for Sociology and Social Research, Department of Socioeconomics at the WU Vienna

List of Publications

The following is the itemized and categorized list of relevant publications.

EnerCoach Case Study

- [1] F. Cech, “*Tackling Algorithmic Transparency in Communal Energy Accounting through Participatory Design*,” in ACM Communities and Technologies Conference 2021, ser. C&T '21: Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech. ACM, 2021, pp. 258–268. Available: <https://doi.org/10.1145/3461564.3461577>
- [2] F. Cech, “*Beyond Transparency*,” in Companion of the 2020 ACM International Conference on Supporting Group Work, ser. GROUP '20: The 2020 ACM International Conference on Supporting Group Work, vol. 41. ACM, January 2020, pp. 11 – 14. Available: <https://doi.org/10.1145/3323994.3371015>

AMAS Case Study

- [3] D. Allhutter, F. Cech, F. Fischer, G. Grill, and A. Mager, “*Algorithmic profiling of job seekers in Austria: How austerity politics are made effective*,” *Frontiers in Big Data*, vol. 3, pp. 1 – 28, February 2020. Available: <https://doi.org/10.3389/fdata.2020.00005>
- [4] D. Allhutter, F. Cech, F. Fischer, G. Grill, and A. Mager, “*Der AMS-Algorithmus: Eine soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*,” ITA/ÖAW, Tech. Rep., 2020. Available: <https://doi.org/10.1553/ita-pb-2020-02>
- [5] B. Wagner, P. Lopez, F. Cech, G. Grill, and M.-T. Sekwenz, “*Der AMS-Algorithmus*.” *Zeitschrift für kritik - recht - gesellschaft*, no. 2, p. 191, 2020. Available: <https://doi.org/10.33196/juridikum202002019101>

Comparative Case Study and **A³ framework**

- [6] F. Cech, “*The Agency of the Forum: Mechanisms for Algorithmic Accountability through the Lens of Agency*,” *Journal of Responsible Technology*, vol. 7, p. 100015, 2021. Available: <https://doi.org/10.1016/j.jrt.2021.100015>

Auxiliary Publications

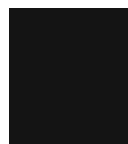
- [7] F. Cech, “Exploring emerging topics in social informatics: An online real-time tool for keyword co-occurrence analysis,” vol. 10540 LNCS, January 2017, pp. 527 – 536. Available: http://link.springer.com/10.1007/978-3-319-67256-4_42
- [8] F. Cech and M. Wagner, “Erollin’ on green: A case study on Eco-Feedback Tools for eMobility,” in *Proceedings of the 9th International Conference on Communities & Technologies - Transforming Communities*, ser. Communities and Technologies 2019, June 2019, pp. 121 – 125. [Online]. Available: <https://doi.org/10.1145/3328320.3328402>
- [9] S. Human and F. Cech, “A Human-Centric Perspective on Digital Consenting: The Case of GAFAM,” in *Human Centred Intelligent Systems*, A. Zimmermann, R. J. Howlett, and L. C. Jain, Eds. Singapore: Springer Singapore, 2021, pp. 139–159. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-5784-2_12
- [10] H. Tellioglu, M. Habiger, and F. Cech, “Infrastructures for sense making,” in *InfraHealth 2017: Proceedings of the 6th International Workshop on Infrastructures for Healthcare*. European Society for Socially Embedded Technologies (EUSSET), June 2017, pp. 1–4. [Online]. Available: <https://dl.eusset.eu/handle/20.500.12015/2908>

Contents

| | |
|---|-----------|
| Kurzfassung | ix |
| Abstract | xiii |
| Contents | xix |
| 1 Introduction | 1 |
| 1.1 Personal Motivation | 7 |
| 1.2 Research Context | 9 |
| 1.2.1 Critical Algorithm Studies | 9 |
| 1.2.2 The Need for Interdisciplinarity in CAS | 13 |
| 1.3 Research Questions | 14 |
| 1.4 A Critical Perspective on Criticising Algorithms | 17 |
| 1.5 Dissertation Structure & Chapter Overview | 18 |
| 2 On Algorithms | 23 |
| 2.1 Conceptualizing Algorithms | 23 |
| 2.1.1 A Brief History of Terminology | 24 |
| 2.1.2 Socio-Technical Systems | 25 |
| 2.1.3 Socio-Technical Assemblages, Actor-Networks and Co-Production | 28 |
| 2.1.4 Algorithms from a Functional Perspective | 32 |
| 2.2 Bias and Discrimination | 35 |
| 2.2.1 Classifying Bias in Algorithmic Systems | 37 |
| 2.3 Algorithmic Transparency | 38 |
| 2.3.1 System Transparency vs. Ex-Post Explainability | 39 |
| 2.3.2 Transparency Challenges | 40 |
| 2.3.3 The Problem with Machine Learning and Transparency | 42 |
| 2.4 Algorithmic Accountability | 47 |
| 2.4.1 Public Accountability: Actor, Forum and Account | 50 |
| 2.4.2 Accountability in the Context of Algorithmic Systems | 55 |
| 2.4.3 Macro-Accountability vs. Micro-Accountability | 70 |
| 2.4.4 Human Agency and the Accountability Process | 76 |
| 2.4.5 Accountability, Moral Responsibility and Computing | 80 |

| | | |
|----------|--|------------|
| 2.5 | The Wicked Nature of Accountability and Transparency | 88 |
| 2.6 | Chapter Summary | 89 |
| 2.7 | Chapter Conclusions | 91 |
| 3 | Methodology | 95 |
| 3.1 | Overall Approach | 95 |
| 3.1.1 | Case Study Selection | 96 |
| 3.2 | Case Study Methodologies: EnerCoach | 98 |
| 3.2.1 | Auto-ethnographic Considerations & Disclosure | 100 |
| 3.2.2 | Phase 1: Data Sources and Analytic Methodology | 104 |
| 3.2.3 | Phase 2: Intervention through Participatory Design Workshops | 111 |
| 3.3 | Case Study Methodologies: AMAS / AMS Algorithm | 113 |
| 3.3.1 | Practical Implementation | 115 |
| 3.4 | Comparative Case Study Methodologies | 121 |
| 3.4.1 | Application to CAS | 122 |
| 3.5 | Chapter Summary | 123 |
| 3.6 | Chapter Conclusions | 124 |
| 4 | Case Study: EnerCoach | 129 |
| 4.1 | Exploratory Vignette | 130 |
| 4.2 | Prior Research | 133 |
| 4.3 | Socio-Technical Description of the EnerCoach System | 136 |
| 4.3.1 | Stakeholder Analysis | 138 |
| 4.3.2 | Technical Implementation | 145 |
| 4.4 | Transparency | 158 |
| 4.4.1 | Requirements | 158 |
| 4.4.2 | Deficiencies | 163 |
| 4.4.3 | Underlying Reasons | 168 |
| 4.5 | Participatory Design Workshop | 172 |
| 4.5.1 | Part 1: Collaborative Exercise | 172 |
| 4.5.2 | Part 2: Concrete Measures | 181 |
| 4.5.3 | Post-Workshop Implementation and Evaluation | 183 |
| 4.6 | Chapter Summary | 185 |
| 4.7 | Chapter Conclusions | 186 |
| 5 | Case Study: AMAS - The AMS Algorithm | 191 |
| 5.1 | Exploratory Vignette | 192 |
| 5.2 | Prior Research | 195 |
| 5.3 | Socio-Technical Description of the AMAS System | 199 |
| 5.3.1 | History, Goals and Aim | 200 |
| 5.3.2 | Stakeholder Analysis | 202 |
| 5.3.3 | Technical Description | 205 |
| 5.3.4 | Operationalization in the Context of the AMS | 214 |
| 5.4 | Critical Issues: Bias, Discrimination, Transparency and Accountability | 220 |

| | | |
|----------|---|------------|
| 5.4.1 | Technical, Pre-Existing and Emergent Bias | 220 |
| 5.4.2 | Potential Discrimination of Jobseekers | 226 |
| 5.4.3 | System Transparency | 228 |
| 5.4.4 | Ex-Post Explainability | 231 |
| 5.5 | Chapter Summary | 234 |
| 5.6 | Chapter Conclusions | 236 |
| 6 | The A³ Framework | 241 |
| 6.1 | Comparability of Case Studies | 242 |
| 6.1.1 | Differences between the Case Studies | 242 |
| 6.1.2 | Similarities between the Case Studies | 244 |
| 6.1.3 | Applicability in the Context of CAS | 247 |
| 6.2 | The A ³ Framework | 250 |
| 6.2.1 | Preconditions and Assumptions | 251 |
| 6.2.2 | Procedural Accountability Model & Guiding Questions | 252 |
| 6.2.3 | Applying the Framework | 260 |
| 6.2.4 | Evaluation in Context with Other Frameworks | 276 |
| 6.3 | Chapter Summary | 288 |
| 6.4 | Chapter Conclusions | 289 |
| 7 | Conclusio | 293 |
| 7.1 | Summary Contributions | 295 |
| 7.1.1 | Theoretical Contributions | 295 |
| 7.1.2 | Practical Contributions | 303 |
| 7.2 | Limitations | 308 |
| 7.3 | Future Work | 312 |
| 7.4 | Final Remarks | 315 |
| A | Appendix | 317 |
| A.1 | EnerCoach Stakeholder Interview Guideline | 317 |
| A.2 | EnerCoach Reporting Sample Screenshots | 319 |
| A.3 | AMAS Case Study Document Index | 338 |
| A.4 | AMAS Case Study Explanation Texts | 349 |
| | List of Figures | 351 |
| | List of Tables | 353 |
| | Glossary | 355 |
| | Acronyms | 357 |
| | Bibliography | 361 |



Introduction

The ongoing *Digital Transformation* of society has profoundly impacted almost all aspects of our daily lives. How we work, how we communicate with each other, how we govern each other and are being governed, or how we seek out information and express ourselves is increasingly enabled and structured by, and dependent on, digital technologies. Technological innovation has shifted a significant portion of human behaviour and interactions into the realm of the digital, and has had transformative impacts on medicine, communication, and public administration, to name just a few examples.

At the core of many of these technologies lie principles of *automation*, enacted and manifested by *algorithms* and *algorithmic systems*. These algorithms have become so powerful and so ubiquitous that some scholars [11, 12, 13, 14, 15] have provocatively proclaimed the advent of *The Age of the Algorithm*. In this time, nearly all human endeavours are mediated, supported, regulated, determined, structured and even replaced by algorithms and algorithmic systems.

Algorithms decide which social media posts get attention and which disappear unseen. They determine which search results appear at the top of the page, and which ones slide into oblivion on page two. What information we see—and when we see it—as we navigate the digital sphere, is curated by algorithms. In the age of constant digital surveillance, every action we take online also feeds these algorithms new data, and determines the targeted advertisements we will see next. As annoying and worrying as these examples may be, these content curation systems, consumer recommender systems or social media scoring are only the most surface-level of algorithmic systems impacting our lives.

On a more fundamental level, algorithms also turn towards us humans as the primary subject of their attention. Algorithms are used to predict crime in the form of predictive policing or criminal recidivism risk assessment. They calculate our credit scores and our risk of being affected by health issues.

They predict our chances of finding a job, they decide which job adverts we see, and simultaneously assess how well we would do in the jobs we apply to—or even if we should be fired from the job we already have. If we try to think of a profession in which algorithms do not play a role (yet), we struggle to come up with examples. Case in point: When this dissertation is published, algorithmic systems will dutifully track and aggregate how often it is being downloaded and by whom it will be referenced. And yet another system will, based on this data, assess the *h-index* and other impact factors of my body of work, and determine the relative, quantifiable worth of my contributions to the academic world in terms of citations.

Even behind the scenes of modern civilization, algorithms make up the digital infrastructure that runs our energy, gas, water and sewage systems. They calculate the fastest route to take from A to B, they tell us which mode of transport to take, and they orchestrate public transport systems. They calculate the grotesque, optimized shapes of congressional districts in the United States of America to nudge the results of elections in one or the other direction.

Considering Benjamin Franklin’s famous quote “[...] *but in this world nothing can be said to be certain, except death and taxes.*” [16, p.410], one might hesitate to assert with *certainty* that there are no aspects of society left untouched by algorithmic systems. Such hesitancy, however, will promptly be contradicted by the fact that, indeed, algorithmic systems that claim to predict our life expectancy and analyse our tax returns already exist [17, 18].

This list of examples could be extended almost *ad infinitum*, given the ubiquity of algorithmic technologies, from the most visible, controversial or newsworthy systems to the most simple and, in their banality, often rather invisible ones. This permeation of society with algorithmic systems is hardly surprising when we consider the *promises*, as lofty as they are profound, made about the benefits of these various algorithmic systems. Algorithms supposedly deliver unprecedented efficiency and speed, insights into previously unsolvable problems based on the analysis of enormously large data sets, and improved fairness and objectivity of previously human decision-making processes prone to bias and discrimination are some of the most prevalent arguments for the benefits of automation. In reference to the latter promise, in many fields and professional contexts, algorithmic systems already augment our human decision-making processes, either by providing additional information or by suggesting specific choices. Where possible, they sometimes even take over entirely, relegating us to *humans-on-the-loop* that only verify what decisions the system otherwise makes.

These promises and potential benefits of algorithmic systems notwithstanding, a deeper look beneath the surface of the techno-utopian narratives dominating these *algorithmic imaginaries* [19, 20] reveals a darker side of algorithmic technologies. The fundamental *lack of transparency* of black-boxed, opaque or otherwise inscrutable systems is particularly problematic when, at the same time, these systems hold enormous power over humans, both as individuals, entire societies and anything in between. As much as these systems claim to know about us and the world, we often know surprisingly little about them,

starting with the simple awareness of their existence. The very same technologies promising to combat human prejudice and bias often not only replicate these very same sources of discrimination, but may even exacerbate and add new types of biases, and apply them at a scale and speed far beyond anything a single, biased human could ever achieve alone.

At the same time, this trend towards increased automation has profound impacts on human agency: as new systems codify processes formerly performed by human actors, they introduce new limitations of what we, as human actors, *can and can not do*. From automated decision-making to statistical profiling, from risk assessment to predictive policing: every new algorithmic application designed with the promise of giving us a more powerful and faster tool to achieve our goals also further erodes our own agency to question their conduct, overrule a decision, or adapt these automated processes according to our own individual needs.

This growing number of *critical issues* has led to a growing chorus of urgent calls for an improved *accountability* and *transparency* of algorithmic systems. As we delegate more power to algorithmic technologies, ensuring their fairness, scrutinizing their conduct, and judging whether or not they are deserving of our trust becomes an ever more pressing issue.

To do so, however, requires more than just applying our established, traditional mechanisms of accountability, be that legal, political or professional accountability. The intrinsic qualities of algorithms, the way they are co-produced, and how they are embedded in the world indicates that we must find new ways of holding these complex socio-technical assemblages to account for their conduct. While there is little dissent on this point within and across related academic disciplines including [Human-Computer Interaction \(HCI\)](#), [Computer-Supported Cooperative Work \(CSCW\)](#), [Science and Technology Studies \(STS\)](#), Political Science, or Governance Studies, a coherent, inter-disciplinary, and agreed-upon definition of *algorithmic accountability* and *algorithmic transparency* has yet to emerge.

Even more fundamentally problematic is the *terminological anxiety* [21] inherent in the cross-disciplinary discourse on *algorithms*: if we can not even agree upon what exactly an “algorithm” or “algorithmic system” is across disciplinary divides, how can we attempt to scrutinize them in all their socio-technical complexity? Similarly, the relation of these terms and concepts to the proximate issues of *algorithmic literacy* and *meaningful human agency* remains the subject of active debate and ongoing research efforts sprawled out across many scientific fields and disciplines.

Even more unclear than the problem spaces spanned by the need for *algorithmic accountability* and its related challenges are *practical and applicable solutions* to address them. Frameworks or guidelines on how to improve *algorithmic accountability* that can be applied across various application contexts and a wide range of technologies—including, but not limited to, [Artificial Intelligence \(AI\)](#) and [Automated Decision-Making \(ADM\)](#)—are still scarce, as are larger, overarching strategies to ensure a safe and human-centric development and deployment of algorithmic technologies.

Finally, as educators, academics and as society as a whole, we must develop a common understanding of the value of *algorithmic accountability* as a *virtue* if we aim to impress upon future generations the norms of an ethical, critical and reflective approach to the design and use of algorithmic systems.

Tackling Algorithmic Accountability

In this dissertation, I venture to address some of these issues from the perspective of the emerging, inter-disciplinary field of [Critical Algorithm Studies \(CAS\)](#).

From the descriptions given above, the following overarching question emerges:

“How can algorithmic systems, in all their heterogeneity, complexity and various application domains, be analysed, designed and improved to satisfy higher standards of accountability towards its stakeholders, affected humans and society at large?”

To approach this question, the first step must be to establish a common understanding of the definitions, conceptualizations and terminology surrounding algorithms and algorithmic systems to overcome the aforementioned *“terminological anxiety”* that is keeping us from pinpointing what exactly we refer to when we speak of *“algorithmic systems”*. To this end, this dissertation provides a thorough discussion of the various meanings and conceptualizations of the terms *algorithm* and *algorithmic system*, and suggests the use of theoretical models such as socio-technical systems, assemblage thinking and [Actor-Network Theory \(ANT\)](#) that allow us to consider algorithmic systems not just in terms of their technical components, but also their context of application, their human and non-human actors, as well as the immaterial aspects such as embedded norms and values.

Considering the uncertain relationship between *algorithmic accountability* on the one hand, and the challenges of *transparency*, *bias* and *discrimination* on the other, a structured analysis of the specific problems that the use of algorithmic systems introduce is a necessary precondition to tackle these problems through the inclusion of insight from the various relevant disciplines beyond computer science alone. Finding common ground on what it means for a system to be *transparent*, *opaque* or *biased* thus enables us to integrate inter-disciplinary approaches when analysing these issues. Providing structured taxonomies of the various issues subsumed under these terms also helps identify the most promising strategies to address them. Thus, this dissertation includes such a broad discussion of *algorithmic transparency* and *ex-post explainability*, as well as the challenges of *bias* and *discrimination* in algorithmic systems, and presents useful taxonomies to conceptualize them.

Building on these foundations, this dissertation makes use of prior work in *public accountability* [22, 23] to conceptualize *algorithmic accountability* as a relational, procedural interaction between a *forum* and an *actor*, in which the *actor* has an *obligation* to provide

an *account* and may face possible *consequences*. The translation of such cross-disciplinary definitions of accountability to the context of algorithmic system also requires *adaptation*, since the *fluid*, *complex* and *heterogeneous* nature of algorithmic systems also introduce new challenges related to determining aspects like cause, effect, agency, blame and responsibility for a given outcome.

As a consequence of the fast-paced nature of automation technologies compared to their non-machinic predecessors, traditional approaches to accountability processes are also woefully inadequate as a measure to prevent *immediate* harm. Since some algorithmic systems, as explicated in the introduction above, are capable of making potentially discriminatory decisions about millions of humans per second, accountability processes measured in months or years will simply not suffice to prevent harm. To address this issue of *immediacy*, this dissertation introduces the notion of *micro-accountability* as an important and overlooked perspective in which an *ad hoc* accountability process can be initiated by individual stakeholders affected by an algorithmic system in the moment.

To adequately conceptualize this process between an individual human as *actor* and, potentially, a *forum of one*, we must also consider which factors influence how individual humans can participate in such a process. Turning to established theories of human *agency*, Bandura's [24] **Social Cognitive Theory (SCT)** of human *emergent interactive agency* offers the most potential due to its inherent human-centricity and broad applicability. Given the rapidly increasing number of algorithmic systems, however, it is rather evident that even a comparably fast human-to-human *micro-accountability* processes will not be enough to achieve algorithmic accountability *at scale*. To offer a potential path forward, as I argue in this dissertation, we will have to accept that accountability processes might have to include a *non-human actor* as a compromise, albeit only in select cases. By considering the philosophical implications of detaching *moral responsibility* from *moral agency* for non-human actors, I thus introduce the concept of *artificial accountability* as a new form of accountability processes enacted between a human *forum* and a *non-human actor*.

As the final point in these theoretical considerations, it is important we learn more about the *nature* of the problems of *algorithmic accountability* and *transparency* to find the most promising avenues towards practical solutions. To this end, this dissertation characterizes them as *wicked problems*: intrinsically lacking a definitive formulation, they defy simple solutions and require iterative, non-linear and gradual approaches, and ultimately can never be solved conclusively, leaving us to strive for *satisficing* solutions as the best possible strategy to address them. This dissertation embraces this characterization both in its theoretical and practical implications.

Gaining Practical Insights

Having established the theoretical foundations of what algorithmic systems are, what challenges they introduce, and how we can characterize the various processes by which we might, in principle, hold them to account, answering the primary research question

then requires some practical insights into real-world algorithmic systems and their *wicked* challenges of transparency and accountability. Given the different conceptualizations of algorithms as socio-technical systems and socio-technical assemblages, it follows that any plausible analysis of algorithmic systems must fundamentally transcend the purely technical components of its assemblage. Thus, I embark on two in-depth, holistic and methodologically diverse investigations of two case studies, namely the EnerCoach energy accounting system and the [Arbeitsmarkt-Assistenz-System \(AMAS\)](#) unemployment profiling system created by the [Public Employment Service Austria \(AMS\)](#).

The first case study, EnerCoach, is a collaborative online energy accounting system currently in use by over 600 communities in Switzerland, whose sophisticated data entry and reporting functionalities aim to support communities in their fight against climate change and promote sustainable energy practices. Through a situated algorithmic (auto-)ethnography [21, 25, 26] of the *EnerCoach* system based on qualitative interviews, code reviews and other auxiliary data sources, I identify the system's crucial challenges related to *system-level transparency* and *ex-post explainability*. Addressing the challenges identified in the ethnographic approach, I then employ an interventionist approach founded in participatory design methodologies as an attempt to empower the stakeholders of the system itself to *co-design* concrete and appropriate measures to improve the system.

The second case study, [Arbeitsmarkt-Assistenz-System \(AMAS\)](#), is an example of a controversial statistical profiling system introduced by the [Public Employment Service Austria \(AMS\)](#), whose purpose it is to predict the chances of jobseekers on the labour market and classify them into three groups with a supposedly positive, neutral or negative outlook. For [AMAS](#), this dissertation includes the results of a collaborative research project founded in a *qualitative document analysis* of more than 134 internal and public documents related to the [AMAS](#) system. By describing the system's socio-technical configuration, its stakeholders and organisational embedding within the [AMS](#) to arrive at a *thick description* [27], the system's numerous critical issues, including challenges of *bias* and *discrimination* as manifested by the system, become evident. The core of this case study analysis focuses on its lack of *system-level transparency* and *ex-post explainability* and the impact of these issues on the system's *algorithmic accountability*.

Both systems analysed in these case studies share significant challenges in terms of transparency, ex-post explainability and accountability, but also represent some of the diversity of the algorithmic landscape through their differences in terms of application context, underlying technologies, operationalization and the potential impact on the stakeholders. From these productive tensions between those differences and similarities, however, we can synthesize important learnings derived from these two case studies in the form of a comparative case study.

In order to transfer these insights to the larger, heterogeneous landscape of algorithmic systems "*in the wild*", this dissertation responds to the need for accessible and versatile tools by introducing the [Algorithmic Accountability Agency Framework \(A³ framework\)](#); a generalized tool to evaluate and assess algorithmic accountability, offering a structured process through a set of guiding questions. Building on Bandura's [24] concept of

emergent interactive agency as an analytic lens, the framework models both *micro-* and *macro-accountability* processes through the lens of *human agency*.

To showcase its usefulness as a widely applicable tool for both evaluation and assessment of algorithmic accountability processes, this dissertation then applies the [Algorithmic Accountability Agency Framework \(A³ framework\)](#) to each case study. Besides the analytic capabilities of the tool, this showcase also illustrates its potential to support a critical discourse to encourage the ideation of *concrete socio-technical measures* to improve these accountability processes.

In summary, and to address the primary research question posed above, this dissertation provides the theoretical, terminological and conceptual foundations to analyse algorithmic systems as the complex, heterogeneous and diverse socio-technical assemblages that they are. These foundations offer us the necessary analytic tools to understand the complex and shifting relations of power and influence occurring between the various human, non-human and abstract or immaterial components of these systems. Furthermore, they allow for a structured approach towards the *wicked* problems of *algorithmic accountability* and *transparency*, and teach us the value of the *productive contradictions* between such cross-disciplinary concepts as moral responsibility, human *emergent interactive agency* and post-humanist theories like [Actor-Network Theory \(ANT\)](#) or assemblage thinking.

By applying these theoretical foundations towards two real-world examples of algorithmic systems in the form of the two case studies, this dissertation presents important insights into the value of stakeholders participation and its potential to elevate them to a *critical audience* [28]. It also highlights the consequences of *design-by-policy* and the concrete effects that deficiencies in *algorithmic transparency* and *accountability* have on the system's stakeholders and their *agency*. Furthermore, the case studies highlight the importance of directing our attention towards the seemingly *banal*, but nonetheless ubiquitous and equally impactful algorithmic systems instead of solely focusing on high-profile or controversial case studies and technologies.

Finally, from the synthesis of these practical insights, the [A³ framework](#) emerges as a versatile, practical and accessible tool to analyse and assess, but also to improve the design of *algorithmic accountability* processes. As such, it offers a path forward towards a world in which we can utilize the full potential of safe, accountable and transparent algorithmic systems by safeguarding our meaningful *human agency* in this “*Age of the Algorithm*”.

1.1 Personal Motivation

This dissertation and my overall interest in the topics of *algorithmic accountability* and its related issues is founded on three contributing factors. First and foremost, as a researcher and new faculty member at the TU Wien in late 2016, I was tasked with the establishing the *Centre for Informatics and Society* as an inter-disciplinary research and knowledge centre together with my advisor, Assoc. Prof. Dr. Hilda Tellioglu. Beyond

the administrative challenges in doing so, our conceptual work led me to realize the importance of a critical interrogation of algorithmic systems and related technologies given their ubiquity, scope and impact. At the same time, I was given the opportunity to collaborate with my colleagues, Fabian Fischer and Gabriel Grill, on the design and implementation of a university course focusing on some of the critical issues related to algorithms and algorithmic systems. This collaboration culminated in a course titled “*Critical Algorithm Studies*”, based loosely on a reading list¹ of the same name curated by Tarleton Gillespie and Nick Seaver [29]. In the following years, I grew painfully aware of the lack of consistent theoretical foundations for *algorithmic accountability* and *transparency*, as well as the absence of practical and widely applicable frameworks for addressing these issues. Trying to provide my students with current, balanced and understandable source material to study these topics presented a significant challenge, and further emphasized the gap between the various academic disciplines concerned with either algorithms or accountability and transparency. Additionally, my own research interests in **Computer-Supported Cooperative Work (CSCW)** and **Human-Computer Interaction (HCI)** often left me searching for guiding tools to (1) situate the artefacts we were creating for our projects within the larger academic context of *technology and society* and (2) evaluate their potential and impacts from a critical perspective, let alone provide concrete suggestions to improve them in terms of *accountability* and *transparency*.

The second contributing factor was my continued engagement with algorithmic technologies as part of my work in the private sector. As a software engineer and senior back-end developer at WIENFLUSS, a web design and software development firm focused on sustainable, accessible technologies, I became increasingly aware of the widening gap between the theoretical underpinnings of ethical software development and the requirements and limitations of practical software engineering in the private sector. The challenges presented by this dissociation were not the result of unwillingness or lack of awareness for ethical, sustainable development practices, but could be directly related to a lack of applicable frameworks and analytic tools to inform such practices. My work on the EnerCoach energy accounting system made these problems particularly tangible as I was trying to incorporate my academic expertise on issues of *transparency* and *ex-post explainability* into my professional engagement with real-world software projects.

Third, the announcement of the **AMAS** system as a profiling system for the unemployed prompted me to take a closer look at algorithmic systems as a form of administrative governance, and immediately revealed the potential for wide-ranging negative impacts in the form of bias and discrimination. In our critical analysis of the system, our newly formed research group incorporated many of the methodologies and concepts discussed in the field of **Critical Algorithm Studies (CAS)**. Yet, in terms of *transparency* and *accountability*, I felt we lacked the requisite tools to not only assess and evaluate this system, but also to help generate potential solutions to address this system’s shortcomings.

Reflecting on these three observations, I realized the need for theoretically well-founded, yet practical research into *algorithmic accountability* and *transparency*, that would also

¹The origins and impact of this reading list is discussed in detail in the next Section **1.2.1**.

generate real-world applicable tools and suggestions to improve the situation. Taking stock of the state of the art, it became clear to me that, in order to reach these goals, research into *algorithmic accountability* needed to be *critical* and *reflective*, *holistic* in its theoretical foundations, fundamentally *inter-disciplinary* in its methods and approaches, and finally, *practical* and related to real-world examples. Out of these observations and requirements, the concept and research focus for this dissertation was born.

1.2 Research Context

The following section ventures to give a broad overview of the research context this dissertation is situated in. Starting with an outline of the nascent field of [Critical Algorithm Studies \(CAS\)](#), the chapter introduces the most important disciplines, works and scholars of this field to set the stage and academic background for the research presented in this dissertation later on. Given the complex, inter- and trans-disciplinary nature of this research context and the methodologies employed in the case studies, the subsequent Section [1.2.2](#) provides arguments for the value of inter-disciplinary approaches when studying algorithmic systems, especially when considering the wicked problems [\[30\]](#) of *accountability* and *transparency*.

1.2.1 Critical Algorithm Studies

The research conducted as part of this dissertation is situated within the field of [Critical Algorithm Studies \(CAS\)](#). [CAS](#) finds one of its first mentions in a curated reading list published by the *Social Media Collective's* Tarleton Gillespie and Nick Seaver, with the intention to “[...] collect and categorize a growing literature on algorithms as social concerns.” [\[29\]](#). Spanning publications from the domains of Computer Science, [Science and Technology Studies \(STS\)](#), Sociology, Anthropology, Geography, Communication, Media Studies and Legal Studies, among others, they provide insights into a number of topics related to algorithms, including competing (historical) definitions of the term algorithm, implications of values and biases embedded within algorithmic systems and their underlying ideological world-views, human-algorithm interactions and various inter- and trans-disciplinary methodological approaches to studying algorithmic systems. [Figure 1.1](#) shows an overview of some particularly relevant disciplines represented² within Critical Algorithm Studies as a whole, and in the original reading list.

Despite the heterogeneous set of disciplines contributing to the field, a unifying aspect of the publications on Gillespie and Seaver’s reading list is their relationship to the discipline of computer science and its predecessors (which represent, arguably, the origin of the

²The disciplinary overview was compiled by sampling both the original reading list, the syllabus used in the *Critical Algorithm Studies* seminar taught by the author at the TU Wien, as well as from the literature referenced in this dissertation. It is neither meant to be a definitive literature survey of [CAS](#), nor is it a scientifically collated, representative sample; rather, it simply highlights the diversity of disciplines contributing to the field, and is meant to help situate this dissertation within the larger context of [CAS](#).

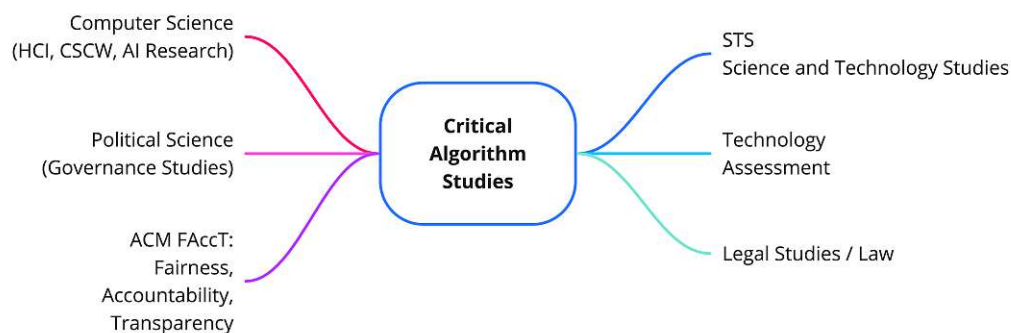


Figure 1.1: Critical Algorithm Studies and related disciplines and fields.

term *algorithm*). Most of the venues, conferences and journals these publications are attributed to could be described as *adjacent* to computer science, or at best as *overlapping*. With few exceptions (e.g., seminal papers on defining terminology [31] or introducing cross-disciplinary concepts into computer science itself [32]), the majority of publications are discussing algorithms from the standpoint of their own disciplines, and engage with computer science literature on algorithms only in a limited capacity. While some of the larger publication venues in human-computer-interaction—such as the *Conference on Human Factors in Computing Systems (CHI)*—do feature in the list ([33, 34]), they focus on specific examples of algorithmic systems from the standpoint of HCI or, more rarely, CSCW.

While the list provides a great starting point for further exploration of the topics at hand, it is by no means to be seen as a comprehensive literature review or even representational for the current state of CAS. The number of contributions to the field has grown substantially since the list was last updated in 2016, as Figure 1.2 illustrates, even only for those that mention “Critical Algorithm Studies” by name. These limitations notwithstanding, the list has since been widely circulated, inspired other similarly curated lists and found its way into a number of Computer Science and STS curricula. The list also served as the original inspiration for the seminar on “*Critical Algorithm Studies*” taught by myself at the TU Wien.

To take a more recent look at the state of CAS, and to illustrate the diverse nature of the contributions to the field, the ACM Conference on Fairness, Accountability and Transparency in Algorithmic Systems (FAccT)³ (formerly ACM FAT*) lends itself as a premier example: Key topics of the 2021 ACM FAccT conference include areas as diverse as *Algorithm Development, Data and Algorithm Evaluation, Human Factors, Privacy and Security, Humanistic theory and critique* and *Social and organizational processes*, to name a few [36]. The steadily rising number of submissions (from 162 in 2019, to 290 in 2020 and 328 in 2021) underscores the growing relevance of the conference as much as the relevance of the field in general. While the published papers of recent iterations of the

³See <https://facctconference.org>

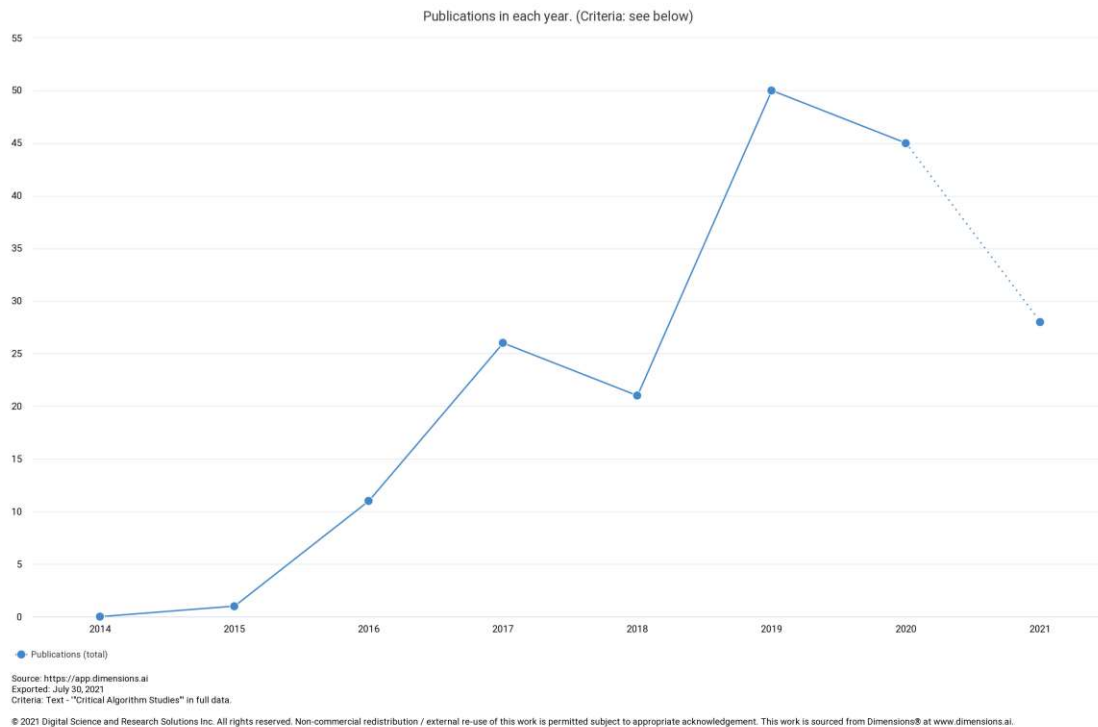


Figure 1.2: Number of publications with keyword “Critical Algorithm Studies” on the analytics platform Dimension.ai [35]

conference show diversity in the variety of key topics and areas that **CAS** is concerned with, the conference also features a strong focus on algorithms under the umbrella term of **Artificial Intelligence (AI)** or **Machine Learning (ML)**, often abbreviated into an amalgamated term as **Artificial Intelligence/Machine Learning (AI/ML)**. This leads to an interesting schism within the **FACcT** community itself: the focus on critical issues in **AI/ML**—particularly bias and discrimination—manifests into two disparate approaches. On the one hand, a number of contributions from computer science scholars engage with these problems through technical solutions such as bias mitigation in data sets [37] or the development of ‘fair’ algorithms used in the field [38]. On the other hand, many **FACcT** publications show a keen awareness of the limitations of a techno-deterministic approach, or, as Raji et al. put it, the “*technocratic paradigm [that] dominate[s] computer science*”, and lament this approach as “[...] *problematic in its ignorance and dismissal of sociological or critical theory.*” [39, p.517]. Consequently, and similar to the **CAS** reading list by Seaver and Gillespie [29], a significant portion of the field as represented in the **FACcT** conference favours inter-disciplinary approaches and embraces methodologies from the Social Sciences and Humanities.

The tension between these two approaches to critically engaging with algorithmic systems and their issues—one stemming from the Social Sciences and Humanities, and the other

originating from within subfields of Computer Science—parallels the tension between CAS and Critical Data Studies (CDS). Prominent CDS scholars like Cathy O’Neil, herself the author of the widely cited book “*Weapons of Math Destruction*” [40], bemoan the lack of a “*distinct field of academic study that takes seriously the responsibility of understanding and critiquing the role of technology — and specifically, the algorithms that are responsible for so many decisions — in our lives.*” [41]. As Moats and Seaver [42] argue, this claim may seem questionable, particularly considering the growing body of work either directly attributable to CAS or at least tangentially related to the study of algorithms and their impact on society. Furthermore, the decades-long history of research in the field of STS [42, p.2], including work on algorithms and algorithmic systems, presents an undeniable counter-argument to this claim. Moats and Seaver interpret O’Neil’s arguments as *boundary work* [43] to distinguish the ‘real’ critical work in algorithms—as situated within the more technical realm of CDS—from the “*more fundamental questions about the enterprise of algorithmic knowledge or modes of decision making*” [42, p.2] being investigated from the perspective of other academic disciplines.

These inner- and inter-disciplinary tensions notwithstanding, the contributions to the ACM FAccT conference and journals such as *Big Data & Society*⁴, *New Media & Society*⁵, *Science, Technology, & Human Values*⁶ or *Information, Communication & Society*⁷ showcase the value of taking a *critical* approach to studying, evaluating and designing algorithms—indicated by the implicit nod towards *Critical Theory* through the name of the field and paralleling academic movements such as *Critical Race Theory* or *Critical Geography*. In the tradition of the Frankfurt School representatives Max Horkheimer, Theodor Adorno and Herbert Marcuse, the overall corpus of work in Critical Algorithm Studies exemplifies Horkheimer’s requirements for a critical theory to be “*explanatory, practical, and normative, all at the same time.*” [44, p.1]. Thus, CAS strive to analyse and explain the various functionalities, underlying values, and power hegemonies pervading algorithmic systems while remaining committed to taking a strong normative stance on practical issues relating to the design, utilization and evaluation of algorithmic systems and applications.

In summary, CAS covers the study of all algorithmic systems, from criminal risk assessment and profiling [45], profiling of jobseekers [3], surveillance technologies [46], energy accounting systems [1] or credit scoring [47], to name just a few examples. In this spirit, this dissertation adopts the subject matters, terminologies and methodologies of CAS to advance our shared understanding of a variety of algorithmic systems and their related issues.

⁴See <https://journals.sagepub.com/home/bds>

⁵See <https://journals.sagepub.com/home/nms>

⁶See <https://journals.sagepub.com/home/sth>

⁷See <https://www.tandfonline.com/toc/rics20/current>

1.2.2 The Need for Interdisciplinarity in CAS

Given the not-so-recent omnipresence of computer software and its applications throughout many parts of human societies, it is surprising that the underlying algorithms have only comparatively recently been the focus of inter-disciplinary inquiry from the point of view of academic disciplines beyond Computer Science, Engineering or Mathematics. As pointed out by MacDonald [48], one of the first proposals to study algorithmic systems (i.e., early social media) from a social science perspective can be traced back to Lev Manovich in his book *The Language of New Media* in 2001, where he argues:

“To understand the logic of new media, we need to turn to computer science. It is there that we may expect to find the new terms, categories, and operations that characterize media that became programmable. From media studies, we move to something that can be called “software studies” — from media theory to software theory.”

[49, p.48]

While limited in its original scope due to the focus on *new media* before the rise of *social media*, this proposal already implies the necessity of an inter-disciplinary approach to the inquiry into the logic and impacts of algorithms—the approach that many of the more recent publications in the field of CAS have taken. Manovich’s approach also exemplifies one of the main foci of the field, namely to study the influence algorithms have on modern culture and society. MacDonald [48] mentions, among others, multiple studies into a competition that the streaming service *Netflix* ran to improve its recommendation algorithm between 2006 and 2009, and investigations of the cultural influence Google’s PageRank algorithm had on the way websites are being optimised for maximum impact [50].

As algorithms gained more attention from other disciplines, including Social Sciences and Humanities (e.g., STS, Political Science or Legal Studies), the challenges of studying these systems became more and more clear. Neither a purely technical approach nor an approach solely grounded in the Social Sciences can comprehensively account for the myriad of effects, impacts, interconnections, and relations of power introduced by these new technologies. Even more concerning, such a comprehensive understanding is one of the necessary foundations upon which “*legitimate and effective algorithmic governance*” [51, p.1] must rest—a governance not only *by*, but also *of* algorithms, is sorely needed as we see algorithmic technologies “[...] *influence, shape and guide our behaviour and the governance of our societies*” [51, p.1], as Danaher et al. put it so succinctly. Not only are we collecting ever more data on a growing number of electronic devices as the internet of things is permeating our world, but producing actionable knowledge from such data to support humans in fighting diseases such as the COVID-19 pandemic [52] or fighting the climate catastrophe [53] requires ever more sophisticated algorithmic systems. As these technologies mature and evolve from *informing* to *supporting* or *automating decision-making*, we also see a shift from human-curated decision-making support systems

towards fully automated, algorithmically curated decision-making systems [54]—a trend that some scholars have provocatively dubbed “*algocracy*” (e.g., [55, 56, 57, 51]).

In a workshop on algorithmic governance, Danaher et al. [51] developed the, to date, most comprehensive research agenda based on 12 major categories of barriers hindering this effective and legitimate algorithmic governance. Beyond those barriers—among them *opacity of algorithms*, *capacity/knowledge* among public servants, technologists and lawyers/legal systems, *privacy and informed consent*, *ethical awareness* and *technological uncertainty*—the participants identified one additional, separate meta-barrier: the challenge of *interdisciplinarity*. While all participants agreed on the necessity for inter-disciplinary cooperation in tackling these complex research questions and challenges, they also noted a prevalent culture of dismissing different perspectives in the fields related to CAS, exacerbating the knowledge gap between technologists and non-technologists, and leading to the development of algorithmic systems that do not satisfy legal and ethical standards [51, p.14]. To counter these gaps, the workshop participants suggested a range of research methods conducive to inter- and trans-disciplinary cooperation and coordination: first and foremost (inter-disciplinary) case studies, but also more participatory methodologies such as action research, or ethnographic studies and document analysis [51, p.10], as well as “[...] *the use of actor-network theory to better understand institutional and legal complexity* [...]” [51, p.15].

Following their recommendations, both of the case studies presented in this dissertation are thoroughly committed to this inter-disciplinary approach, making use of many of the suggested methodologies and theoretical approaches. Chapter 3 provides a structured and detailed description of the methodologies utilized, and the following Section 1.3 detailing the research foci of this dissertation also exhibits many of the topics and themes within and adjacent to the research agenda originally put forth by Danaher et al. [51].

1.3 Research Questions

This dissertation addresses the timely and pressing issues of algorithmic transparency, algorithmic accountability and human agency. At its core, it follows the credo of the TU Wien—“*Technology for People*”—in its human-centric approach to answering research questions within and between the realms of *Technology* and *Society*.

In particular, the research presented in this work answers the following *primary research question*:

Primary Research Question *How can algorithmic systems, in all their heterogeneity, complexity and various application domains, be analysed, designed and improved to satisfy higher standards of accountability towards its stakeholders, affected humans and society at large?*

Subsequently, this overarching primary research question is to be answered through the following secondary and supplemental research questions:

SRQ1 *What are different conceptualisations of algorithms, algorithmic transparency and algorithmic accountability?*

The field of **CAS** remains fraught with *terminological anxiety*, as Seaver [21] so pointedly put it: a wealth of competing, sometimes overlapping or even contradictory definitions for the subject matters at hand—from *algorithm* or *algorithmic system* itself to the issues of *algorithmic transparency* and *algorithmic accountability*—exists and makes a holistic view of these matters challenging. Any research situated within and across the boundaries of these subjects is inevitably at risk of being, at best, misinterpreted or, in the worst case, dismissed by those venturing outside the confines of their own specialized disciplines. This is particularly troublesome given the stated need for inter-disciplinary engagement between researchers in natural and technical sciences (such as computer science or informatics) on the one hand, and social scientists on the other. If we cannot establish a common ground of understanding of the subjects we research, inter-disciplinary cooperation and reaping the reciprocal benefits of such research are doomed to fail. To investigate the issues at the core of this dissertation, it is consequently paramount to establish such a common understanding of the terms, methods and paradigms utilized as part of its strategy of inquiry. To date, coherent conceptualizations of algorithmic systems that include and transcend the disciplinary boundaries of computer science and its sub-disciplines of **HCI** and **CSCW** on the one hand, and **STS** on the other, are scarce. Thus, the first step in approaching the *wicked problems* [30] of *algorithmic accountability* and *transparency* must be to establish such a common understanding. Based on a the terminological history of the term *algorithm*, a critical literature review of conceptualizations of technology and its interplay with society, **SRQ1** aims to provide answers to the questions

- **SRQ1.1** *What are established definitions of accountability and transparency?*
- **SRQ1.2** *What challenges fuel the wicked nature of these issues?*
- **SRQ1.3** *How must these definitions be adapted to become applicable to algorithmic systems?*
- **SRQ1.4** *How are algorithmic transparency and algorithmic accountability related to each other?*
- **SRQ1.5** *What roles do human and non-human agency play in these issues?*

and finally,

- **SRQ1.6** *Which conceptual and theoretical paradigms are most suited to the analysis and improvement of algorithmic systems in regards to their accountability and transparency?*

SRQ2 *What measures can be taken to improve the transparency and accountability of algorithmic systems?*

Having established a common understanding of terminology and core issues related to accountability and transparency of algorithmic systems, the need for an improved accountability of these systems in light of (1) their widespread use, (2) their growing societal impact, and (3) increasingly ubiquitous nature in almost all areas of society becomes obvious. To provide actionable results to address these issues, this dissertation aims to investigate and evaluate different approaches to increase system transparency, ex-post explainability and accountability of two exemplary algorithmic systems in the form of two case studies. Drawing on a variety of methodologies, including *technical analysis*, *situated ethnography*, *document analysis* and methods of *qualitative inquiry*, the case studies provide insights to answer **SRQ2** and its subsequent questions

- **SRQ2.1** *What design methodologies support or hinder the accountability and transparency of algorithmic systems?*
- **SRQ2.2** *What actions can system stakeholders take to improve the accountability of their systems?*
- **SRQ2.3** *What technical, socio-technical and procedural measures can be taken to support transparency and accountability processes when interacting with the socio-technical assemblage of an algorithmic system?*

SRQ3 *What guidelines can be developed to improve algorithmic accountability for future algorithmic systems?*

Finally, to generalize the results yielded by the empirical work on the case studies in SRQ2, learnings can be derived that synthesize into guidelines and recommendations for the future development of both algorithmic technologies in an academic research context, as well as concrete algorithmic systems for real-world applications. Answering **SRQ3** involves gaining insights on the subsequent research questions

- **SRQ3.1** *How can accountability requirements be formulated and adapted based on the application context?*
- **SRQ3.2** *What analytic lenses can help evaluate hindering factors to successful accountability processes?*

and lastly

- **SRQ3.3** *How can a guiding framework be designed to suggest material solutions to improve the accountability of a broad range of algorithmic systems across different domains of application?*

1.4 A Critical Perspective on Criticising Algorithms

Considering the prevalence of the word “*critical*” both within this dissertation, throughout related research, and, indeed, in the very name of the field this work is attributed to, an unsurprising yet common misconception persists that the approach exhibited by **CAS** embodies a fundamental techno-pessimism or even techno-phobic stances. While understandable, it is important to reject this notion, both as a disclaimer for this dissertation and the larger field. A critical approach to algorithms entails a nuanced reflection on the positive and negative aspects of a given technological assemblage and the power structures it is both embedded in and contributes to. Neither should such an analysis overemphasize the dangers, issues and challenges that arise from the use of technology, nor should it ignore them in favour of evangelizing the purported benefits of said technology. However, a balanced perspective does not necessarily mean that a given technology, algorithmic system or application must be described through an equal set of positive and negative aspects: there are certainly examples abound where the lofty claims of proponents of a system have been proven to be of little substance or downright baseless, and where critical voices have rightfully focused on the problematic aspects. To distinguish between a well-founded critique from an imbalanced, overly critical and biased perspective can, indeed, sometimes be a difficult challenge—hence to common stereotype that all research done in **CAS** (or other fields that contain the word ‘critical’) is simply the manifestation of a politically motivated agenda of academic ‘spoilsports’ seeking to discredit the hard work of other scientists. In its most extreme forms, the word *critical* alone may trigger hate speech and aggression, as recently exhibited by the public controversy around *Critical Race Theory* [58, 59] fuelled by right-wing media and conspiracy theorists in the United States of America.

To clarify my position and the goals of this dissertation, and to avoid any misunderstandings, I want to press the fact that neither the theoretical foundations, the methodologies, the case studies and their results, nor the synthesis and conclusions drawn in this work are based on the assumption that algorithmic technologies, in their myriad of configurations, are categorically ‘bad’, ‘dangerous’, or ‘problematic’—or, in the contrary, per definition ‘good’, ‘safe’ or ‘harmless’. Any impression that this dissertation focuses too strongly on criticism of technology in general or algorithmic systems in particular is solely the result of limited space and time to list the various undisputed benefits that modern technology has provided us with. There is no doubt that algorithmic technologies, including the often controversially discussed examples of **AI/ML**, can make important contributions to global issues and challenges, such as the fight against poverty and inequality, efforts to mitigate the looming climate catastrophe, or battling disease and human suffering.

In the end, this dissertation also embodies my personal conviction that algorithmic technologies are not bad *per se*, but carry risks and dangers that, at this time, significantly limit their potential for good. We are still failing to alleviate these risks in a satisfactory manner, and lack the tools and understanding to address the resulting issues in a scientific and systematic way. To reach our goal of realizing the true potential of these technologies, we must point out their shortcomings as clearly and precisely as possible. After all, if we

do not take an unflinching look at technology’s problematic aspects, how could we hope to improve it?

1.5 Dissertation Structure & Chapter Overview

This dissertation addresses the previously outlined research questions and topics in a structured manner illustrated through the flow chart depicted in Figure 1.3. In this chart, the interdependencies between the chapters are represented through bold arrows, illustrating how they build upon each other. Furthermore, the chart visualizes the contributions the various chapters make to answering both the *primary* and *secondary* research questions through the dotted arrows, linking chapters to either specific SRQs or the overall, primary research question as a whole.

With the exception of this introductory chapter and the concluding final Chapter 7, each of the following chapters ends with both a chapter summary and chapter conclusions. Conceptually, the summaries offer a more detailed descriptive overview of the chapter content, while the conclusions provide larger analytical insights gained in the chapter and correlate them to insights from the previous chapters as well.

Chapter 2, “On Algorithms”, discusses the various competing conceptualizations of the terms *algorithm* and *algorithmic system*, including its historical and technical roots, socio-technical systems, co-production and assemblage thinking approaches, as well as a functional perspective on algorithms. The following sections outline the related issues of *bias* and *discrimination* as well as describe different readings of and challenges related to *algorithmic transparency*. At the core of the chapter lies the discussion of *algorithmic accountability* based on Bovens’ work on public accountability, introducing the relational concepts of *forum*, *actor*, *obligation* and *account*. In this chapter, I also introduce the perspective of *micro-accountability* to expand on Bovens’ definitions largely focused on *macro-accountability*. Considering the role of humans in these micro-accountability processes, I introduce Bandura’s concept of *emergent human agency* and situate it within the other theoretical foundations of this dissertation. Next, in a short excursion, I discuss the notion of non-human actors and the philosophical implications this notion presents for *moral responsibility* and *moral agency* in computing, and introduce the concept of *artificial accountability* as an extension of the taxonomy presented by Bovens to include non-human actors. Finally, I characterize the problems of *accountability* and *transparency* as members of a class of *wicked* problems, and draw conclusions for research approaches and methodologies based on this classification. Chapter 2 directly addresses **SRQ1** and its sub-questions, and also provides theoretical insights for the conclusions drawn in Chapter 7.

Chapter 3 starts with an outline of the overall methodological approach of this dissertation, including the guiding principles that informed the selection of methodologies and case studies. Following this, I detail the specific methods of inquiry and their implementations that were used in the two case studies and the synthesis of the **A³ framework**. For the EnerCoach case study, I first disclose and discuss the auto-ethnographic nature of

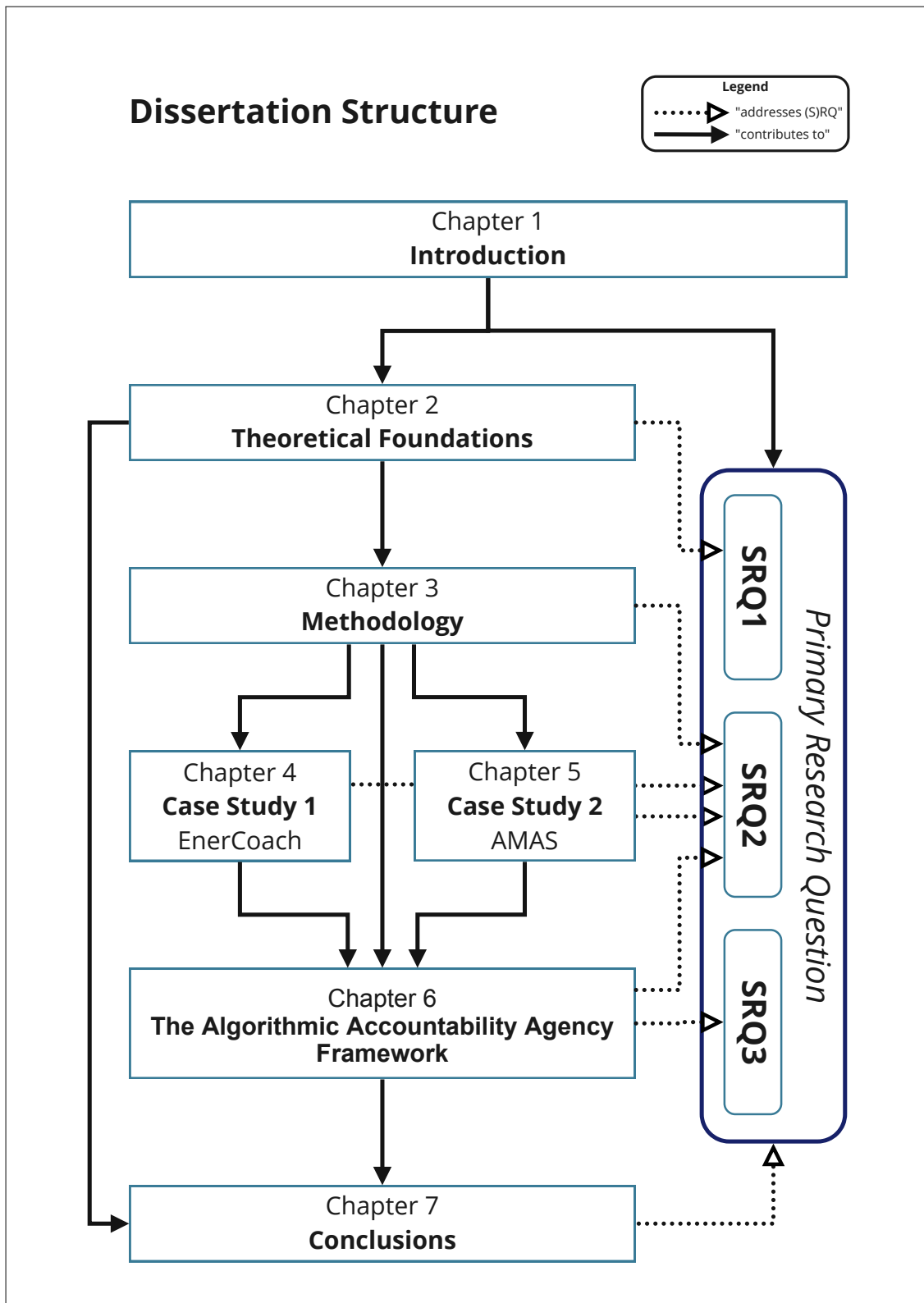


Figure 1.3: This chart illustrates the flow of chapters and their contributions to the primary and secondary research questions.

the study and describe the strategies to ensure the validity of its results, followed by a description of the analytic and interventionist approaches used in the two phases of the case study. For [AMAS](#), I disclose my collaborators in the original research project and their contributions, describe the overall methodological considerations and their practical implementation in the form of a document analysis. Finally, I present the theoretical foundations for the approach of the comparative case study that led to the development of the [A³ framework](#), and make supporting arguments for the choice of this method. Chapter [3](#) also contributes to the answers to **SRQ2** insofar as it describes the interventionist, participatory design methodologies used to develop accountability and transparency measures for the EnerCoach case study.

Chapter [4](#) introduces the first case study of the EnerCoach system. After an initial *composite exploratory vignette* [\[60\]](#) to provide some background on the case, I then provide an overview of domain-specific prior research on energy accounting and civil technologies. At the core of this chapter lies the socio-technical description of the EnerCoach system, including a detailed stakeholder analysis and a description of its technical implementation. Following this, I discuss the analytic results of the case study by outlining the transparency requirements, deficiencies and their underlying reasons of the system. I then report on the results of the interventionist part of the study in the form of a *participatory design workshop* and evaluate the results. As such, this chapter directly contributes to the answers to **SRQ2** and its sub-questions.

Chapter [5](#) presents the analysis of the second case study, the [AMAS](#) system. Analogous to Chapter [3](#), I make use of an exploratory vignette to outline my interest and engagement with the system and provide background information situating the case study contextually and temporally. Following this vignette, a socio-technical description of the [AMAS](#) system discusses the system in terms of its history, goals and aim, presents a stakeholder analysis and attempts a technical reconstruction of its inner workings as best as possible based on the available documents. As the final part of this description, I outline the planned operationalization of the [AMAS](#) system as part of the consultation process between jobseekers and caseworkers. The second part of this chapter focuses on a critical description of the system in relation to the issues of bias, discrimination, system-level transparency and ex-post explainability and the implications of these issues for the system's requirements towards accountability. Like the previous chapter, Chapter [5](#) also directly contributes to the answers to **SRQ2** and its sub-questions.

Chapter [6](#) presents the results of the comparative case study synthesizing the learnings from two case studies into the [A³ framework](#). Based on a detailed description of similarities and differences of the two cases, I present arguments for why a comparison of these two cases is worthwhile and promising in the context of algorithmic accountability, and reflect on their applicability to the larger context of [CAS](#) as a whole. Having established these rationales, I then introduce the [A³ framework](#) and define the necessary assumptions and preconditions underlying its use and application. I then apply the framework to each case study in the form of two scenarios for prototypical accountability processes, to support my arguments for its wide applicability and showcase its potential as an assessment

tool. Closing the chapter, I evaluate the framework in context with previous frameworks for *algorithmic accountability*, situate it adjacent to established models in [HCI](#) such as the *Human-Artefact Model*, and discuss its applicability as part of *algorithm audits* and [Algorithmic Impact Assessments \(AIAs\)](#). Chapter [6](#) thus contributes additional insights for **SRQ2**, but most importantly, directly answers **SRQ3**.

In *Chapter 7*, I provide summary remarks and synthesize overall learnings derived from the previous chapters. In particular, I explicate the theoretical contributions this dissertation makes for the conceptualization of algorithmic systems, the methodological implications derived from the two case studies, as well as the conceptual and practical contribution represented by the [A³ framework](#). Finally, I disclose any limitations of this dissertation and its constituent parts, and suggest further topics for research and an outlook towards future lines of inquiry. As a conclusion, this chapter addresses all secondary research questions and directly summarizes the answers to the primary research question that this dissertation provides.

On Algorithms

The terms ‘algorithm’ and ‘algorithmic systems’ cover a wide variety of meanings, and find different use depending on their context. To develop a coherent and critical analysis of algorithmic systems and the issues they raise (including, but not limited to *algorithmic accountability* and *algorithmic transparency*), this chapter discusses various ways in which algorithms can be conceptualized, based on a history of the term algorithm itself and leaning on theoretical [STS](#) concepts as well.

The subsequent sections then put the spotlight on the most critical issues relating to algorithms, their use as well as their impact on humans, culture and society at large: bias, discrimination, fairness and equality; algorithmic transparency and finally, algorithmic accountability itself.

2.1 Conceptualizing Algorithms

As Seaver argues in his excellent article “Algorithms as Culture” [\[21\]](#), the term *algorithm* is fraught with overlapping and competing definitions, leading to what he identifies as “terminological anxiety” within [CAS](#). From highly technical definitions within computer science and its predecessors to broader, more practical definitions spanning multiple disciplines, “algorithm” takes on a variety of contextual meanings and defies simple definitions. Nonetheless: a common understanding of the socio-technical artefacts with which [CAS](#) concerns itself will be necessary to contextualize both the methodologies of inquiry used in this dissertation as well as the results produced by the studies.

To this end, this section provides an introduction to the different conceptualizations of *algorithms* and *algorithmic systems*. Starting with a brief history of the term algorithm from a more technical perspective, the following sections frame algorithms and algorithmic systems through different lenses and disciplines, from algorithms as socio-technical

systems, socio-technical assemblages and [ANT](#), to algorithms viewed from a functional perspective.

2.1.1 A Brief History of Terminology

The term *algorithm* most likely traces back to the name of the Persian scholar and mathematician al-Khwārizmī (*780 - †850), whose contributions to algebra¹, arithmetic and trigonometry are foundational to mathematics [\[62, 63\]](#). However influential to mathematics and the term *algorithm*, his life and works predate the first computers by more than a millennium [\[64\]](#). While various translations (and bastardizations) introduced the term *algorithm* during the middle ages and enlightenment periods, it did not reach wider popularity until the second quarter of the 20th century and the corresponding advent of early computing. Hilbert’s Entscheidungsproblem [\[65\]](#) and Alonzo Church’s and Alan Turing’s (coinciding) publications [\[66, 67\]](#) of what would become known as the *Church-Turing-Thesis* refer to mathematical concepts that are—within the field of computer science—widely understood as the theoretical underpinning of modern algorithms and, indeed, modern computer science. As Gurevich, however, points out, the wide-spread notion that the *Church-Turing-Thesis* and subsequently, the overwhelming evidence produced in its favour, settled the definition of what an algorithm is, is incorrect, since an algorithm is “[...] *much, much more than the function it computes.*” [\[68, p.1\]](#). But even Gurevich stays within the narrow confines of logic and computer science, limiting his arguments to Turing machines, Kolmogorov (pointer) machines, abstract sequential state machines (ASMs) or parallel and distributed algorithms. He does, however, make one distinctive point: Even within logic and theoretical computer science, there are algorithms “[...] not covered directly by Turing’s analysis” [\[68, p.8\]](#), specifically *interactive, non-discrete* or those who require *abstract data structures as inputs*. It is here that we find first glimpses of a trend to widen the definition of algorithms, to account for examples that interact with their environment. Even though Gurevich refers to broad classes of theoretical algorithms that are still formally defined (among them randomized, asynchronous and nondeterministic algorithms), the notion that an algorithm would interact with the context it operates in—be that humans or other artificial or natural influencing factors—already points to the wide umbrella term that ‘algorithm’ has become in the last decades.

With the advent of more widely accessible programming languages and the subsequent spread of programming practice beyond academic or research settings, these narrow definitions started to soften further. Donald Knuth, in his seminal book “The Art of Computer Programming” [\[69\]](#), introduced a set of requirements any program must meet to be regarded as an algorithm: *finiteness, definiteness, input, output* and *effectiveness*. *Finiteness* requires that the program terminates after a finite number of steps, which must be precisely defined (*definiteness*); the program must operate on a specified set of objects as *input* and produce *output* in a specified relation to the input, and finally, the

¹The term ‘algebra’ itself is derived from one of his books, “The Compendious Book on Calculation by Completion and Balancing”, earning al-Khwārizmī the recognition as the “father of algebra” [\[61, p.77\]](#)

operations must only take a finite length of time even for a human to calculate to be *effective*.

While this definition more closely relates to a current understanding of what algorithms and algorithmic systems are than the more theoretical definitions listed above, it still leaves the algorithm almost solely within the domain of programming, and thus, computer science in all its theoretical and practical manifestations. In order to describe and analyse algorithms and algorithmic systems in a more holistic way and to allow a critical analysis of their impacts as well, a broader definition is required. The following sections offer such broader perspectives.

2.1.2 Socio-Technical Systems

Given the ubiquity of digital technologies permeating our society in its ongoing digital transformation, algorithms take on a multitude of roles that transcend the boundaries of narrow technical definitions. While the kind of algorithms that Gurevich or Knuth describe play an integral role in these larger-scale systems, solely analysing them as technological artefacts disconnected from the larger technical, social and procedural context they are deployed in is simply not sufficient to critically assess their impacts. To do so, a different, broader and more inter-disciplinary perspective is necessary: algorithms as *socio-technical systems*.

Socio-technical systems² are a theoretical approach to characterize the complex interplay between humans and technology. The term itself was first introduced by Trist, Bamford and Emery in the 1960s as part of their work at the Tavistock Institute investigating English coal mines and their workers [70]. Much like today’s larger Digital Transformation of society, the study focused on the coal mining industry’s transformation through mechanization and automation in post World War II Britain, and described the self-organizational practices of coal workers as they organized their own work in relation to new technologies and techniques becoming available. At the time, the general trend for work organization in coal mines followed Taylor’s scientific management theory (i.e., *Taylorism*) [71] and relied on bureaucratic principles. Trist et al. studied a specific coal mine whose management and miners diverged from these principles—and did so by ‘disobeying’ the technological imperative to positive effects on morale as well as on the economic success of the mine. They concluded that this reorganization process would constitute the “[...] ‘*emergence of a new paradigm of work*’ [...] *in which the best match would be sought between the requirements of the social and technical systems*” [70, p.41].

They also note that the nature of this process is one of *self-regulation*, i.e., a perpetual re-negotiation between the technical and social components of such a system, and deduce

²Socio-technical systems are sometimes abbreviated as ‘STS’, which can lead to confusion with Science and Technology Studies. To avoid such overlap, STS refers only to Science and Technology Studies in this dissertation.

the term ‘socio-technical system’ from this process in the context of an enterprise:

“The technological component, in converting inputs into outputs, plays a major role in determining the self-regulating properties of an enterprise. It functions as one of the major boundary conditions of the social system in mediating between the ends of an enterprise and the external environment. [...] The technological component has been found to play this mediating role and hence it follows that the open system concept, as applied to the enterprise, ought to be referred to the socio-technical system, not simply to the social system [...]”

[70], p.43]

Relating these observations to today’s algorithmic systems is fairly straightforward. Like the technical equipment and social processes studied by Trist et al., most algorithmic systems rely on inputs—either from humans or other systems—and produce some sort of output. Where, in Trist’s case, input and output may have been physical in nature (be that material or human resources), for algorithmic systems, it is the data that they operate on, and the interactions that happen between humans and the system that represent the transfer of data as input and output.

The way socio-technical systems work also limits the types of interactions that are possible for human actants when engaging with the technical components; their design shapes the way humans approach the tasks they tackle with the help of (or sometimes against the original intention behind) the system. For Trist’s examples, those limitations involved, for instance, the technical requirements of the machinery involved in a task—such as speeds of conveyor belts for optimum production rates—or external factors such as the hardness and grain of the coal vein being mined [70], p.181]. In algorithmic systems, these limitations manifest in the design of user interfaces, the types of inputs the system accepts, and the data structures and methodologies it involves. To illustrate, take the example of the criminal recidivism risk assessment system Correctional Offender Management Profiling for Alternative Sanctions (COMPAS); operating on correlations between personal information of individuals charged with a crime and prior data on how these data relate to the chance of becoming repeat offenders, the system outputs a (widely criticised and strongly biased) risk score for recidivism [45]. The human actants—in this case, both the accused individual and judges in a court of law in the United States of America—have limited ways of interacting with the technical component. The accused must answer a specific set of questions and has no way of supplying information beyond the technical requirements of the algorithmic system, and the judge can only review the output risk score, but not gain further information on the causal connection between personal data points of the accused and the predictive risk score itself. Thus, the system’s design directly influences the way humans can interact with it, and thus shape human behaviour.

Conversely, the way humans utilize the system also impacts the system itself on multiple levels. In Trist’s studies, the human actants (a cohort of coal miners) had certain freedoms in their task organization and could configure the technical components to fit the requirements of their tasks, and used this agency to develop more efficient and less strenuous procedures [70, p.41]. In the case of algorithms, certain algorithmic technologies implicitly integrate human feedback into their outputs, thus changing as they are being used: most notably, these include machine learning techniques such as recommender systems [72] that dynamically adapt to their environment over time, based on user inputs. Secondly, the use or abuse of certain systems impacts the development of future iterations of the same system. Take Google’s now retired “PageRank” algorithm—the system that rated the relevance and importance of websites based on the number of incoming and outgoing links—as an example: After it became clear that, in order to be listed high up in the search results, a website needed to score a high PageRank, entire industries working to create and maintain so-called link farms, spam blogs and content farms appeared seemingly overnight. Consequently, Google adapted its methods for ranking websites in what Pasquale calls “*the endless cat-and-mouse game of search engine optimization*” [73, p.65]. Similarly, Google’s AdWords system—the algorithmic system matching Google searches with online ads, and the prime example for what Zuboff [74] calls *surveillance capitalism*—went through iterations of evolution as well: starting out with pricing schemes based on *click-through rates* (as opposed to how many users see the ad), AdWords has evolved into a complex machinery of automated auctions, performed within milliseconds every time a user requests a webpage that features Google ads [74, p.83]. While the reasons for these changes may not be directly linked to specific and singular actions by humans interacting with the system, they serve well as examples to dispel the notion that algorithmic systems are solely technological artefacts that can be investigated and analysed detached from their environment.

Framing algorithmic systems as socio-technical systems is a useful conceptual tool to understand the complex interplay between humans and technology. Nonetheless, it is worth noting that the concept of socio-technical systems as originally described by Trist has undergone a certain evolution since the original publications in the 1960s. While still applicable to a wide range of algorithmic systems today, the strong focus on the relationship between humans and technology in a work context still remains. For a number of examples—including profiling and risk assessment systems—this approach can be helpful to analyse the power hegemonies established and perpetuated by the use of such technologies; for other examples that do not implicitly fall under the larger context of [CSCW], this framing limits the perspective by presupposing a relationship between human workers, technological components and an *enterprise*. For many modern algorithmic systems (e.g., recommender systems [72]), the *workers* may be willing or even unwitting users of a larger technological system such as Amazon’s website. While they arguably are still contributing to an enterprise, their role is of a different nature than that of a worker or employee of this enterprise, which calls into question the applicability of the concept of socio-technical systems in these instances.

To expand the socio-technical concept and to include these even wider contexts, I turn to the concepts of *socio-technical assemblages* and [Actor-Network Theory \(ANT\)](#).

2.1.3 Socio-Technical Assemblages, Actor-Networks and Co-Production

Assemblage thinking originates with Gilles Deleuze and Félix Guattari [75, 76, 77]; the English term *assemblage*, while etymologically of French origin, is itself a translation (subjected to quite some critique³) of the original *agencement*. The idea of an assemblage is hard to pin down, a fact that is reflected in Deleuze's own words:

“What is an assemblage? It is a multiplicity which is made up of many heterogeneous terms and which establishes liaisons, relations between them, across ages, sexes and reigns—different natures. Thus, the assemblage’s only unity is that of co-functioning: it is a symbiosis, a ‘sympathy’. It is never filiations which are important, but alliances, alloys; these are not successions, lines of descent, but contagions, epidemics, the wind.”

[79, p.69]

What becomes clear through this definition is the symbiotic nature of assemblages: mutual benefits, or what Deleuze calls *co-functioning*, are the essential qualities of assemblages that give them coherence. What separates assemblages from the idea of socio-technical systems—beyond the limitation of the latter originating from the context of capitalist enterprises—is how interwoven the constituent parts and relations of an assemblage are, and how much they influence each other. Deleuze exemplifies this with the assemblage "MAN-HORSE-STIRRUP" [79, p.70] by describing the impact the invention had on warfare: *“Man and the animal enter into a new relationship, one changes no less than the other [...]”*. However impactful technology may be, it would be a mistake to put technology itself at the centre of socio-technical assemblages, as Deleuze himself is quick to point out:

“An assemblage is never technological; if anything, it is the opposite. Tools always presuppose a machine, and the machine is always social before being technical. There is always a social machine which selects or assigns the technical elements used. A tool remains marginal, or little used, until there exists a social machine or collective assemblage which is capable of taking it into its ‘phylum.’”

[79, p.70]

Assemblage thinking finds itself adjacent to another conceptual approach: [Actor-Network Theory \(ANT\)](#). Social constructivist scholars like Bruno Latour [80] or John Law [81]

³See, e.g., [78] for an in-depth discussion.

proposed **ANT** as a way of looking at entities in the world solely constituted through their relationships to each other, including human and non-human actors alike, and assuming they have *agency*—the capacity to act—within this relational space. Countering critique that agency requires intentionality (e.g., [82]), proponents of **ANT** point out that the concept of agency in actor-networks does not presuppose this intentionality for all actors, particularly not for non-human agents. In short, Law pointedly sums up **ANT** as a “[...] *ruthless application of semiotics.*”⁴ [81, p.3]. While both **ANT** and assemblage thinking have found numerous proponents (particularly in Science and Technology Studies), some philosophical debate exists on the similarity or difference between the two approaches, notwithstanding the fact that John Law himself once stated that “*there is little difference between Deleuze’s agencement (awkwardly translated as “assemblage” in English) and the term “actor-network”*” ([83], citing [84, p.147]). To reproduce the decade-spanning philosophical discourse on this topic would transcend the scope of this chapter and, indeed, this dissertation, but characterizing both approaches as applicable and partly overlapping when using them as analytical devices to study algorithms is quite sufficient. Furthermore, relating some of the attributes of assemblages and actor-networks to algorithmic systems reveals some striking similarities to underscore this point.

Relationality, Productivity and Heterogeneity of Algorithmic Systems

To explicate, the concept of assemblages is characterized by Müller [85] through a number of constituent features, among them *relationality*, *productivity* and *heterogeneity*. Starting with the *relational* nature of assemblages, algorithmic systems themselves are remarkably interconnected in terms of their constituent parts; from technical artefacts such as software or the hardware it runs on, to the data it processes and the humans interacting with it, and finally, to the larger organisational or societal context they are embedded in. It is the relationship between these parts that gives the system as a whole agency and meaning, and transforms it into different systems when these relationships change. For instance, facial recognition technologies are being used by police to identify potential criminals captured on CCTV cameras, but the nature of the systems in place differ greatly depending on the legal context of the country that uses them and the people affected by them, even if the system’s technical architecture may be one and the same across different instances. Even within one instance, the system may have different impacts and meaning to different people affected by them; a person falsely accused of having committed a crime due to a misidentification will have a very different relation to the system than a person hoping to identify the perpetrator that robbed them. This example also serves to showcase the *multiplicity* of actor-networks according to **ANT**: one and the same system may take on different roles, meaning and have concurring—sometimes even conflicting—impacts, depending on the actors involved.

⁴Emphasis from the original text.

The second characteristic of assemblages, *productivity*, is certainly fitting for algorithmic systems as well. As Müller [85] writes:

“They produce new territorial organisations, new behaviours, new expressions, new actors and new realities. This also means that they are not primarily mimetic; they are not a representation of the world.”

[85, p.29]

Algorithmic systems have been an integral element of the Digital Transformation of society; they do not simply translate pre-digital practices into an algorithmically supported one, but fundamentally transform the context in which they are deployed, from business to health care, public administration or education. The final point in Müller’s description—that assemblages are *not* a representation of the world—may seem controversial when applied to algorithms. A large swath of Big Data applications and machine learning methodologies rely on the fundamental assumption that they, indeed, are a (somewhat) truthful representation of the world. Besides the wealth of evidence available to show how fraught with bias these systems can be (see the following Section 2.2 for a more detailed overview of these problems), this assumption is problematic in more foundational ways as well. McQuillan [86] makes this argument most clearly when suggesting that data science is more than simply a set of scientific methods of inquiry:

“A broader framework for corrective action can be generated by seeing that data science is in fact more than the sum of its parts; that it represents a new way of structuring thought that draws allegiance from older historical currents and, as an organising idea, redefines observations and norms; and that it has a social momentum derived from both its metaphysical and machinic aspects.”

[86, p.7]

He concludes his observations by positing that data science is an “[...] *automated form of applied philosophy: a machinic neoplatonism*” [86, p.7]. In this sense, algorithmic technologies based on data science methods do more than just represent reality: they interpret it, mould it based on the choices made by their creators and the optimizations inherent in their techniques, and, in their applications for the purpose of prediction, classification or risk assessment, impact the constituent parts of their assemblages as well as society as a whole.

The third characteristic—*heterogeneity*—highlights the different natures of the entities involved when applied to algorithmic systems. Müller’s assertion that, in assemblage thinking, there are no assertions as to “[...] *what can be related — humans, animal[s], things and ideas — nor what is the dominant entity in an assemblage [...]*” [85, p.29] demands a multi-faceted look at algorithmic systems transcending the obvious actors (e.g., humans and technology), and encourages taking into consideration further immaterial

aspects of algorithmic systems, such as the underlying ideas or legal constraints as relevant entities worth investigating. Conceptualizing algorithmic systems as assemblages thus helps taking different perspectives into account when untangling their socio-technical components.

One of the major differences between assemblage thinking and ANT as explicated by Müller and Schurr [83] is the nature of change within either ANT or assemblages: “ANT describes change without rupture, or fluidity, whereas assemblage thinking describes change with rupture, or events.” [83, p.1]. Here, both approaches can provide additional value to an analysis of algorithmic systems. On the one hand, an approach based on assemblage thinking may reveal specific and relevant events and impacts (e.g., the different stages of development and deployment of an algorithmic system in a specific context, as well as fundamental changes in the assemblage when new entities, such as users or a legal challenge, are introduced). On the other hand, the fluidity of actor-networks helps illustrate the constant change algorithmic systems are in and highlights the *ontogenetic nature* of algorithms, as they are “[...] always in a state of becoming [and] teased into being: edited, revised, deleted and restarted, shared with others, passing through multiple iterations stretched out over time and space.” [87, p.5]. Consequently, both ANT and assemblage thinking are applicable as a lens to analyse algorithmic systems and the process of their genesis and development, and may be applied situationally.

The Co-Production of Algorithmic Systems

Lastly, the relationship between algorithms and society warrants a closer look in regards to their genesis. Simply put, the question arises whether algorithmic systems as socio-technical artefacts are determining the values and structure of society they are embedded in, or whether society is the driving force determining which forms new technologies such as algorithmic systems take. These two stances—*technological determinism* or *social constructivism*—are two competing viewpoints that, each in their own way, fail to account for some of the developments and hegemonies of power created by or influencing the development of algorithmic systems. A purely techno-deterministic viewpoint falls short in explaining the way technological developments are conceived, while a purely social constructivist stance limits the analysis of how new technologies (e.g., surveillance technologies) impact the social order. To resolve this contradiction, scholars in STS propose a different stance, namely one of *co-production*. Jasanoff [88] summarizes this proposition thusly:

“Briefly stated, co-production is shorthand for the proposition that the ways in which we know and represent the world (both nature and society) are inseparable from the ways in which we choose to live in it.”

[88, p.2]

Jasanoff argues that scientific knowledge (e.g., the foundations for new algorithmic technologies and applications) cannot be created “[...] independent of political thought and

*action[...]”, and likewise that social institutions do not “[...] passively rearrange themselves to meet technologies insistent demands.” [88, p.15]. To understand the interplay of power and influence, we must conceptualize the relationship between technology and society as fluid and complex. Applied to algorithmic systems and, for instance, their regulation by legal actors within society, we can observe how these institutions are both influenced by new algorithmic technologies in their actions to regulate, and how the (proposed) regulations, in turn, influence the directions that research and development of algorithmic systems takes. The case study presented in Chapter 5, the **AMAS** system or, colloquially, “AMS algorithm”, serves as an illustration of this duality: Existing regulation in Austria was presented by the **AMS** as the legal foundation for both the use of **Personal Identifiable Information (PII)** data and the mandate to develop the system, while the **Austrian Data Protection Agency (DSB)** reacted to the announcement of the system with a legal challenge itself, citing a lack of legal coverage for the use of profiling systems such as the **AMAS**. The algorithmic system in question thus finds itself in a space fraught with tension between existing legal standards shaping its design, and future legislation potentially aimed at limiting the use and further development of such systems.*

These observations on the co-production of algorithmic systems has consequences for the approach to analysing algorithmic systems as well. To take into account this reciprocal relationship, any holistic analysis of algorithmic applications must try to avoid following one-directional narratives, such as ‘System X was a response to this (societal) problem’ or, in reverse, ‘The advent of technology/system Y produced the following issues’. In many cases, both aspects of this relationship between the social context of creation and application on the one hand and the technologies on the other are intertwined and concurrent narratives, and warrant an equally interrelated and comparative analysis. The case studies presented in this dissertation and the methodologies of inquiry used for their analysis are a reflection of this stance.

In summary, algorithmic systems—from the simplest software to the most complex AI application—can be viewed through different lenses, highlighting different qualities of their constituent parts, relations, actions and genesis. Both assemblage thinking and **ANT** are applicable concepts to disentangle these components of algorithmic systems, and serve as useful tools to conceptualize algorithmic systems beyond their technological aspects alone, and in wider contexts than just an enterprise. Finally, when taking account of the way algorithmic systems come into being and how they are situated within the myriad of societal actors impacting the way they manifest, the concept of co-production as coined by Jasanoff [88] is a useful perspective to avoid simplistic, one-directional explanations that fall short in their descriptive power for the analysis of algorithmic systems.

2.1.4 Algorithms from a Functional Perspective

Given the ubiquity of algorithmic systems impacting our daily lives both visibly and invisibly, any reasonable analysis of algorithms must include not just the technical artefact itself, but also its connections to the social world it is embedded in, as explicated in

the previous sections. For instance, practitioners of computer science—programmers, systems designers, project managers—have long since adopted a more informal definition of algorithms as, broadly, technical solutions that contain a series of steps for organizing or acting on given input to achieve a desired outcome: a solution that often is not *certifiably correct* in the sense that more rigorous definitions of an algorithm would require [21].

But it is specifically this rather informal definition of algorithms that often comes with promises of positive impacts on the social context it is being designed for, including less discrimination, less subjectivity and fewer errors compared to procedures executed by humans. To evaluate those claims, a fourth perspective on algorithms beyond the technical, socio-technical, and assemblage thinking or ANT perspectives emerges: the *functional* perspective. No longer solely looking at what algorithmic systems do *internally*, this approach focuses on what functions algorithms *fulfil in the world*.

As Barocas et al. [89] summarize, algorithms thus can function as *talismans* to ward off criticism of procedures [90, 40], a particular form of (technologically supported) decision-making [91], as an “*epistemology onto itself*” [92] or a type of rationality of social ordering [93]. For example, Citron shows how automated decision-making systems jeopardize norms of due process through combining individual adjudications with rulemaking—the former through their everyday use, the latter through the way programmers implement them:

“Programmers inevitably alter established rules when embedding them into code in ways the public, elected officials, and the courts cannot review.”

[94, p.1249]

Through their very nature, algorithms also represent a certain commitment to procedure, functioning as mechanisms that introduce—and often privilege—quantification, proceduralization and automation in otherwise human endeavours. Given their need for specificity in input, the application of algorithms in contexts hitherto controlled by humans requires an abstraction of complex, non-linear processes. “Street-level” bureaucrats that would, for instance, make exceptions or include their personal subjective assessment in their interpretation of rules or policy are being replaced in the name of efficiency by algorithmic systems that do not allow that same level of flexibility [94, p.1263].

Due to their complex nature, algorithms may also function as *black boxes* [73], making them either completely invisible to the general public, or at least inscrutable and subsequently unaccountable (for a more in-depth analysis of the larger issue of *algorithmic transparency*, see Section 2.3). But regardless of their opaque nature, algorithms are often seen as a technical solution to socio-political problems. Case in point: the enthusiasm with which, in 2013, celebrities and political experts alike touted “Big Data Will Save Politics” (cf. Figure 2.1), a stance that would be proven (almost hilariously) wrong only a few years later as the world watched the 2016 U.S. presidential elections being manipulated by Cambridge Analytica [95, 96, 97].



Figure 2.1: Cover of the MIT Technology Review, Vol. 116 No. 1, illustrating the misplaced hope in Big Data to solve complex socio-political problems.

Finally, algorithms and algorithmic technologies have grown to be functionally indistinguishable from the ideological standpoints they represent, effectively creating an *algorithmic ideology*. The implementation of a given system often obscures its underlying values, overshadowed by lofty claims about the system’s benefits. Two examples illustrate this particularly well: Fitness trackers or self-tracking systems and blockchain technologies and applications. In the case of fitness trackers, the implicit promise of a better, healthier lifestyle if only the user would have access to these detailed measurements and analytics promotes the underlying assumption that said healthy lifestyle is an individual responsibility [98]. This assumption, however, is countered by numerous studies that show how strongly socio-economic factors influence a person’s health independently from their other risk scores [99, 100]. Secondly, blockchain technologies are often promoted as a replacement for the existing solutions to the societal issue of trust between strangers. Generally, this problem is answered through established, trusted third parties, such as banks or governmental institutions. It comes as no surprise that many of the arguments made for the use of blockchain technologies - such as cryptocurrencies like Bitcoin - are closely linked to libertarian notions of a general distrust towards such institutions [101]. Both examples show the functional use of algorithmic technologies and their applications as carrying ideological values and their utilization as arguments for a certain, underlying political ideology.

2.2 Bias and Discrimination

With their continuing permeation of almost all aspects of modern society, algorithmic systems have come under scrutiny for a variety of reasons, first and foremost due to their potential for bias and subsequent discrimination supported by or proceduralized through algorithmic systems. Risk assessment and profiling systems based on vast amounts of data are increasingly utilized to classify, make predictions and decisions about people. From recidivism risk assessment to predictive policing, to the distribution of social services, these systems have real and sometimes immediate effects on people’s lives, often under the assumption that—contrary to humans undertaking the same or similar tasks—they are more efficient and fair. However, upon close inspection, many examples of skewed and biased systems have been at the centre of academic discourse, particularly in communities such as the [FAccT](#) conference series.

Bias in algorithmic systems occurs when that system *systematically* and *unfairly* discriminates against certain groups of individuals in favour of others [\[102\]](#). Both conditions must be true: neither do we consider random, isolated errors that unfairly favour affected individuals as bias, nor do we consider systematic slants within an algorithmic system as bias *if* the use of that system does not subsequently discriminate against certain sub-groups in an unfair manner. What exactly constitutes ‘fairness’ is a complicated issue and subject to much debate among scholars [\[103, 104\]](#). Approaches vary depending on the underlying technology; recently, particular attention has been paid to fairness in machine learning through some quantitative approaches aimed at evaluating bias in datasets and the resulting ML-applications (e.g. [\[37\]](#)), as well as mitigation strategies aimed at the engineering side of algorithm development (e.g., [\[105, 106, 107, 108, 109, 110, 38\]](#)). However, as critics such as Skirpan and Gorelick [\[104\]](#) point out, much of the current literature hinges on a very limited definition of fairness, namely that of a system producing *disparate impacts* for (sub-)populations affected by it. To remedy this, they propose an expanded definition of fairness based on three categorical questions that require the inclusion of contextual knowledge to answer: “*Is it fair to make X ML system?*”, “*I want to make X ML system, is there a fair technical approach?*” and “*I made X ML system, are the results fair?*” [\[104, p.2\]](#). While bias mitigation strategies can play a role in answering the second and third questions, they are but one technical approach among a breadth of technical, socio-technical and social approaches to evaluate and improve the fairness of machine learning or, more generally, algorithmic systems.

One of the most widely cited case studies of a flawed algorithmic risk assessment system is the [Correctional Offender Management Profiling for Alternative Sanctions \(COMPAS\)](#) recidivism risk assessment tool developed by Northpointe Incorporated⁵. Based on a number of personal variables, the tool labels criminal defendants as high, medium or low risk of being charged with further offences in future, essentially claiming to predict future criminality. As a series of groundbreaking reports by ProPublica [\[45\]](#) showed in 2016, [COMPAS](#) was found to be systematically discriminating against people of colour in

⁵The company later rebranded as “Equivant”.

a variety of ways. Most prominently, COMPAS's erroneous classifications—false positive and false negative predictions—were massively biased in favour of white defendants, falsely flagging black defendants as high-risk more than twice as often as falsely flagging white defendants. Likewise, white defendants classified as low-risk would turn out to be re-offenders significantly more often than black defendants labelled low-risk [45, p.3]. These findings should be seen as even more serious considering the comparably low accuracy of the tool: according to Angwin et al.'s analysis, the overall predictive accuracy of the tool was as low as 62.5%, meaning that the tool performed only 12.5% better than flipping a coin. And yet, the COMPAS tool was and, as of the publication of this dissertation, still is being used across the United States of America in pre-trial courts to support judges in determining, for instance, whether a defendant should be released on bail before trial or not.

While the case of COMPAS shows the disastrous consequences a skewed system can have for individuals, it also serves as an illustrative example of the way such systems impact whole domains of society, often exacerbating social problems by creating a streamlined feedback loop that enforces and continuously re-enforces inequality. COMPAS's predictions stem from data straight out of the criminal justice system in the U.S. itself, and by systematically disadvantaging black defendants, their cases are more likely to become the source of even more skewed predictions for future iterations of the recidivism risk scoring. Far from being the only case, this pattern can be observed for a variety of algorithmic decision-making, risk-scoring, profiling and classification systems, from credit scoring, welfare eligibility scoring, child welfare risk assessment to personality tests for hiring to micro-targeting algorithms for political campaigns [111, 40]. The impact of these feedback loops can be, in and of itself, subject to bias towards marginalized groups, as Eubanks argues:

“Marginalized groups face higher levels of data collection when they access public benefits, walk through highly policed neighborhoods, enter the health-care system, or cross national borders. That data acts to reinforce their marginality when it is used to target them for suspicion and extra scrutiny. Those groups seen as undeserving are singled out for punitive public policy and more intense surveillance, and the cycle begins again.”

[111, p.12]

A particularly telling example of such feedback loops can be found in various predictive policing systems that use personal data to target specific individuals. While some predictive policing tools are aimed at determining geographical hotspots, these individually-targeting profiling tools correlate personal data of citizens with criminal statistics in order to predict the potential for future offences, and is often used to focus police attention on these individuals as a means to prevent criminal offences [112, 113]. This practice has been widely criticised by scholars [114, 115, 116] as highly problematic due to the inherent nature of the data feedback loop between prediction and police

action: Naturally, arrest records occur where arrests are made, and if an automated system influences where the police focus their attention, these data are reinforcing the system’s predictions for specific people or locations, leading to further police attention and arrests, and so forth. If the original data fed into these systems is already flawed, i.e., biased towards certain populations, socio-economic backgrounds or ethnicities, then these predictions will—in line with the *garbage (data) in, garbage (data) out* principle—do little to improve the situation and likely exacerbate pre-existing discriminatory practices.

2.2.1 Classifying Bias in Algorithmic Systems

Various frameworks (e.g., [117, 118, 119]) have been proposed to help analyse and classify bias in algorithmic systems. While they all have their merits, they are also often specific to a given context or algorithmic methodology (e.g., social media algorithms [118]), and, subsequently, are only applicable for a subset of algorithmic systems. On a broader scale, Friedman and Nissenbaum [102] proposed a high-level classification system for bias in computer systems, which the case study on the **AMAS** system utilizes as an analytic lens as well. They classify bias in computer systems into three separate types: *pre-existing bias*, *technical bias*, and *emergent bias*.

Many of the examples outlined in Section 2.2 illustrate *pre-existing bias* as one of the root causes for a biased outcome. As both “*societal institutions, practices and attitudes*” [102, p.334] and individual human assessments are inherently subject to a variety of bias, these can be embodied by an algorithmic system through the data they either operate with, or get trained on. Friedman and Nissenbaum differentiate two variations of pre-existing bias: individual pre-existing bias can get introduced into a system through individuals that have significant impact on the design, implementation or application, and societal bias that informs the overall goals, concepts and trajectory for the system in question. Both types of bias can also be introduced *explicitly* as a conscious effort, or *implicitly* and despite best intentions to avoid them [102, p.334]. The example of predictive policing presented in the previous Section 2.2 exemplifies this danger particularly well: as all crime-related data is inherently incomplete and influenced heavily by politics and policy of a criminal justice system, any system built upon this data will necessarily exhibit similar bias.

As all algorithmic systems are, to some extent, abstracting reality into a technical or mathematical form that can be processed by computer systems, they are prone to introducing *technical bias* as part of that process. The practice of *modelling*—for instance, when translating human constructs to “*quantify the qualitative, discretize the continuous, or formalize the nonformal*” [102, p.335]—requires developers to make various *value judgments*, which will inevitably favour one interpretation of reality over another. For example, a risk assessment system correlating personal attributes such as age to other variables might need to merge datasets into age groups instead of taking individual age into account. The choice of how to model these age brackets can have a significant and disparate impact on individuals being assessed through the system, particularly if their age falls close to a hard threshold (e.g., over / under 30 years of age). Similarly,

the technological limitations of computer systems often prohibit the use of valuable information in decision-making, such as personal motivation or mental health, due to its qualitative and difficult-to-grasp nature. The omission of such data sources in algorithmic decision-making—which human decision-makers include (often subconsciously) in their assessments—can be both a *source* of machine bias, but can also function as a *remedy* for human bias. Subsequently, the impact that the presence or absence of this information in algorithmic decision-making may have requires our close attention and a careful and nuanced analysis.

Finally, the occurrence of *emergent bias* underlines the immense impact the context of application and operationalization of algorithmic systems can have on its fairness and overall performance. This bias remains difficult to predict at the time of implementation of algorithmic systems, as its emergence is the result of “*changing societal knowledge, population, or cultural values*” [102, p.336]. An example of a system particularly prone to such emergent bias are Clinical Decision Support Systems (CDSS), which are “[...] *unavoidably biased towards treatments included in their decision architecture.*” [120, p.8] and can, subsequently, be cumbersome to adapt to new medications or treatments becoming available. Another example for a changing societal value resulting in emergent bias is the introduction of a third gender option for citizens in Austria in 2018 [121], which created tensions for a number of (administrative) algorithmic systems that would not recognize such a third option. This particular example is discussed in more detail in the case study of the AMAS system [3, 4] presented in Chapter 5.

2.3 Algorithmic Transparency

Given the multitude of challenges arising from the widespread use of algorithmic technologies with particular impact on humans, including the issues of bias and discrimination outlined above, it is not surprising that one of the issues brought up most frequently in both scientific and political discourse is algorithmic transparency. However, before algorithmic transparency became a focal point of interest, scholars from various fields already identified the challenge of *automation transparency* in complex systems as a pressing issue related to trust and security. The rise of automation in the industrial sector, for instance, simultaneously gave rise to worries about the ability of human operators of complex systems (e.g., in manufacturing, nuclear energy production or aviation) to grasp the inner workings of these automated components, particularly in the case of errors that required human intervention. At the time, scholars like Norman [122] argued that the issue lies not with the level of automation or the safety of the systems alone, but specifically with the system not providing an appropriate level of “*continual feedback and interaction*” [122, p.589]. Synonymously, this issue was described as *opacity* or *lack of feedback* [123, p.94], or, from a positive point of view, as the *observability* or *informativeness* of an automated system [124, p.4-5]. Over the course of the last decade, the discussion has somewhat shifted away from these specialized contexts of automated (industrial) systems, towards more fundamental questions relating to transparency of algorithmic systems—not least due to the fact that automation technologies have found

their way into everyday life, making the issue of transparency or explainability relevant to more diverse contexts of application and society at large. As Pasquale [73] discusses at length in his book “*The black box society: The secret algorithms that control money and information*”, it is no longer just technical experts or operators of sophisticated machinery who are in dire need of insight into the inner workings of the automated tools they work with, but nearly everyone else as well. Whether we are looking to get approved for a loan, get insurance, or simply perform an online search, our inputs are processed by ever more complex and opaque algorithms that score and rank our data, perform risk assessments, and exchange data with a number of other (business) entities, which may be nigh impossible to determine. The stark tension between this “*decline in personal privacy*” [73, p.4] and the increasing inscrutability of the algorithms processing that data makes algorithmic transparency not just an issue of industrial safety, but one of societal power hegemonies as well. When algorithmic systems are employed to make decisions (or at least support humans in making these decisions), they become more than a descriptive tool: they exercise authority and hold normative power, or—in the words of Danaher—they become *algocratic systems* [54]. Finally, while the exact relation between algorithmic transparency and accountability is still subject of active inquiry—including the work presented in this dissertation—one thing remains undisputed: holding algorithmic systems accountable requires at least some level of insight into their workings, or, in other words, black-boxed systems are mostly unaccountable.

2.3.1 System Transparency vs. Ex-Post Explainability

The existing wealth of literature on automation and algorithmic transparency (e.g., [125, 126, 127, 128, 129, 73, 130, 28]) also provides a diverse set of definitions what ‘transparency’ could and should mean. For the context of this dissertation and the prior work on algorithmic accountability and transparency it builds upon (i.e., [3, 4, 5, 2, 1, 6]), I adopt an overall approach based on work by Mittelstadt et al. [131], distinguishing *model* or *system transparency* on the one hand, and *ex-post explainability* on the other.

Broadly speaking, the overall *system transparency* refers to the possibility of understanding the inner workings and processes of a given algorithmic system, its constituent parts and their relations and interactions, including the interactions between human and non-human actors in the socio-technical assemblage making up the system. When taking a more technical stance on analysing an algorithmic system, the term *model transparency* is more appropriate and describes the mathematical and statistical methods and paradigms utilized in the system, as well as their specific configurations that allow the transformation of inputs into outputs. As Mittelstadt et al. summarize, three distinct aspects of transparency can be sought to gain a “*mechanistic understanding of the functioning*” [131, p.2] of either the model as a whole (*simulatability*), its constituent components (*decomposability*) or the training algorithm (*algorithmic transparency*)⁶.

⁶This terminology creates conflicting definitions for the term ‘algorithmic transparency’. Due to the fact that, in the context of this dissertation, *algorithms* and *algorithmic systems* denote complex socio-technical assemblages, and not the technical artefacts alone, the term ‘algorithmic transparency’ is

Thus, *system transparency* as used in this dissertation both includes and transcends pure model transparency, by including non-technical aspects necessary to comprehend an algorithmic system and its functioning in its entirety.

Conversely, *ex-post explainability* refers to the possibility of tracing the output of an algorithmic system to its given inputs, or, more generally, explaining why a system exhibited a certain behaviour. This type of transparency applies to almost all algorithmic systems producing specific instances of outputs, including (but not limited to) Automated Decision-Making (ADM) or Automated Decision Support (ADS) systems, classification, ranking, risk assessment and profiling systems. In particular, the growing field of Explainable AI (XAI) [131] is mostly concerned with researching how to provide *ex-post explanations* for machine learning methods and paradigms. For a more detailed description of the challenges and issues posed by machine learning systems for algorithmic transparency in both of its forms, see the subsequent Section 2.3.3.

2.3.2 Transparency Challenges

The need for improved transparency of algorithmic systems is a relatively uncontroversial stance among scholars of CAS and related fields (e.g., [132, 133, 134, 28, 135]). The question of *how* to achieve such transparency, however, is much less clear, given the numerous challenges to transparency posed by these systems. Figure 2.2 illustrates the taxonomy of challenges and issues synthesized from the literature referenced in this and the following section. Burrell [127] identifies three distinct forms of opacity prevalent in those algorithmic systems she deems “*socially consequential*” [127, p.1], e.g., algorithmic systems whose outputs have direct impacts on human lives (including the areas of application mentioned above, as well as news trends, market segmentation and advertising, spam filters or credit card fraud detection). First, many algorithmic systems are *intentionally opaque* [127, p.3] (see Figure 2.2) in order to maintain an advantage over competitors: after all, financial interests may be tied tightly to the performance of certain algorithmic systems or the promises made to customers based on those systems. Disclosing all relevant information about the COMPAS system ([45], see Section 2.2 for a more detailed description of the case), including the exact variables and possible values used, as well as publishing the data utilized to train the predictive model would have given competitors an edge in developing their own systems, to the detriment of the financial success of Northpointe Inc.; in these cases, the companies in question see the (involuntary) disclosure not as a measure to improve transparency, but rather an act akin to industrial espionage and put safeguards in place to prevent such a disclosure.

Beyond these issues of trade secrecy, both the authority and effectiveness of these systems sometimes hinges on their *inscrutability*: for instance, search-engine-optimization is an adversarial way to ‘game the system’ that led to the “*endless cat-and-mouse game*” [73, p.65] between search engine operators and advertisers trying to achieve a higher rank for

consequently used to describe both system and model transparency as well as ex-post explainability, superseding the definitions by Mittelstadt et al..

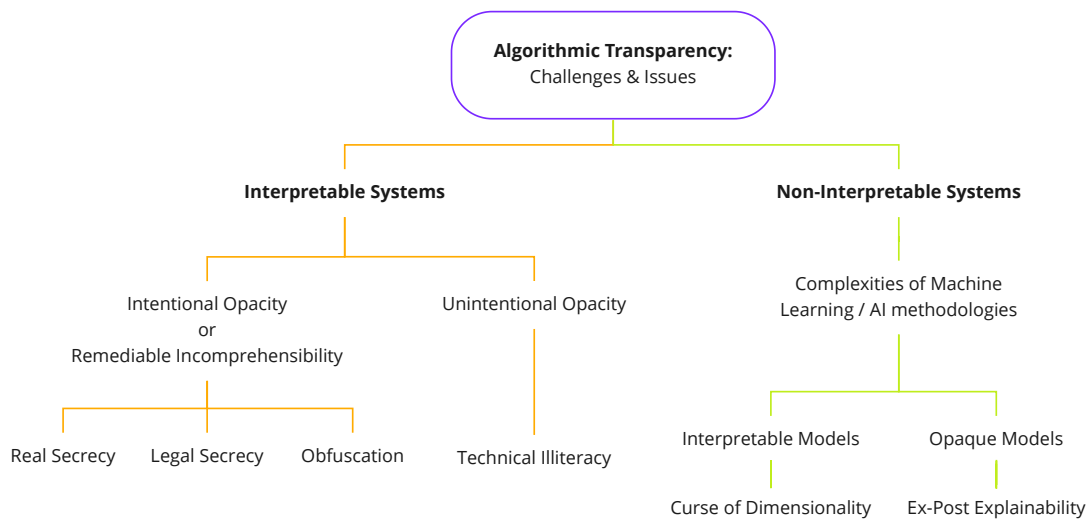


Figure 2.2: Structured, taxonomic overview of challenges and issues related to algorithmic transparency.

their clients’ web pages (see Section 2.1.2 for a more detailed description of this practice). To maintain the authority of search engine results and be able to reasonably claim to deliver the results most *relevant to the user*, as opposed to delivering the *results most beneficial to the advertisers*, search engine operators such as Google, Microsoft’s Bing or Yahoo are actively trying to tightly control which details about their algorithms become accessible to the general public. Similarly, algorithmic systems designed to classify spam email or detect cybersecurity threats and intrusions are prone to circumvention once their exact functionality—be that in the form of code or training data—become widely known [127], and their inner workings subsequently remain intentionally opaque.

Lastly, Pasquale [73] identifies three strategies of intentional opacity he subsumes as “*remediable incomprehensibility*” [73, p.7]: *real secrecy*, *legal secrecy* and *obfuscation* (see Figure 2.2). *Real secrecy* ventures to “[...] *establish a barrier between hidden content and unauthorized access to it*” [73, p.6]—in other words, any efforts to restrict access in the same way we lock doors or use passwords to protect our homes or privacy, fall under this category. *Legal secrecy*, in contrast, implements legal requirements of secrecy, such as the obligation of health care providers not to disclose private health information to third parties. Applied to algorithms, both real and legal secrecy are at play when companies restrict access to specifications or data they do not want to or are legally prohibited from sharing. Finally, *obfuscation* takes on a special role in the context of algorithms. Defined by Pasquale as the “*deliberate attempts at concealment when secrecy has been compromised*” [73, p.6], aiming for this type of opaqueness may be a viable strategy to hide the inner workings of a given algorithmic system even when forced to disclose information about it (for instance, by court order or as part of an algorithmic audit process). By providing too much information, or providing information in an

unstructured or incomprehensible way, the process of comprehending an algorithmic system's functionality may simply become such a laborious task that neither single individuals nor teams of experts can do so with reasonable effort. At the same time, this kind of deliberate obfuscation is hard to prove and easy to defend in light of the complexity of some algorithmic systems. Given that the development of these systems often involve entire departments of software developers, project managers, scientists or engineers—essentially what Nissenbaum [136] denoted as the *many hands problem*—it is often difficult to argue that a full disclosure of relevant information about these systems should not entail the equivalent of thousands of pages of data and documentation.

Returning to Burrell's taxonomy, the second form of opacity she distinguishes—*technical illiteracy*—often guarantees the success of the intentional strategy of obfuscation (see Figure 2.2). The specialized skills required to develop, comprehend or analyse algorithmic systems (such as software engineering, data mining, statistics, and programming, to name but a few) make the knowledge gap between those that develop algorithms and those that are affected by them particularly evident [127]. For the overwhelming majority of the population, a comprehensive understanding of how automated targeted advertisement auctions work and why they end up seeing a particular ad as part of their Google search results is as unattainable, and—arguably—unnecessary, just as a comprehensive understanding of how the aircraft transporting them home for the holidays really works. However, the safety rules and regulations imposed upon passengers are more easily explainable to the general public than why Google requires the collection and processing of their personal data when they make use of their web search. Furthermore, the average citizen can rely on the fact that the aviation sector is well-regulated with the safety of passengers as the main concern, whereas the same cannot be said about most algorithmic technologies. But even when putting aside any aspirations towards a full transparency of algorithmic systems for *everyone*, and looking at specific groups of professionals (e.g., policy experts, government workers, algorithmic auditors or (investigative) journalists), the knowledge gap remains a prohibitive barrier to the understanding necessary to do their jobs. As Diakopoulos [137, 138] argues, while journalists could theoretically step up to educate and explain algorithmic systems to the general public, they often lack the technical background, resources or access to gain a sufficient understanding of the systems at hand. Given the fact that, for the larger and more complex multi-component algorithmic systems, unintentional opacity can be a real challenge even for other domain experts within the same organisation that is developing the system, the chances to bridge this knowledge gap for outsiders appear slim in most cases.

2.3.3 The Problem with Machine Learning and Transparency

The final challenge to algorithmic transparency as identified by Burrell [127] transcends intentional opacity and algorithmic (il-)literacy alike, and is a direct consequence of the *inherent complexities and attributes* of modern machine learning approaches (see Figure 2.2).

This challenge to transparency is exacerbated by a general shift from *top down* to *bottom up* algorithms, as Danaher et al. explicate:

“[O]ne of the most important high-level shifts in the design of algorithms in recent years is the move from ‘top-down’ algorithms (in which a programmer or team of programmers exhaustively defines the ruleset for the algorithm) to ‘bottom up’ machine-learning algorithms (in which the algorithm is given a learning rule and trained on large datasets in order to develop its own rules). This shift is important because the use of bottom-up algorithms creates certain problems when it comes to the transparency and opacity of algorithmic governance systems, particularly when such algorithms are incorporated into already-opaque governance structures.”

[51, p.3]

What Danaher et al. mean when they reference “certain problems” is the sheer immensity of data processing required by machine learning systems that far exceeds human comprehensibility, or what Zarsky [139, 140] calls *non-interpretable* systems. While *interpretable* systems may be based on rule sets defined by software engineers and programmers that themselves can become almost incomprehensible due to their size (i.e., the number and complex interplay of their rule sets), they nonetheless can still be “reduced to a human language explanation” ([54, p.244], citing [139, p.293]), however long or impractical that explanation may be.

Non-interpretable systems, on the other hand, defy those explanations simply because the variables its outputs are based on (i.e., *features*) are often determined by the system itself as part of the data mining process (i.e., *feature selection*) and ranked based on their predictive value. The number of features available for selection and learning—the algorithm’s *dimensions*—is also the source of what Richard Bellman famously called the “curse of dimensionality” [141, p.ix] (see Figure 2.2). Originally, this expression pointed to the fact that even simple equations become very difficult to solve at scale as the number of variables increases. For modern machine learning, the *curse of dimensionality* reveals new challenges: As each added dimension increases the size of the input space (i.e., the number of possible combinations of input data points), the number of training data examples would have to increase massively to still allow reasonable similarity-based (e.g., nearest-neighbour) generalization. As Domingos explicates, for a (modest) number of 100 features, even a massive 1 trillion training data sets only covers a miniscule 10^{-18} of the input space [142, p.82]. At the same time, machine learning approaches are based on the assumption that the selected features and their correlations do represent—to some extent and limited by the necessary process of abstraction inherent to modelling—reality. Limiting the number of features to improve its comprehensibility thus can limit the usefulness of the system, while increasing the features makes the interpretability of the resulting classifiers increasingly challenging.

Domingos illustrates this issue of a ‘dimension explosion’ particularly well for algorithms based on *K-nearest-neighbour* approaches:

“It’s not uncommon today to have thousands or even millions of attributes to learn from. For an e-commerce site trying to learn your preferences, every click you make is an attribute. So is every word on a web page, and every pixel on an image. [...] The first problem is that most attributes are irrelevant: you may know a million factoids about Ken, but chances are only a few of them have anything to say about (for example) his risk of getting lung cancer. And while knowing whether he smokes is crucial for making that particular prediction, it’s probably not much help in deciding whether he’ll enjoy seeing Gravity.”

[143, p.186]

While the curse of dimensionality is a common problem for many machine learning paradigms, it is far from the only issue affecting the potential for transparency. Belle and Papantonis [144] provide an excellent overview of current technologies in machine learning and the requisite strategies for XAI to remedy inherent opacity. They distinguish between *transparent models*—i.e., models that allow a “*human-level understanding of the inner workings of the model*” ([144, p.3] citing [145])—and *opaque models*, whose processes elude human-level understanding due to their complexity and incomprehensibility (see Figure 2.2). *Transparent models* include regression models, decision trees, k-nearest-neighbour and rule based learners; their classification as ‘transparent’ stems from the fact that they fulfil at least one of the three aspects of model transparency (*simulatability*, *decomposability* or ‘*algorithmic transparency*’ [144, 131]) as outlined in the introduction to Section 2.3. Conversely, *opaque models* include random forest, support vector machines and, most prominently, multi-layer neural networks or deep learning (see Figure 2.3 for an illustration of this taxonomy).

To illustrate this discrepancy between transparent and opaque models, consider the following two examples. First, a *machine learning classifier* is trained to identify movie preferences of a given user, by calculating a *similarity measure* (i.e., the k-nearest neighbours) based on a given number of *features*. To make a classification, it inspects the variables assigned to the user in question, finds the data point that is closest, and assigns the the user in question to the same class as the one of the closest other data point [144, p.3]. This process can be described in a human language, and may or may not (depending on the number of features) be easily done by hand to reach the same result. Contrast this first example with the following (very simplified) description of a Convolutional Neural Network (CNN) used to classify the contents of an image as a second example. To train the model using supervised learning, a number of pre-classified images are used to compute optimal values for a convolutional filter matrix that, in a series of matrix multiplications, extracts image features (such as textures, shapes or edges) by sliding over the input feature map, and a series of pooling matrices are utilized to reduce the resolution

Map of Explainability Approaches

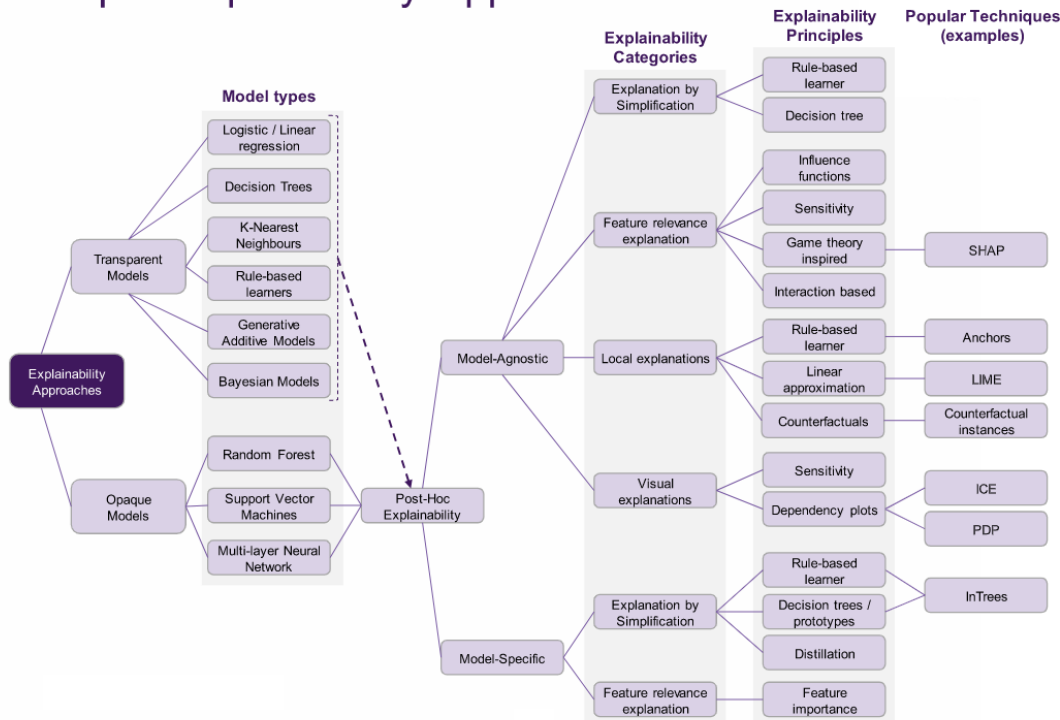


Figure 2.3: Illustration of current approaches to **XAI** according to Belle and Papantonis [144, p.5]

of the feature maps [146]. After a number of these convolution and pooling steps, one or more fully-connected layers perform higher-level reasoning to produce a classification based on probability values for the features extracted previously by the convolutions [147]. To finalize the model, a loss function modelling the wrong classification of features is minimized through stochastic gradient descent. Having determined these optimal parameters (e.g. filter type or size), the same process is applied to a new image, which produces the same type of output layer with classification probabilities. As illustrated by this extremely simplified and surely reductive description of the general functionality of a **CNN**, these steps of optimization and minimizing of the loss function may yield excellent classification performance, but does not lend itself to human understanding, neither in terms of model transparency nor in terms of ex-post explanations. In other words, while answering the question “Why was this user classified as being interested in horror movies?” only requires pointing to the number of similar users also interested in horror movies, answering the question “Why does the CNN classify this image as containing a cat?” involves, at best, an increasingly byzantine number of mathematical explanations.

To address this shortcoming of machine learning methodologies, the burgeoning field of XAI is offering a number of approaches claiming to provide human-level ex-post explanations. For an overview over current approaches, the taxonomy provided by Belle and Papantonis in Figure 2.3 [144, p.5] is a good starting point; a full evaluation of the state of the art of XAI advances surely exceeds the scope of this dissertation. What is more relevant to the overall question of algorithmic transparency in light of the increasingly prevalent application of machine learning technologies remains the question whether or not these approaches can, for real-world applications, provide the levels of transparency and explanation necessary to gain the *deserved trust* of users of such systems. As Gilpin et al. [103] note, a number of attack vectors with the potential for manipulation of deep neural networks have been identified. They allow fooling such systems into changing the output classification by introducing imperceptible alterations in images; the same holds true for neural networks used in Natural Language Processing (NLP). Given the potential areas of application for these technologies in, for instance, policing through facial recognition, the danger of discrimination and bias (see Section 2.2) looms large and warrants particular attention to the viability of existing explanatory models. Giplin et al. propose two opposing evaluation criteria for such explanations: *interpretability* and *completeness*. *Interpretability* is a measure of how well an explanation describes the internals of the system via “*descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user.*” [103, p.2], whereas *completeness* is the measure of how accurate that description is to the technical internals. In the example mentioned above for image classification through a CNN, an interpretable explanation might be pointing to the similarity in shapes and edges between different cat ears as the reason for the classification, which might be highly understandable for humans, but at the same time mostly inaccurate and reductive regarding the myriad of convolutional matrix operations leading to the final classification. Conversely, while “*a perfectly complete explanation can always be given by revealing all the mathematical operations and parameters in the system*” [103, p.3], this explanation will be all but incomprehensible to the average user of such a classification system.

Choosing, for the sake of interpretability, to provide simplified explanations in lieu of more accurate, but complex ones, may also yield unexpected ethical dilemmas. Herman [148] describe these more interpretable approaches to explanations as *persuasive*, and asks “*When is it unethical to manipulate an explanation to better persuade users?*” [148, p.3]. This trade-off between accuracy and interpretability (or ‘persuasiveness’) requires careful consideration of this question, as any reductive explanation could potentially be misused to manipulate users into a false belief of understanding, and subsequently influence their decision-making.

However difficult balancing these decisions may become in future, the current state of advancements in XAI is far from these more abstract questions. In a survey of 23 papers submitted to a workshop on XAI, Miller et al. [149] determined that most publications included no expertise from disciplines beyond AI research (e.g., social science research, cognitive science, psychology and human behavioural studies) when presenting solutions

to explainable AI, and that actual behavioural evaluation of these models is sorely absent as well.

They conclude in their paper, pointedly titled “*Explainable AI: Beware of Inmates Running the Asylum*”:

“While the re-emergence of explainable AI is positive, this paper argues most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users. [...] [L]eaving decisions about what constitutes a good explanation of complex decision-making models to the experts who understand these models the best is likely to result in failure in many cases. Instead, models should be built on an understanding of explanation, and should be evaluated using data from human behavioural studies.”

[149], p.1]

To date, few existing models for *ex-post explanations* or *model transparency* in machine learning satisfy these requirements for human-centric conceptualizations of ‘explanations’. Furthermore, the rapid advancements and diversification of methods in both machine learning and classical, rule-based algorithmic systems make the development of a ‘unified theory’ of algorithmic transparency with a wide applicability to different types of systems and applications relatively unlikely. In other words, while it may be possible to describe—in a human-language explanation—*how* a given system processes its data, it is not always possible to formulate the *why* as clearly, and any solution will inherently have to submit to certain trade-offs between *accuracy* and *understandability*.

2.4 Algorithmic Accountability

Having established the various heterogeneous conceptualizations of algorithmic systems, from narrow and technical to broad, socio-technical and assemblage thinking perspectives (see previous sections 2.1.1, 2.1.2, 2.1.3, and 2.1.4), and given the challenging issues of bias, discrimination and transparency arising from the use of these systems (see previous sections 2.2 and 2.3 respectively), the need to hold these systems accountable is becoming increasingly obvious. Nonetheless: while scholars on (algorithmic) accountability are, by and large, in agreement on this point (e.g., [138, 150, 151, 152, 153, 136, 154, 155]), the reasoning for *why* algorithmic accountability is of such importance, and what exactly constitutes an accountable algorithmic system warrants a closer and more detailed look.

Drawing on the field of accountability studies and research into accountable governance, Bovens [156] distinguishes two concepts of accountability: accountability as a *virtue*, and accountability as a *mechanism*. First and foremost, accountability as a *virtue* is a “normative concept, [...] a set of standards for the behaviour of actors, or [...] a desirable state of affairs.” [156], p.949]. Framed as these substantive norms, accountability in this form is characterized by the multiplicity of its dimensions, including *transparency*,

liability, controllability, responsibility and *responsiveness* [157]. While neither Bovens nor Koppel specifically apply these dimensions to the domain of algorithmic systems in general or algorithmic governance in particular, their analysis nevertheless provides a valuable perspective that is equally applicable to algorithms. The importance and challenges of algorithmic transparency have already been established in the previous Section 2.3. *Liability* in the legal sense comes into play where the use of an algorithmic system results in what Bovens describes as “[...] *incidental cases of tragedies, fiascos and failures*” [156, p.954], and requires legal processes of assigning culpability. *Controllability* of algorithmic systems is in and of itself a multifaceted issue, given the legal, regulatory, and technical challenges they face. Controlling algorithms is rapidly becoming an even more difficult matter in light of their complexity and the widening knowledge gap between the domain experts designing, implementing and maintaining them on the one hand, and the administrative bodies tasked with such control. Finally, considering *responsibility* and *responsiveness* as virtuous qualities is the source of many a philosophical debate on the morality of artificial agents (e.g., [158]), posing questions on whether or not algorithmic systems could be seen as exercising either agency or intent⁷.

Secondly, accountability can be conceptualized as a mechanism between an *actor* and a *forum*, which Bovens ([156], citing [159]) traces back to its historic roots during the reign of William the Conqueror (William I) after his conquest of England: In order to establish both their autonomy and their fealty to the crown, property owners were required to literally ‘give a count’ of their properties to be listed in so-called Domesday Books, thus declaring they were “*capable and willing to adhere to a moral obligation to be called to account for [their] actions as they relate to a principal’s claim on those actions.*” [159, p.16]. This obligation to provide both explanation and justification for a certain conduct remains at the core of Bovens’ [22] more general definition of accountability described in detail in the following Section 2.4.1. Contrary to accountability as a virtue, this mechanistic view points to the *analytic* and *descriptive* nature of accountability: fulfilling these obligations can only happen *ex-post*, i.e., after the actor has engaged in a certain conduct. Subsequently, this view is only applicable for algorithmic systems after they have produced some kind of output or performed some type of processing; holding an algorithmic system to account *ex-ante* (for instance while it is still being designed) would not be covered by this definition. This limitation, however, highlights the complexity of ascribing ‘agency’ and ‘conduct’ in an assemblage of human and non-human actors: While a given system *in statu nascendi* may not be held to accountable standards, the humans and organisations working to create it may well face interrogation on their conduct, such as their adherence to ethical standards or best practices of industrial software development.

For both these conceptualizations of accountability—as a *virtue* or as a *mechanism*—Bovens [156] explicates the importance of accountability in the context of democratic governance. Following a strategy of appropriation for algorithmic systems outlined above, his

⁷In line with Floridi and Sanders [158], I provide arguments for the former, and against the latter in Section 2.4.5

arguments fit well within this context as well.

First, in the conceptual framing of accountability as a *virtue*, accountability provides *legitimacy* to governmental officials and public organisations: as western governments “*face an increasingly critical public*” [156, p.954] when they exercise their authority, so do algorithms and algorithmic systems. Whether or not an algorithmic system is seen as trustworthy, and, consequently, whether the results it delivers are seen as legitimate, can depend on the ability of humans interacting with it to exercise what Passi et al. [160] describe as *deliberative accountability*: “*collaborative negotiations between diverse forms of trained judgments and performance criteria*” [160, p.6]. An algorithmic system that is, by design or by nature of its underlying technologies, *inscrutable*, can lose both the users’ trust and its claims to a legitimate exercise of power, as the case study on the **AMAS** system and the public discourse in response to its unveiling demonstrates quite clearly.

Secondly, accountability as a *mechanism* fulfils a number of functions applicable to algorithmic systems as well, among them providing “*public catharsis*” [156, p.954] in the aftermath of calamitous events they contribute to, as well as allowing popular control of governmental authority, and providing feedback loops that encourage reflection and learning [156, p.955]. When algorithmic systems fail to achieve their objectives, or when their use comes with unwanted consequences (such as biased and discriminatory practices), accountability mechanisms offer a way to voice these grievances and criticisms, ideally in combination with the potential for consequences in order to effect change. As algorithmic systems are increasingly being deployed by governmental agencies and industry organisations alike to fulfil crucial and often highly sensitive functions—including distributing scarce resources as part of the welfare state [3], selecting and ranking job applicants [111], or score communities’ sustainable energy consumption practices [1]—accountability mechanisms are processes that facilitate a public discourse which enables both reflection on and learning from failures, in order to prevent such negative outcomes in future. Correspondingly, the absence of such processes allows organisations to avoid culpability and responsibility by utilizing the existence of an algorithmic system as a *talisman* to ward off criticisms⁸ [90]. This trope of computer systems as the *inscrutable culprit* has become so widely accepted it even found its way into popular culture, as exemplified by a now famous scene in the British sketch comedy show ‘Little Britain’, where a bank teller reacts to any and all customer inquiries by ostentatiously typing nonsense on her keyboard and, giving no further justification, responds with her catchphrase “Computer says no” [161] (see fig. 2.4).

To remain within the scope of this dissertation, algorithmic accountability will be mainly discussed from a practical, mechanistic standpoint—in other words, taking the stance Bovens describes as accountability as a *mechanism*. While the conceptualization of accountability as a virtue can provide valuable insights into the larger nature and role of algorithmic technologies and their reception and function in society, the goal of this dissertation is to provide an analysis of existing accountability practices through the

⁸For a more detailed discussion of the *functional* perspective on algorithms see Section 2.1.4.

⁹Image credit: ©HBO Everett / Rex Features



Figure 2.4: Variant of the popular meme “Computer says no” based on the sketch comedy show *Little Britain*⁹[162, 161].

case studies, and recommendations to improve algorithmic accountability processes. To that end, the mechanisms by which accountability is performed and exercised, and the factors supporting or hindering these processes must be the focal point around which the discussion is centred.

2.4.1 Public Accountability: Actor, Forum and Account

Algorithmic accountability has garnered significant attention from academic communities in recent years [23], but a coherent and shared understanding of what exactly algorithmic accountability *is* remains elusive. Given the fact that a similar shared definition does not—and arguably, cannot (see [156])—be derived even within the field of *accountability studies* in general, this is hardly surprising for the complex, inter-disciplinary field of *algorithmic accountability* either. Nonetheless, scholars researching algorithmic accountability appropriate Bovens’ [22] high-level definition of *public accountability*, which has found wide acceptance within their communities (e.g., Computer Science, STS or CAS).

While Bovens’ definition has been described as *metaphorical* [163] and might seem—at first glance—both wildly generic and contextually specific to public accountability, its use of metaphoric language makes it malleable, yet descriptive, and serves well to illustrate the involved entities and relationships of an accountability process. In other words, “[...] when it comes to how the underlying decision-making processes are described, Bovens’ work is at the fore.” [163, p.3].

Bovens thus conceptualizes accountability as

“[...] a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences.”

[22, p.9]

This definition opens up various lines of inquiry deserving of close attention, i.e., ‘Who are the *actors* and the *forum*?’, ‘What characterizes their *relationship*?’, ‘Where does the *obligation* for *explanation* and *justification* stem from, and what forms should they take?’ and lastly, ‘What is the nature of this *judgment* and *consequences*?’. To answer these questions, I will summarize Bovens’ taxonomies of public accountability in the following paragraphs before discussing the transference of his definitions to algorithmic accountability; Figure 2.5 visualizes this taxonomy for reference.

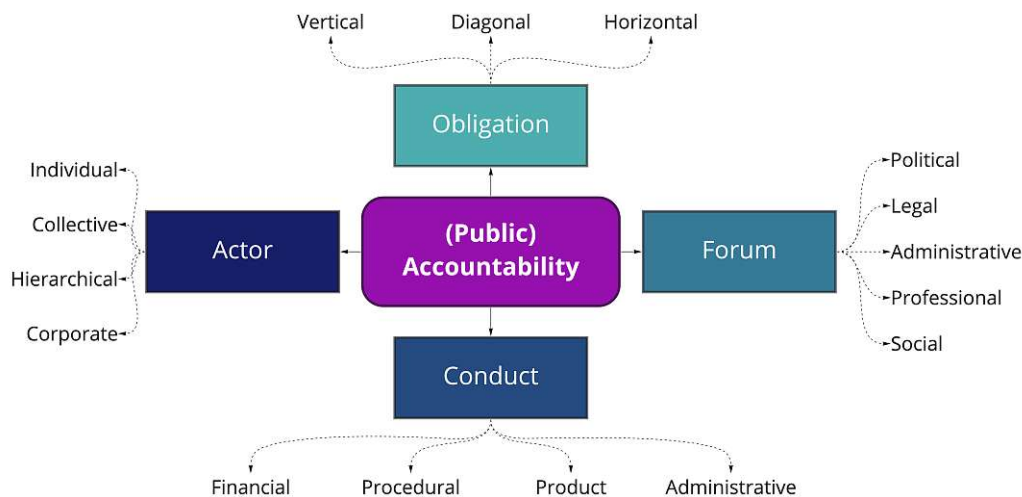


Figure 2.5: Illustration of Bovens’ [22] taxonomy of public accountability

2.4.1.1 Actors in Public Accountability

First and foremost, the role of the *actor* can take multiple forms. Bovens distinguishes four types of accountability dependent on who the actor being held to account is: *individual accountability*, *collective accountability*, *hierarchical accountability* and *corporate accountability* [22, p.19]. *Individual accountability* is the simplest form, where single individuals are being judged for their own conduct, either as being wholly responsible or responsible in proportion to their contribution to the larger conduct of an organisation they are part of. For proportional responsibility, the problem of *many hands* (first mentioned by Thompson [164] for public officials, later coined by Nissenbaum [136] for

computer systems) can complicate this process: determining the exact proportion of their contribution, and subsequently the fair share they should be held accountable for, can be difficult to impossible for large organisations or ventures. *Collective accountability*, on the other hand, makes any member of an organisation responsible for the conduct of the entire organisation—a stance that, as Bovens concedes, is “barely reconcilable with legal and moral practices and intuitions current in modern western democracies.” [22, p.19]. Only in specific cases would it be morally (and legally) acceptable to hold any single individual responsible for the collective conduct of the organisation they belong to, given that most organisations follow either implicit or explicit hierarchies that make the roles and responsibilities of different members incomparable. For instance, assuming that any employee of the AMS—from janitorial staff to caseworkers up to management—could be equally responsible for the organisation’s conduct as a whole, is ludicrous. For these cases, *hierarchical accountability* is far more prevalent, where certain and specific members carry an *a priori* responsibility for the organisation, based on their leadership role [22, p.19]. For instance, the CEO of a company would be held accountable for the overall conduct of her company, including the conduct of all its employees. While this type of accountability does exist in the private sector, in a legal sense, *corporate accountability* is far more common, where the entire organisation in its independent legal status [22, p.18] is held accountable in lieu of individual humans.

2.4.1.2 Fora in Public Accountability

Secondly, the question ‘...accountable to whom?’ leads to a classification of *fora*, which determine the nature of the accountability relationship and, subsequently, the nature of the relationship between *actor* and *forum*, including the types of judgment and consequences that can be rendered by said forum. Bovens distinguishes five types of accountability based on the nature of the forum: *political accountability*, *legal accountability*, *administrative accountability*, *professional accountability* and lastly, *social accountability*. *Political accountability* plays out in democracies due to the delegation of sovereignty by citizens to their elected officials, whose responsibility as the *forum* in the accountability relationship is to hold their own, appointed officials (e.g., cabinet members) to account, which in turn take that responsibility for the civil servants or administrative bodies they control [22, p.16]. The result is a *chain of accountabilities*, with the head of government being the penultimate, and the voters being, in a sense, the ultimate forum that renders judgment by casting their votes. Wielding a very specific kind of power, the judiciary system (i.e., courts) represents the *forum* in a *legal accountability* process. Both civil courts and specialized administrative courts hold organisations and individuals accountable for their conduct, measured by the standards of applicable law. In this sense, accountability is a purely reactive or *ex-post* process, which can only be initiated after an event occurred that warrants legal prosecution; in contrast, *administrative accountability* includes auditors, inspectors and controllers as the forum, who execute accountability processes potentially before questionable conduct occurs [22, p.17]. The fact that these three types of accountability fora are mostly situated within the sphere of government, politics and administration—although internal corporate auditing represents the exception to this

rule—is testament to the research context in which Bovens’ definition is situated. The final two types, *professional* and *social* accountability, transcend this context. *Professional accountability* includes fora from the professional context of the actor (e.g., engineers, lawyers, or doctors), and the standards by which the conduct of the actor is judged are those of their professions: for instance, medical associations may enforce a code of conduct for doctors accredited through them as part of their professional accountability. For *social accountability*, the relevant types of fora are even more broadly defined and include all stakeholders of the actor’s conduct that may have a vested interest in holding them to account. Examples include interest groups, civil society organisations or social justice advocacy groups, the media, and, in the broadest sense, the general public and society as a whole.

2.4.1.3 Public Accountability Relationships and Accounts

Depending on the types of accountability as determined by the nature of both *actor* and *forum*, their accountability relationship will be characterized through (1) the *conduct* in question, and (2) the nature of the *obligation* to justify and explain. As Bovens [22, p.20] summarizes, the aspects of an actor’s *conduct* subject to these inquiries by the forum can be varied, and depend mostly on the type of forum. Bovens mentions *financial conduct*, *procedural conduct*, *legal conduct* and *conduct of the product* as examples. For instance, in the case of courts as the forum, the conduct requiring justification will be a *legal* one, whereas a forum of shareholders of a company holding the CEO accountable for the performance of the company would investigate *financial* conduct. Another particularly relevant distinction of two different types of *conduct* offered by Bovens citing Day and Klein [22, 165] is *procedural* vs. *product* conduct, with the former holding the actors accountable for the way a certain result was reached, and the latter for the actual result or ‘product’ itself.

In terms of the nature of the *obligation*, Bovens posits a broad spectrum between *vertical accountability* and *horizontal accountability* on each ends, with *diagonal accountability* situated in between. A strong, hierarchical obligation (e.g., legal or political accountability) indicates *vertical accountability*, whereas no formal obligation to render an account (e.g., *social accountability*) indicates purely *horizontal accountability*. For the cases in between, where a formal obligation may be imposed by a third party (e.g., an auditor is being tasked with investigating the conduct of a government agency by a commission in a parliament), Bovens uses the third notion of accountability obligation, namely *diagonal accountability*.

Both the *conduct* in question and *obligation* to justify said conduct determine the nature and form of the *account*. Financial conduct, for instance, may be accounted for by providing in-depth transactional data in combination with reasoning for the choices made, whereas legal conduct may be justified through a legal defence in court as the account. What constitutes the appropriate account for a given case is, inherently, part of the judgment rendered by the forum; providing a well-formulated account for *procedural* conduct to a court that normally requires an explanation for the *legality* of a certain

conduct will not succeed in most cases [22]. The one constant characteristic of the account in *public* accountability as defined by Bovens lies in its *public* nature or openness:

“Account is not rendered discretely, behind closed doors, but is in principle open to the general public. The information provided about the actor’s conduct is widely accessible, hearings and debates are open to the public and the forum broadcasts its judgment to the general public.”

[22], p.11]

The other reading of *public* in *public accountability* applies to the conduct being judged, which must be (at least peripherally) be situated in the public domain. Beyond public institutions, this may include private enterprise as well, if their conduct includes exercising public privileges, or if they do so with the support of public funds [22], p.12].

2.4.1.4 Consequences in Public Accountability

Finally, Bovens definition is quite non-committal in regards to the outcome of the accountability relationship, stating only that “[...] *the actor may face consequences.*”¹⁰ [22], p.9]. While Bovens concedes that there is some debate among scholars (e.g., [166]) whether a complete and successful accountability process presupposes the potential for sanctions, he still argues that the “[...] *possibility of sanctions — not the actual imposition of sanctions — makes the difference between non-committal provision of information and being held to account.*” [22], p.10]. He proposes the use of the term ‘consequences’ in lieu of *sanctions* to take into account the variety of possible outcomes of a less formal nature than *sanctions* would be. At the same time, ‘consequences’ is meant, in this sense, to be a fundamentally neutral term to “[...] *avoid bias towards negative forms of scrutiny*” [22], p.10] and thus allowing for the possibility of positive consequences (e.g., praise or rewards) in addition to the more commonly assumed negative ones. Finally, Bovens asserts of the broad range and wide variety of what may be considered a consequence, from highly formalized (e.g., “*official authorisations, financial rewards, fines, disciplinary measures, civil remedies or even penal sanctions*” [22], p.10]) to completely informal consequences such as negative publicity.

As a researcher in [CAS], I am inclined to agree to this general stance for the purpose of *algorithmic accountability* as well: Any accountability process that does not carry at least the potential for consequences would be doomed to falling short in terms of providing incentives for a better conduct toward the relevant actors. After all, the utter lack of consequences for not adhering to guidelines of ethical software development (e.g., [167], [168]) have underscored the weakness of such measures time and time again [169]. At the same time, it is important to consider the possibility of positive consequences as a way to incentivize ethical conduct through accountability processes, or neutral ones simply aimed at a change of conduct detached from moral judgement. In summary,

¹⁰Emphasis added for clarity by the author.

the versatility and variability of the term ‘consequences’ in Bovens definition is one of the core reasons for the usefulness of his conceptualization of accountability across disciplines and beyond public accountability alone. Considering consequences as a hard requirement for algorithmic accountability, while, at the same time, allowing for nuance and the existence of grey areas in practice through a broad and encompassing definition of said consequences offers a viable compromise to retain the potential for change in accountability processes without limiting either the nature of that change nor the mode of effecting it.

2.4.2 Accountability in the Context of Algorithmic Systems

Translating insights from the discipline of accountability studies as outlined above to algorithmic systems requires some careful consideration. The following section aims to do just that, by providing a more detailed look at algorithmic accountability and the roles and relationships of *actors* and *fora* in this context.

In their excellent systematic literature review “*What to account for when accounting for algorithms*”, Wieringa [23] analysed 242 English scientific articles between the years 2008 and 2018 that refer to *algorithmic accountability* and related terms, spanning multiple disciplines (including literature from CAS, CDS, legal studies, computer science and governance studies). Through their review, they note the increasing attention the topic has received within the last years, partly because of new legislative initiatives such as the General Data Protection Regulation (GDPR) [170], as well as national governmental initiatives (e.g., [171, 172]). This increased awareness of the need for heightened accountability of algorithmic systems notwithstanding, they also bemoan the lack of coherent understanding of the terminology, and venture to situate the existing literature they survey within accountability theory [23, p.2]. Using Bovens’ [22] definition relating *actor*, *forum* and *account* as an analytic lens, they explicate the various interpretations yielded by the literature review of *who* the actors and forum are, as well as *what* constitutes the account. The following section discusses the key points they infer from the literature, situates these key points within the larger scope of this dissertation, and synthesizes additional insights through merging them with additional literature not included in Wieringa’s review. For the complete details of their findings, the interested reader may refer directly to their publication [23].

2.4.2.1 Actors in Algorithmic Accountability

By appropriating the concept of public accountability for the context of algorithmic systems, the *actors* held to account for their conduct shift from politicians, administrators and members of governments towards actors as part of the socio-technical assemblage that makes up the algorithmic system. Wieringa [23, p.2] (citing Yu et al. [173]) note an important distinction of responsibility for harm inflicted by the system, namely when it is working *correctly* or *incorrectly*. To explicate on their point, arguments for reduced responsibility of software developers could be made if they implemented a system designed and conceptualized by another entity (e.g., management of their employer or a different

organisation altogether), which subsequently causes harm *as specified*. This argument would, in essence, follow the arguments made by soldiers that they were “*just following orders*” when using lethal force in times of war, which itself has found both staunch defence and rebuttal from ethicists (see, e.g., [174] for arguments in favour and [175] for arguments in opposition to this stance), and remains the subject of complex and heated philosophical debate. Whichever way a potential forum might lean in this debate, this argument only applies if the system is working *as intended*, hence the importance of distinguishing these two types of responsibility in this case.

As noted previously in this dissertation, the problem of *many hands* [136] complicates matters when attempting to identify the relevant actor(s). This issue is just as applicable to algorithmic systems as it is to public accountability, particularly given the number of individuals and organisations contributing to the design, implementation and operationalization of many algorithmic systems. In the time between the first inception of an idea for an algorithmic system and its deployment and operationalization, potential actors can include research scientists, project management, software developers, and users. The number of people involved can literally include thousands of people involved in the socio-technical assemblage, particularly if the system in question is being widely distributed and finds use across the globe. A credit scoring system may be based on bleeding edge research from a handful of scientists, conceptualized and implemented by a team of 50 or more project managers and software developers, and, upon deployment, used by tens of thousands of bank employees every day. To determine (e.g., in the case of a catastrophic failure), who in this chain of actors should be held accountable, and to which extent, is incredibly difficult, and may not be possible at all given the aforementioned challenges of intentional and unintentional opacity of such a system.

To simplify this process, Wieringa [23, p.3] identify three roles of (human) actors in the existing literature, with varying levels of responsibility: *decision-makers*, *developers* and *users*. *Decision-makers* carry responsibility for the overall design of a system, and would, in an ideal accountability process, be held to account for the various value-laden judgments expressed by algorithmic systems. This may include decisions on what level of precision is acceptable for systems based on statistical analysis, or how certain hard-to-quantify aspects of reality should be modelled (e.g., how the *trustworthiness* of a potential debtor should be quantified?). *Developers*, on the other hand, have to contend with the fact that they find themselves in the unique position to both understand these design decisions, and are responsible for translating them into the technical implementation of the system thereafter. Depending on the level of technical knowledge the *decision-makers* have, the developers may be forced to “*implicitly or explicitly make value judgments which are woven into the algorithmic system.*” [23, p.3] themselves, implying the need for a certain sensitivity to these questions of ethical nature on the part of the *developers*. Finally, the *users* of algorithmic systems would face various levels of scrutiny for their conduct depending on their involvement in the production of the output (e.g., an algorithmic decision in the case of ADM systems). To avoid confusion, it should be made clear that *users* in the context of algorithmic accountability are only those humans interacting with

a given system if they have meaningful control over the system. To exemplify: A bank employee using a credit-scoring or risk-assessment system would qualify as an active user, whereas the customer trying to negotiate terms for a loan would only qualify as a passive *patient* of the system, and not be considered accountable for the system’s conduct, since they have little to no agency to influence the system’s behaviour beyond their (often unwitting) contribution to the system’s input data.

Regarding the agency of the active *user*, however, a common mode of distinction based on the quality of human agency are *human-in-the-loop*, *human-on-the-loop* and *human-out-of-the-loop* systems [23, 54, 176, 177]: While *human-in-the-loop* systems leave the human user entirely in control of any actionable decision and merely support their decision-making process, *human-on-the-loop* systems take action themselves by default, but allow their human operators to halt or override any action undertaken by the system should the need arise. The final type—*human-out-of-the-loop* systems preclude human agency entirely and function wholly autonomously without human oversight. While this typology may sound convenient, it also suggests a more clear distinction between the three subtypes than may be possible in practice. As Wagner [177] explicates most clearly, there may be a stark difference between human agency *on paper* and *plausible* or *meaningful* human agency, depending on external factors not necessarily related to the technical implementation of an algorithmic system. He identifies two primary challenges to the notion of human agency in the context of algorithmic systems: *quasi-automation* and the *assumption of binary liability*. First, many systems can be described as *quasi-automated* if the role of humans *on-the-loop* is essentially limited to ‘rubber-stamping’ an automated decision due to time constraints, various degrees of qualification or liability, or limited levels of support given by the system to comprehend its decision. In other words: “*If the only function of the human operator is to regularly agree with the machine and only very rarely disagrees with it, it is highly likely that the human operator’s agency is insufficient.*” [177, p.115]. Second, Wagner identifies the *assumption of binary liability* as a particular challenge to this model: if it must be either the human *or* the machine who is at fault, the various grey areas introduced by *quasi-automation* will often lead to assigning blame to a single human operator in lieu of having to deal with the complicated issue of legal accountability of a non-human actor (i.e., the system itself). This is particularly worrisome given the number of examples listed by Wagner where companies use a de-facto automated system, but employ low-paid human “operators” to make meaningless decisions under enormous time pressure in order to deflect liability from themselves towards these employees [177, p.116]. Considering these arguments for the question of responsibility of human actors as part of an accountability process, the typology of humans *in*, *on* or *out* of the loop should be critically interrogated and must always be taken *cum grano salis*.

Lastly, the classification of roles presented by Wieringa [23, p.3] excludes any notion of *non-human agency* or actors. Given the conceptualization of algorithmic systems as *assemblages* or complex *actor-networks* presented in Section 2.1.3, their assumption falls short of accounting for the myriad ways in which algorithmic systems exercise their power and agency, embody and prescribe values, and perform morally charged

actions. However, simply defining algorithmic systems as actors within the accountability relationship as described by Bovens [22] would raise numerous questions on the nature of (moral) responsibility, and offers no easy solutions to this conundrum. For the sake of clarity and structure, this notion of *non-human agency* and its implications are discussed in more detail in Section 2.4.5; for this overview of the state of the art of algorithmic accountability, the systems themselves are not considered actors in the sense of Bovens' definition.

2.4.2.2 Fora in Algorithmic Accountability

Similar to the considerations on the actors outlined above, the *forum* in algorithmic accountability may take different forms and must fulfil different requirements compared to public accountability. First and foremost, Kemper and Kolkman [28] introduce the concept of a *critical audience* for notions of algorithmic transparency, and, subsequently, algorithmic accountability:

“Measures toward algorithmic accountability are most effective if we consider them a property of socio-technical assemblages of people and machines. Within such assemblages, the value of transparency fundamentally depends on enlisting and maintaining critical and informed audiences.”

[28], p.12]

Relating their plea for this critical engagement to the *forum*, it becomes clear that regardless of the type of forum—be it *political, legal, administrative, professional* or *social*—*critical engagement* presupposes an ability of the forum to understand and analyse the actor's conduct in question. This ability can be strongly limited by the challenges algorithmic transparency imposes, as explicated in Section 2.3.2. However, even under the assumption that all of those challenges *could* plausibly be addressed for a given system, scholars like Ananny and Crawford [129] have argued that transparency alone is fundamentally insufficient to guarantee accountability. Their argument rests on the same conceptual approach to algorithmic systems as *assemblages* or *actor-networks* presented in Section 2.1.3, and is the most succinct formulation of the underlying problems:

*“If the truth is not a positivist discovery but a relational achievement among networked human and non-human agents, then the target of transparency must shift. That is, if a system must be seen to be understood and held accountable, the kind of “seeing” that an actor-network theory of truth requires does not entail looking **inside** anything—but **across** a system. Not only is transparency a limited way of knowing systems, but it cannot be used to explain—much less govern—a distributed set of human and non-human actors whose significance lies not internally but relationally.”¹¹*

[129], p.11-12]

¹¹Emphasis by original authors.

Consequentially, in order to be able to fulfil their role as a *critical audience*, the forum must be able to not only understand and interrogate an algorithmic system’s internal functionality, including its technical specifications, code, underlying data and so forth, but more crucially, grasp the relational aspects of how the system interacts with the other nodes in its actor network. To concretize this challenge with an example: A hypothetical audit board tasked with investigating a facial recognition system used by the police to find and identify presumed perpetrators cannot achieve a meaningful accountability process unless they can investigate and understand *how* the system will be used, by *whom* it will be employed and in *which* situations and under which constraints—even if they get otherwise complete and unfettered access to the system’s code and underlying training data. This kind of holistic and in-depth understanding suggested by Ananny and Crawford stands in stark contrast to Wieringa’s [23, p.4] assertion that the EU’s **GDPR** regulation points us towards the individual citizen affected by algorithmic systems as such a critical audience: what plausible chance would a random individual have to gain such an understanding or claim their ‘right to an explanation’, if all they can reasonably be expected to know is that they were wrongly identified as a crime suspect by an otherwise opaque facial recognition system, or given an inexplicably low credit score?

Political Accountability of Algorithms

This example of the forum as an *individual* sits on the opposite side of the spectrum than the types of *collective* or *organisational fora* identified by Bovens and Wieringa [22, 23]. Nonetheless, even for those fora, challenges are plentiful. In terms of *political accountability* (e.g., civil servants being accountable to their superiors or to elected officials as part of the political hierarchy), algorithms complicate matters due to their increased use in governmental and administrative processes. As bureaucrats start delegating certain administrative and decision-making tasks to algorithmic systems (the **AMAS** case study being a particularly illustrative example of such delegation), their superiors will inevitably struggle to hold them accountable for an outcome they evidently did not produce alone, yet they may be liable for. This problem is exacerbated by the issue of (limited) human agency discussed in the previous Section 2.4.2.1: as administrations go through what Bovens and Zouridis [178] characterise as a shift from *street-level* to *screen-level* and finally, *system-level* bureaucrats, human agency and thus civil servants’ discretionary power gets incrementally diminished from *humans-on-the-loop* to *humans-in-the-loop* and *humans-out-of-the-loop*. In its extreme form of a purely *system-level* bureaucracy, in which algorithmic systems replace civil servants entirely, a meaningful *political accountability* will become increasingly difficult to achieve, as the political fora often cannot fulfil the requirements of a *critical audience* as outlined above, and the system itself cannot provide an account.

Legal and Administrative Accountability of Algorithms

This development may relegate algorithmic accountability towards other types of *fora*, including *legal*, *administrative* or *professional* ones. Both legal and administrative fora

(i.e., courts and administrative agencies such as governmental audit boards) face steep challenges to fulfil their roles, as Wieringa [23, p.5] note. On the one hand, algorithmic systems used in the context of government or administration make decisions that “*are not up for deliberation, as they are enshrined in law*” [23, p.4]. Presuming these systems are implemented correctly (e.g., according to a specification in accordance with the governing laws), courts can indeed exercise their role as *legal* fora to provide accountability. At the same time, most existing laws that are generally applicable to such algorithmic systems cannot govern all aspects of these systems equally and in all contexts of application, leading to “*gaps in the judicial code*” [23, p.4]. This issue is not a new one within bureaucracies, of course, but was previously addressed through the aforementioned discretionary powers of civil servants tasked while executing these laws. This reliance on discretion, however, could lead to a paradoxical or ‘*Catch-22*’ [179] type of situation: Because the law cannot cover any and all possible edge cases arising from the use of an algorithmic system, the courts will defer judgment to the discretion of the bureaucrats, whose conduct as they utilize or even just monitor the system may not allow for a lot of discretion in the first place. Subsequently, a person affected by a decision rendered or support by said system may, again, seek to challenge these decisions in the courts. This cycle of shifting blame and unclear responsibility could be termed a *vacuum of accountability* [180], where potential accountability fora endlessly fail to complete a successful accountability process.

Due to a substantial lack of regulatory power or simple absence of agencies tasked with fulfilling this role, the situation of *administrative* accountability presents itself as even worse. While some initiatives to establish such regulations or agencies do exist (e.g., [171, 172]), their scope and power is still limited, and research evaluating their effectiveness is still scarce. One of the few notable analyses is provided by Metcalf et al. [181], who compare (among others) the regulatory approaches of Canada with that of the United States towards algorithm audits through mandating AIAs. For the *Directive on Automated Decision-Making* [172] published by Canada’s Treasury Board, their analysis is sobering: while the push to mandate such an AIA for all ADM systems developed for government use is laudable, the actual assessment itself is limited to a survey that “*assigns numerical scores in a rubric format to identify risk tiers*” and asks participants to answer binary yes/no questions as broad as “*Are stakes of the decisions very high?*”, “*Are the impacts resulting from the decision reversible?*”, “*Is the project subject to extensive public scrutiny [...] and/or frequent litigation?*” and “*Have you assigned accountability in your institution for the design, development, maintenance, and improvement of the system?*” [181, p.5]. As such, this AIA neither fulfils the requirements for the accountability relation to include the chance to ask questions and provide judgment for the account, nor does it carry the potential for consequences. Furthermore, as Metcalf et al. note, critics rightly called the binary questions a “*shallow form of accountability*” [181, p.5], since it completely avoids any kind of transparency on the inner workings of the system, metrics used to evaluate the risks in question or the operationalization of the system in practice. The second approach Metcalf et al. mention comes in the form of a bill proposed to the US Congress in 2019, dubbed the Algorithmic Accountability Act (AAA)

[182]. Besides the fact that the law was never passed, the current iteration of the bill seemed as toothless as the Canadian directive, albeit aimed at all companies instead of being limited to government institutions. Under the law, companies would be required to perform their own [AIAs](#), but neither were the form or content of the assessment strictly mandated, nor was the publication of the results or potential consequences part of the proposal. In summary, while regulatory initiatives seem to be spearheading the concept of impact assessments to ensure accountability, their current or proposed implementations leave a lot to be desired and have yet to be shown to be particularly useful.

Professional Accountability of Algorithms

Professional accountability transcends the areas of public administration and governance, and assigns the role of the forum to the professional peers of the actor. On the technical side of algorithm development, processes akin to the accountability process have a long-standing history in the form of *code reviews* [183, 184, 185, 186]. These processes can help mitigate the knowledge gaps between actor and forum, since the professional peers performing the review should be domain experts themselves. While a wealth of literature on various best practices illustrates the widespread use of code reviews as an internal quality assurance procedure, code reviews are also widely criticised for their bad practices or, as Doğan et al. [186] call it, *code review smells*. In their literature review, they identify various common issues with code reviews in practice that relate to their value as accountability practice. First and foremost, code reviews are often not mandatory, enforced or otherwise institutionalized within software development companies, which leads to unreviewed or self-reviewed commits to the codebase for small changes, essentially circumventing the accountability process altogether. Secondly, as the professional peers of the developers may be direct colleagues, a widespread practice are so-called *review buddies*, i.e., reviews being performed by the same pairings of actors / forum over and over again and subsequently lacking scrutiny. Other issues identified include (1) the arduousness of the process due to back-and-forth communications between developer and reviewer, (2) signing off on a review under time pressure (“LGTM” or “Looks good to me” code reviews), (3) sweeping and largely unconnected changes that are being reviewed as one, and (4) the fact that frequent reviews slow down the development process significantly. Finally, the last issue identified by Doğan et al. strikes familiar notes in relation to the arguments made by Ananny and Crawford [129]: “*Missing context in Reviews*” [186, p.8] occurs when the reviewers lack the contextual information necessary to consider the consequences of the change or addition to the codebase they are supposed to review. This issue is already a serious challenge to large-scale projects on a purely technical level, given how interconnected and cross-dependent current software development is on libraries and modules. For instance, in the widely used [Node Package Manager \(NPM\)](#) network of libraries for JavaScript applications, security issues introduced through malicious or negligent actors have threatened to expose tens of thousands of pieces of software to attacks through propagation (see Zaidman et al. [187] for a quantitative analysis of this potential impact in the case of [NPM](#)). However, as noted before, the

technical side represents only one aspect of contextual knowledge necessary to perform a plausible accountability process as the forum, and code reviews are, by definition and in practice, mostly limited to this technical side. It follows that a code reviewer will not be the *critical audience* required for such a holistic accountability process in most cases. Consequentially, a code review alone will not satisfy the requirements for a successful accountability process as outlined above, although it may contribute valuable (technical) insights needed to perform a more holistic review.

A different reading of the term *professional accountability* in the context of algorithms might not look at the professional context algorithmic systems are created in, but rather the one they are deployed in. For this kind of accountability, the *actors* employing algorithmic systems would be accountable to their professional peers as a *forum* for *why* and *how* they utilize a given algorithmic system. Taking the case study of [AMAS], such an accountability process might mean that the [AMS] and its decision makers would be facing scrutiny in international fora of employment service organisations for their use of the system. While such accountability processes are certainly possible, their applicability is strongly dependent on the field of application: without the obligation to provide the account or the potential for consequences, professional accountability in this form would be relegated to a voluntary disclosure of internal practices of actors in a given field. Some examples for strong norms demanding professional accountability in certain fields (e.g., journalism) exist, and the use of algorithms in these fields has garnered some critical attention by scholars such as Diakopoulous (e.g., [138, 133]). True professional accountability in the sense of Bovens' definition may arise from such discussions in due time, of course, as professional associations may include rules and guidelines governing what *is* and *is not* an appropriate use of algorithmic systems in these fields. But the fast-moving pace of technological developments and the competitive advantage actors hope to gain from the use of algorithmic technologies will most likely mean those organisations of professional peers capable of imposing professional accountability are forever doomed to play catch-up with the newest, not necessarily publicly disclosed, algorithmic application in their respective fields. The success of professional accountability throughout the various non-technical fields employing algorithmic systems thus also hinges on the field's professional accountability fora and their ability to understand these technologies and their potential consequences, and their willingness to frequently adapt to new developments. In other words, algorithmic literacy and a commitment to holding those employing algorithmic systems to account are the necessary preconditions for this other reading of professional accountability of algorithms beyond the domain of computer science alone.

Another aspect of *professional accountability* consists of the various *ethics guidelines*, *codes of ethics* or *ethics frameworks* that have emerged following the recent attention of professional and academic communities on the ethics of technology research and development. As mentioned before in Section 2.4.1, large organisations of professionals such as the [Association for Computing Machinery (ACM)] have published high-level guidelines (e.g., [167, 168]), urging their members to “*Avoid harm*”, “*Be fair and take action not to discriminate*”, “*Be honest and trustworthy*”, “*Respect privacy*” and “*Perform*

work only in areas of competence” [168, p.4-8], to name but a few. While these guidelines may be valuable to individuals seeking guidance for their conduct as professionals, these codes also lack enforcement, leaving them mostly toothless, as Rességuier and Rodrigues [169] put it succinctly. Mostly limited to what Herkert [188] calls *micro-ethics*, these guidelines also show a distinct lack of applicability to the conduct of collective agents, such as corporations, governments, or administrative bodies. Furthermore, the heterogeneity of professional activities covered by organisations like the ACM necessitates these guidelines to be broad and all-encompassing, thus making them subject to interpretation in specific cases of professional accountability, and allowing them to be instrumentalised as an excuse for questionable conduct.

Finally, a discussion of *professional accountability* in computer science often touches upon the value and importance of educating future generations of practitioners on the norms and standards of our field. The value of critical reflection and interdisciplinary education beyond the purely *technical* remains undisputed, as the international trends in curricular design and often-repeated calls [189, 188] for a stronger focus on ethics education in computer science show. In terms of *professional accountability*, however, the concrete effect of these curricular interventions and shift in educational focus remains unclear, due to what Hess and Fore bemoan as “limited empirical work on ethics education” [190, p.1]. In order for these educational interventions to positively affect future conduct in regards to algorithmic systems, students would need to be given resources that are both versatile enough to fit the vast diversity of fields and application contexts they may work in, but also concrete enough to be applicable to specific instances and cases they may find themselves confronted with. On this scale, the above-mentioned ethics guidelines are situated on the extreme end of versatility: while “*avoid harm*” may be one of the most fundamental and versatile guidelines possible, the vague nature of the term *harm* requires interpretation and will be up for discussion in most instances. Thus, a more specific focus on algorithmic accountability as a *virtue*, as well as the tools to apply and evaluate that construct in real-world instances, is the necessary precondition to equip future generations with the understanding and agency to promote accountable systems and hold their peers to account for their creations.

Considering the strong focus of most of these educational resources on the personal conduct of computer science and engineering students, the resulting accountability processes either concern an *individual* accountability of students in their future careers as engineers (e.g., raising awareness for their responsibilities as creators of algorithmic systems) or, vice-versa, educating them on their responsibility to hold future employers and colleagues accountable for their decisions and conduct. The former aspect promises, ideally, a future engineering community consisting of individuals that are more reflective of their own conduct; the latter may, over time, isolate individual and organisational actors engaging in conduct their colleagues or (potential) future employees consider unethical or problematic. Although the quantitative impact of such developments is hard to gauge, anecdotal evidence suggests this is already happening. At least for the biggest and most high-profile corporate actors such as Meta (formerly Facebook) [191], we see a

shift awareness as they are struggling to hire new talent precisely due to the nature of their business and ethical concerns of future employees. Given the fact that, by and large, such processes play out on the large scale of entire workforces and timescale as large as years (if not decades), I would argue that this kind of accountability is as much *social* as it is *professional*. Whether or not engineers would also avoid working for employers whose business model and practices have not been scrutinized by society at large as much as the actors of *surveillance capitalism* [74] have, remains questionable. Case in point: Meta’s business practices a decade ago were not significantly different from today, and rather common knowledge among computer science graduates, while public awareness of Meta’s questionable conduct before the Cambridge Analytica scandal [96, 97] was much lower. And yet the question of whether or not working for Meta may pose ethical challenges for engineers only recently found their way into the public discourse, turning headlines like 2014’s “*How to get engineers to work for you instead of Facebook or Google*” [192] into “*‘I Don’t Really Want to Work for Facebook.’ So Say Some Computer Science Students.*” [193] in the course of four years.

The example given above illustrates the fundamental challenge of professional accountability for algorithms. Putting the burden on the educators teaching the next generation of engineers means pitting them against an entire industry: After all, we are asking them to deter students from working for employers that may violate our norms on acceptable conduct, while those same employers have enormous power to raise the incentives for their future employees to look the other way. Economic pressure, for instance caused by the enormous debt higher education systems like the one prevalent in the USA impose on their students, will always work against these educational efforts: ethical behaviour in the workforce, of course, is a luxury one must first be able to afford.

Considering how *professional accountability* is operationalized and enforced in other fields (e.g., journalism or the medical field), looking to professional organizations imposing a code of conduct on their field as a solution to directly pressure employers rather than employees may be an alternative worth considering. The difficulties in this strategy, however, lie in the heterogeneous nature of computer science and its related fields of application. For instance, medical associations derive their executive power to enact *professional accountability* from cultural and social agreements more than just from a legal framework; their power to sanction their members for transgressions of their code of conduct (be that individuals or other organisations, e.g., hospitals or other health care providers) rests on the incentives for said members to remain part of the organisation or face ostracization. Journalism, as another example of a field with a historically strong set of rules and guidelines, exemplifies this as well: the virtues of journalistic integrity are founded in a larger societal understanding of the importance of ethical conduct, and the self-regulation of its member associations follows an interpretation of these virtues for practical purposes. As the recent developments in journalism and the controversial debates surround issues like “*fake news*” illustrate, the power of these accountability mechanisms is waning as well, at least partially due to the impact of algorithmic technologies on the field [194, 64].

In the current landscape of algorithm development and the professional field employing computer scientists and software engineers, it is difficult to imagine an association with enough executive power: after all, what incentives can the membership in such an organization or pledging to adhere to its code of conduct offer that would trump capitalist logics of growth and competition? Thus, working towards professional algorithmic accountability means working towards a larger, societal agreement on the importance of algorithmic accountability as a virtue. Starting with future generations of practitioners by introducing these virtues to our students is surely an important first step, but to create an environment in which professional accountability can truly (self-)regulate the practice of algorithm development, educating the general public on these matters must be the larger goal. Examples like the one provides above on the impact of the broader discussion of the Cambridge Analytica scandal offer some hope for the feasibility of such an approach, but also illustrate how far we still have to go. Only when the virtues of ethical, transparent, and accountable algorithm development have become as widely understood and agreed upon as the virtues of independent and objective journalism, or the values embedded in the Hippocratic oath, could we, arguably, expect an effective self-regulation of the field in all its heterogeneity to take form.

Social Accountability of Algorithms

Having identified the different challenges *political, legal, administrative* and *professional* fora of accountability face, the *social* forum remains. The number of social fora paying attention to and demanding accountability for algorithmic systems has skyrocketed in recent years: NGOs such as *Algorithm Watch*^[12], *Privacy International*^[13], *noyb*^[14] or *epicenter.works*^[15] are increasingly focusing on issues of inequality, discrimination and surveillance related to algorithms. These organisations fulfil a vital role for the civil society by spotlighting and publicly responding to new developments, such as the introduction of the **AMAS** system discussed in Chapter 5. In doing so, they highlight the issues and garner media attention, thus increasing the public pressure on the actors to respond and justify their conduct, and they translate between technical and domain experts and the general public [23, p.5-6]. In terms of the accountability relationship, the social forum only wields a limited power to impose consequences in the form of, mostly, bad publicity, or by delegating consequences for potentially illegal conduct towards legal fora. As beneficial as this engagement can be, these organisations face similar issues of (intentional) opacity as other fora, with the added complication of national laws impeding their requests for information. Not all countries have implemented freedom of information laws like the US **Freedom of Information Act (FOIA)**, and depending on the actor, responding to requests for information may be voluntary. To support social accountability in the form of a social contract, Rahwan [195] goes so far as to argue the need for a control paradigm akin to the *human-in-the-loop* concept on a societal

¹²<https://algorithmwatch.org/en/>

¹³<https://privacyinternational.org/>

¹⁴<https://noyb.eu/en>

¹⁵<https://epicenter.works/>

level, which he proposes in the form of a *society-in-the-loop* agenda for systems with a particularly broad scope (e.g., AI algorithms governing autonomous driving).

While the variety of types of fora proposed by Bovens [22] are certainly applicable to holding algorithmic systems to account, they all face different challenges in the accountability process, whether that be a lack of insight and expertise or a lack of power to demand information and justification or impose consequences. Depending on the algorithmic system in question, overlaps between different fora can occur, and the list of potential types of fora may require amendment as well. Nonetheless, the taxonomy presented by Bovens [22] and contextualized for algorithms by Wieringa [23] helps us—as a conceptual lens—to identify these challenges, issues and shortcomings, and paves the way for a closer look at the accountability relationship in the context of algorithmic accountability.

2.4.2.3 The Algorithmic Accountability Relationship: Conduct, Account and Consequences

Formalizing the accountability relationship between the *actor* and the *forum* as defined by Bovens [22] yields a procedural model of three different *phases*: *information*, *deliberation* or *discussion* and *imposing consequences*. Brandsma and Schillemans [196] propose characterizing this model as a three-dimensional “*Accountability Cube*” illustrated in Figure 2.6.

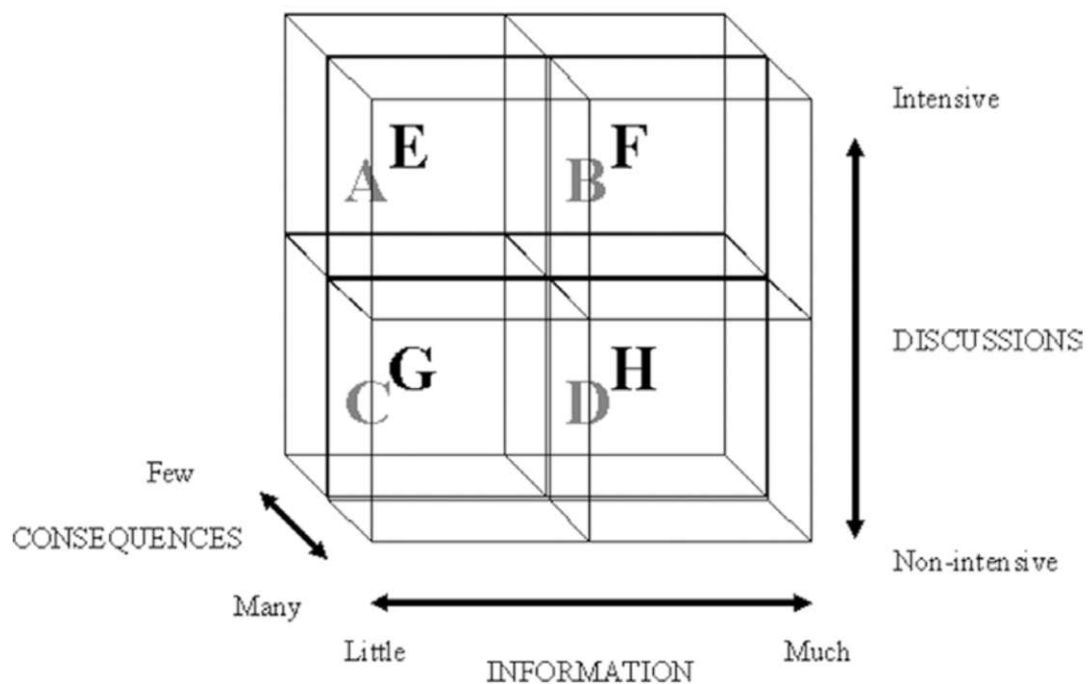


Figure 2.6: Brandsma and Schillemans’ [196, p.9] model of the “Accountability Cube”

The benefit of this visualization lies in its ability to function as an evaluation device by situating the performance results of an accountability process evaluation within the three dimensions of the cube, and allowing a comparative assessment between different accountability processes according to their assigned quadrants. Along with the basic definition of the three phases, they offer their interpretation of a quantitative scale of performance or “intensity” for each phase: from *little* to *much* information, from *non-intrusive* to *intensive* discussions, and from *few* to *many* potential consequences. According to their conceptualization, the quadrant with the “best accountability” represents a process with “much information, intensive discussions and many opportunities to impose consequences” [196, p.8]. Originally designed as a tool for *public accountability* process evaluations, the accountability cube raises some questions in the context of *algorithmic accountability*.

First and foremost, the underlying assumptions incorporate certain values of accountability. It is debatable whether or not an accountability process is really better if the discussions as part of the deliberation phase are more intensive. Whether or not more information equals more accountability is particularly questionable in light of the arguments on intentional opacity and the misuse of the transparency ideal as brought forth by Ananny and Crawford [129]. Secondly, while the accountability cube brings clarity to the process through modelling these distinct phases or aspects, it also suggests a certain independence of the three dimensions, or, as Wieringa argue, that “[e]ach of the phases can be ‘measured’ separately, giving little information does not necessarily entail little discussion amongst the forum, for instance.” [23, p.6]. Countering this argument, it should be noted that, while there may not be a clear causal connection between bad or good performance in different phases, the absence of such a connection cannot be presumed either.

Finally, the accountability cube does not take a normative stance on the relevance or weight of each of the phases: an accountability process scoring high on the information and discussion phases, but with very little to no potential for consequences would still be considered better than one with an average score for all three phases, calling into question the overall normative standards by which we ought to evaluate accountability. Building on prior work by Bovens, Schillemans and ‘t Hart [197], Wieringa [23] summarize the perspectives on public accountability as *democratic*, *constitutional* and *learning* perspectives. Each of these suggests a measure of accountability by its success in either a) assessing the executive branch of government and enforcing good behaviour as *democratic*, b) curtailing corruption and power concentration in the executive branch as *constitutional*, and c) providing office holders and public agencies with feedback to improve their performance as *learning* perspective. None of these perspectives are broadly applicable to algorithmic accountability, and may only be of use when evaluating algorithms in the context of government (e.g., *political* or *administrative* accountability).

These limitations notwithstanding, the three phases proposed in the accountability cube can be a useful analytic lens to characterize the process of accountability: in the *information phase*, the actor is confronted with the request to answer for their *conduct* in the form of the *account*, which the forum deliberates on in the *discussion phase*, passing

judgment and, where applicable, imposing *consequences* in response. The metaphorical and malleable nature of this definition implies the extreme heterogeneity of its terminology, and the variability of when it occurs in the life cycle of an algorithmic system.

Starting with the question of when such a relationship can occur, Wieringa [23, p.6] detail the various positions of scholars situating this process either *before* or *after* a system has been deployed, *during* the development process, or *throughout* the entire life cycle of the system. Arguments for accountability processes *before* deployment point towards limiting negative consequences through impact and technology assessments (i.e., the AIAs discussed in the previous section). Proponents of accountability processes taking place *after* the system's deployment argue that a meaningful account can only be rendered when a system has already been contextualized and situated within its deployment context, which may—particularly in the case of machine learning technologies—change the behaviour of the system significantly between testing/training and its deployment to a production environment. Those in favour of monitoring a system *during* the entire development process put forth arguments related to the complex value-laden decisions made during that time, which should be included in the account. Finally, scholars like Crawford [198] call for an even more holistic approach that should include all of the aspects listed above, given that algorithms, in particular for ADM, always reside “*within a wider sociotechnical field where irrationality, passion, and emotion are expected*” [198, p.11].

Synthesising these positions, Wieringa suggest a modular approach based on the software development life cycle, including *ex ante*, *in medias res* or *ex post* accountability processes. These processes would be applicable, but not necessarily mandatory, during the entire algorithmic life cycle: *ex ante* accountability processes would then apply to the planning and analysis stages of a given system, *in medias res* considerations would come into play during the design, implementation, and testing/integration phases, while *ex post* accountability would occur when the system has been deployed, i.e., during the maintenance phase of the life cycle model. Given the fact that more rigorous approaches may not always be possible, they argue that such a “*modular accountability framework for algorithms could help balance on the one hand the costs, and on the other the public's right to information and explanation.*” [23, p.7].

Using this (albeit—as Wieringa concede—slightly reductive and artificial) model as the basis for analysis, the true complexities and heterogeneity of the terms *conduct*, *account* and *consequences* emerge. During the *ex ante* phases, the conduct needed to be accounted for includes the actions and decisions of all parties involved in the planning and analysis stages. The list of possible parties is long, and includes not just individual decision makers, developers, and users, but also those stakeholder groups that conceptualize the problem the system is meant to solve or the task it should achieve, whose values implicitly guide their decisions. In other words, nothing less than “*development history of the entire assemblage*” [23, p.8] must be considered for this phase. Wieringa's detailed analysis of the questions that may be asked about this conduct, and the possible account rendered as a result, illustrates how difficult the role of both the actors and fora in

this relationship can be: Just to be able to ask the right questions presupposes an intricate knowledge of inter-disciplinary matters, including organisational, contextual and technical knowledge about the case—a knowledge that, arguably, only few people or even organisations can attain. Likewise, even the most cooperative actors would struggle to provide satisfactory answers for many of the decisions made, paths (not) taken and discussions had, particularly given the tacit [199] nature of much of the knowledge driving the design of these systems. Furthermore, answers given at a specific point in time may not hold true forever due to the *ontogenetic nature* of algorithmic systems (see Section 2.1.3). On the other hand, the potential for *consequences* imposed by the forum to effect positive change is the highest at this stage, when decisions are still being deliberated, but have not yet been implemented: establishing, for instance, certain mandatory development guidelines or best-practice models to be followed when implementing the system in the next phase may positively impact the outcomes of the later phases.

The insights gleaned from accountability processes at this stage of development have profound impacts on subsequent *in medias res* assessments during the design, development and testing phases of the system. As the conduct in question in this phase includes the myriad of technical and socio-technical decisions made by developers and decision-makers—including difficult choices and trade-offs related to technologies used, data sources, modelling, UI/UX design, and evaluative criteria such as precision or acceptable error thresholds—a forum as the *critical audience* must be able to trace said decisions back to the initial trajectories set in the planning and analysis phases [23, p.8]. Only then can the current conduct be contextualized and an appropriate judgment rendered: If, for example, technical trade-offs emerging as necessary during development and testing cannot be reconciled with the original intentions, an appropriate conduct might include going back to the drawing board to remedy these tensions. Such a decision, however, might incur serious and sometimes prohibitive financial consequences, and may lead to a technical decision being prioritized over one guided by social or other values. It is therefore crucial that the forum in this phase has the power to impose consequences that, from economic or organisational standpoints, may be seen as controversial, lest the accountability process remain impotent to effect any plausible betterment of the issues identified.

Finally, algorithmic accountability processes during the *ex post* phase of the algorithmic life cycle often take place in times of crisis, i.e., after ostensibly negative outcomes have already occurred—either due to errors or, conversely, because the system was designed to behave this way and the re-contextualisation between *in vitro* testing and *in vivo* deployment led to emerging issues hitherto unforeseen. It is also in this phase that the implicit assumption of moral agency being solely located with the humans interacting with or designing the system is most limiting. This assumption may lead to accountability processes that are primarily focused on assigning blame to the humans responsible, omitting more constructive approaches aimed at remedying the situation. Particularly given hindering factors such as the *many hands* [136] problem, the deliberation and

judgment rendered by the forum is in danger of becoming one-sided and ultimately serve only to appease a need to see public accountability for grievances caused, as opposed to concrete and actionable learnings that would help to improve the system and its future iterations.

2.4.3 Macro-Accountability vs. Micro-Accountability

An important addition to the discussion of algorithmic accountability as summarized above is the question of *scope*. As a keen reader may have observed, the majority of research into public accountability in general and algorithmic accountability in particular is concerned with actors, fora and their relationship on a macroscopic level: groups of people or entire organisations function as actors or fora, and their relationship is being discussed on the grand scale of states, governments, bureaucracies and judicial systems in what could be termed *macro-accountability*. From that point of view, only high-level processes can be described, and interactions between actor and forum remain nondescript in their concrete, minute enactment. Since much of the literature on algorithmic accountability appropriates theories and insights from public accountability, this broad view is hardly surprising. The value of this big picture view notwithstanding, I posit an equally important and much overlooked point of view on the other end of the spectrum: the *micro-accountability* of algorithmic systems.

Micro-accountability takes place when individual humans interact with each other and with algorithmic systems, and issues demanding accountability arise. After all, in many cases of algorithmic systems it is not solely an organisational entity that interacts with the system, but rather specific humans, be they users or operators, or simply people affected by the system's behaviour. To illustrate this distinction, consider the following scenario of a facial recognition system used in policing as mentioned before in the sections [2.1.3](#), [2.3.3](#), and [2.4.2.2](#):

“An AI system designed to identify biometric markers in CCTV camera images showing human faces falsely identifies an individual as present at a crime scene. Police utilize this system to identify the person, detain them for questioning, and confront them with the evidence provided by the system—evidence that the individual in question knows to be wrong, as they are, in fact, innocent.”

However that situation may play out, the system in question and the organisation using it (i.e., the police) will be subject to scrutiny as part of an accountability process. On a *macro* level, that process can play out over months or years, and may involve *fora* such as courts (in case the wrongly accused, or a social justice advocate group acting on their behalf, sues one of the actors), internal or external auditing organisations, and so forth. In contrast, as the person is being detained and questioned, a wholly different *micro-accountability* process may occur, in which the wrongfully accused individual demands to know *how* and *why* the police came to the conclusion that they were present at the crime scene, how the system identified them and if it could be wrong in its determination,

in order to prove their innocence. This process plays out on a much smaller scale, e.g., between two humans and the facial recognition system in question. Assuming this socio-technical assemblage proves to be ideally accountable for the sake of this example, the *actor* (i.e., the police officer) can provide answers to these questions (perhaps supported by the system itself providing estimates of low confidence for this identification), the *forum* (i.e., the accused) accepts the justification and asks, as a *consequence*, to be issued an apology for being wrongfully detained and let go. It should be noted that this ideal outcome is, even for a hypothetical scenario like the one presented above, difficult to imagine and highly unlikely under current social and technological conditions—hence the need for closer investigation of *micro-accountability* processes as outlined in the following paragraphs.

This scenario serves as an illustration of the relevance of *micro-accountability* processes for two reasons. First, given its immediacy after the initial interaction with the algorithmic system being held accountable, this process of *micro-accountability* is likely to happen before any other *macro-accountability* processes at the *ex post* stage of the algorithmic life cycle. As such, it may represent the initial contact between human and non-human actors in an accountability relationship, and can determine human trust in the system on an immediate level. Secondly, and perhaps most importantly, the successful completion of this *micro-accountability* process may preclude a lengthy, costly and perhaps ultimately dissatisfactory *macro-accountability* process. Conversely, if that *micro-accountability* process fails, be that because the forum is unable to pose their inquiry, or because the actor is unable to provide the necessary answers or is unwilling or unable to accept the consequences posed by the forum, the likelihood for such a successful *macro-accountability* process occurring later on are slim.

A Forum of One

In its most reduced form, a *micro-accountability* process may occur between two single humans as the *actor* and the *forum*. While the term *actor* already implies a singular entity we readily associate with the prefix ‘human’, *forum* carries connotations of an abstract or physical space in which a plurality of discourse, deliberation and exchange may occur. Most of the classifications for public accountability fit well within these connotations as they reference abstract fora (e.g., “courts”, “audit boards”, or “social watchdog organisations”) that at least carry the potential for such interpersonal discourse or deliberation between its (multiple) members. In the spirit of *human-centricity*, however, it is important to remember that even those abstract fora can not exist without humans, and indeed, are both constituted and enacted by the humans that belong to it. In order for such a constitutive act, I would argue, an awareness of their identity as being part of a forum is not a necessary prerequisite: the simple act of asking questions and demanding accountability upon the presumption of an obligation to provide such an account, and attempting to impose consequences constitutes them as a forum, just as they constitute the *actor* by identifying and addressing them for a specific instance of an accountability relationship. Based on this conceptualization, the roles within a forum may shift, humans

may enter and leave the forum, the processes of deliberation and the rules governing the imposition of consequences may change, but the basic foundation establishing a forum is still its constitution by human action. In other words, *fora* in accountability processes exist because humans act like one. For an accountability process to be a *successful* one, however, this enactment must occur within the larger scope of a socio-technical assemblage that invests the requisite power and agency into both *forum* and *actor*.

Thus the assumption that a *forum* will most likely manifest as an abstract space constituted by a collective of humans, plausible as that may be in many cases, should not be seen as “*limitative*”, as Bovens [22, p.16] himself explicates quite clearly. Previous work on personal accountability, for instance, even goes as far as considering one’s conscience a type of forum deliberating on the moral quality of one’s own actions: “*Personal accountability is fidelity to personal conscience in basic values such as human dignity [...]*” [200, p.230]. These extreme interpretations of (personal) accountability aside, even in abstract and collective *fora* such as the judicial system, humans enact the accountability process, and may do so alone or as representatives of a greater whole. Case in point: the government bureaucrat performing an audit of another government entity as part of *administrative accountability* processes may do so on their own, and deliberate explanations and consequences collectively only insofar as they follow guidelines or rules that stem from the larger, abstract organisation they represent. Consequently, the idea of an individual constituting “forum of one” in a *micro-accountability* process is not as far-fetched as it may seem initially.

Secondly, we can observe a larger trend towards “*direct*” or “*stakeholder*” accountability, as Meije points out in the context of public accountability:

“Accountability has broadened in recent years as citizens and other stakeholders are offered ever more opportunities to scrutinize an actor’s exploits, to debate results, and to pass judgment. Accountability to citizens and stakeholders has been labeled “direct” accountability, “stakeholder” accountability, or, most often, “horizontal” accountability.”

[201, p.7]

For algorithmic accountability, we can observe similar trends in the way the EU’s **GDPR** identifies the *data subject*—a single individual affected by data processing covered by the GDPR—as the forum in accountability mechanisms. Vedder and Naudts [202] make this argument in reference to the right to object to automated decision-making, stating that “[...] *the exercise of the right not to be subject depends upon strong accountability mechanisms towards the individual.*” [202, p.8]. As it is the data subject themselves that, in order to exercise this right, must be informed about the fact that automated processing is occurring, and are the recipient of the account of the automated decision resulting from this process, the GDPR clearly considers an individual not just capable, but responsible to fulfil the role of a *forum* by and of themselves. Similarly, the right to learn what

information a data controller has about a data subject as granted by the [GDPR](#) also represents an accountability process targeted specifically at individuals as a forum.

Whether or not these trends towards demanding algorithmic accountability as an individual responsibility should be considered as a positive development is, of course, debatable. Naturally, the legitimacy of larger fora, the larger weight of the resulting assessment, as well as the potential for greater and more impactful consequences imposed on the *actors* also suggests a higher chance for effecting sustainable change. Furthermore, larger organisations are, by and large, better equipped to participate in such accountability processes than individual actors. In contrast, *fora* in *macro-accountability* processes will always be slower to react and more cumbersome in their internal deliberation than an individual human being. Furthermore, the immense scope of deployment and subsequently large number of impacted individuals necessitates *macro-accountability* processes that either demand accountability for a large number of affected individuals, or discuss exemplary cases as precedents for later accountability processes. Either approach means that individual cases may “fall through the cracks”, as the threshold for initiating such a larger *macro-accountability* process may be prohibitive for many affected individuals.

Consequentially, the fact that *micro-accountability* processes may require individuals to take up the role of the *forum* is not to be seen as *desirable*, but rather as *necessary* in light of the limitations of *macro-accountability* processes when it comes to protecting the individual. In the end, neither *micro-* nor *macro-accountability* alone will always be sufficient for all cases, and combinations and intermediary forms will be required.

Consequences and Explanations in Micro-Accountability Processes

Considering *micro-accountability* as a variation of *macro-accountability* also requires a closer look at the impact this reduction in scope has on the various constituent aspects of the accountability process beyond just the *actor* and *forum* as individuals. Primarily, this concerns the nature of and possibility for the imposition of consequences by the forum. Just as the nature of consequences correlate to the nature of the forum in *macro-accountability* processes (e.g., the judicial system imposing sentences against a corporation, as opposed to professional peers sanctioning a member of a professional association for transgressions of their code of conduct), the potential consequences in *micro-accountability* processes also have to fit both scope and forum. Neither would it make sense to assume an individual affected by an algorithmic system would plausibly be granted the power to halt the system’s use entirely based solely on their individual judgement of the *actor’s* explanation, nor can it be expected that a layperson questioning the results of, e.g., a credit score or risk assessment system would be able to demand the same specific consequences as an expert in machine learning or statistics might.

At this point it is prudent to shortly revisit Bovens’ [\[22\]](#) suggestion for the use of the term ‘consequences’ in lieu of ‘sanctions’ as discussed in detail in Section [2.4.1.4](#). Bovens’ intention behind proposing a neutral term is the possibility to include a broad range of possible consequences, from formal to informal, and negative to positive, in the definition

of (public) accountability. In my appropriation towards algorithmic accountability, I combined this broad definition with the affirmative assertion that consequences, still, represent a hard requirement for accountability, in order to guarantee at least the potential for change, however minimal or limited that may be. In doing so, ‘consequences’ remains relevant for *micro-accountability* processes as well.

For instance, as a minimum requirement, *micro-accountability* should empower the forum with the agency to escalate the process to a suitable macro-accountability process, be that in the form of legal, professional, social, administrative or other forum at a larger scope. An algorithmic system that, for instance, requires the user to sign a non-disclosure agreement or to waive their rights to take legal action against the system’s provider before using the system would thus limit the potential for the imposition of consequences through escalation to a social or legal forum. Other potential forms of consequences befitting the smaller scope of *micro-accountability* could include, for example, the opportunity to opt-out of the use of a given system (the Austrian Electronic Health Record system ELGA [203, 204] exemplifies that option), to exercise one’s *right of erasure* as mandated by the GDPR [170], or simply the right to reject a given decision or output and demand an alternative process, perhaps executed by a human, be used instead. As these examples show, the nature of possible consequences in *micro-accountability* processes are transformed and reflect the immediacy of the process as well. Where consequences of *macro-accountability* processes may follow paradigms of *punishment* or *sanctions* as a form *deterrence* (e.g., in the case of legal accountability), consequences in *micro-accountability* processes may also be aimed at *protecting* the affected, individual forum, making their imposition a form of *self-defence* against immediate or intermediate harm caused by an algorithmic system.

Following these arguments, we can see how much power some of the *actors* have over these processes. Following Wieringa’s [23] classification, *decision makers* and *developers* of algorithmic systems shape—through their decisions when designing the system—the way *micro-accountability* processes can occur, which types of account or explanations are available, and finally, what consequences are possible or likely. That is not to say that micro-accountability processes do not occur if they are not actively designed, but rather that they may be meaningless or fail due to a lack of opportunity for explanation or consequences. Bovens et al. explicate this difference in their call for “*deliberative*” instead of “*defensive*” accountability:

“[M]eaningful accountability is an adjustment to, or a supplement for, existing forms of accountability. The account-giving is not organized as a mindlessly repetitive phenomenon, impervious to context and change, but is instead calibrated to fit specific circumstances and issues [...] This type of accountability supports, rather than presupposes, sense-making processes. [...] Finally, meaningful accountability is not about compliance with existing rules and regulations, but about whether the organization is effectively serving its mission and about whether, and how, improvements are necessary.”

[205], p.8]

Thus, the first step towards better (micro-)accountability of algorithmic systems will require the *actors'* awareness and willingness to consider these aspects as an important, even integral part of designing algorithmic systems. What Bovens et al. suggest for public accountability holds true for algorithmic (micro-)accountability as well: *actors* have much to gain from successful and meaningful accountability beyond simple compliance. Failing that, however, regulation compelling algorithm developers to consider strategies for *micro-accountability*—including measures to implement such processes—would be a logical next step. Doing so would thus follow a similar pattern to the introduction of consumer protection laws in the past, which compelled manufacturers of physical products—against significant resistance!—to avail themselves to consumers directly for exchange or repair of faulty products [206, 207].

Considering these caveats for micro-accountability, one might reasonably conclude that those *micro-accountability* processes that carry particularly little potential for imposing real and impactful consequences are, in essence, simply calls for a better *explainability* of algorithmic systems. Such a perspective, however, omits a crucial aspect of the *account* in *accountability*: there is a qualitative difference between the mere *capability* to provide an explanation, and the *obligation* to do so. While the latter, naturally, presupposes the former, merely designing and operationalizing algorithmic systems in a way that allow an individual forum to receive an explanation, but without considering that the *actor* has an obligation to do so, would also shift the power to decide what *kinds* of explanation should be provided from the *forum* to the *actor*. After all, why consider complex and difficult to implement forms of explanations (e.g., for AI/ML systems) that might benefit the *forum* in the moment when other, easier forms of explanation could be provided without violating an obligation to provide an account? In the contrary, designing a system for *micro-accountability* must involve carefully considering what types of accounts potential individual *fora* might require and deem satisfactory, and re-evaluating the resulting explanations over time. While the potential for impactful consequences may still be an incentivizing factor for such a behaviour, it should not be the only one, considering the other functions of consequences in *micro-accountability* as discussed above.

Micro-Accountability as Commitment to Human-Centricity

This call for a greater attention to *micro-accountability* processes and mechanisms aligns with the greater trend [9, 208, 209, 125, 177, 210, 211] in HCI, STS and CAS towards *human-centricity* in design and application. Recent initiatives like the Vienna Manifesto on Digital Humanism [212] illustrate this trend well in their demands to refrain from replacing consequential decisions with automation technologies, instead empowering humans through technology to make better, more informed decisions themselves. Similarly, scholars have investigated technologies such as online-consent mechanisms [9], data analytics [209], human-adaptive socio-technical systems [208] and (semi-)automated

decision-making [177] from human-centric perspectives. The core argument for human-centricity in computing in general, and algorithmic systems in particular, pertains to human agency and a responsible, sustainable way of developing computing technologies that serves, supports and expands human agency and capabilities. The credo “Technology for Humans”¹⁶ adopted by the TU Wien reflects this overall stance as well, and informed the underlying approaches of this dissertation. Given this stance, it is surprising how little attention algorithmic micro-accountability processes involving actual humans have received, as opposed to the comparable wealth of research on macro-accountability processes for algorithmic systems.

A serious commitment to human-centricity in computing, I posit, therefore must not relegate accountability to the realm of inter-organisational, high-level processes alone, but should consider the way individual human beings interact with algorithmic systems and attempt to hold them to account: hence the introduction of the *micro-accountability* perspective. Consequentially, we must evaluate measures, guidelines, frameworks and other tools aimed at improving algorithmic accountability by their capacity to integrate aspects of *micro-accountability* with *macro-accountability* processes as well. The **A³ framework** presented in Chapter 6 reflects this stance as well, allowing for a broad applicability in both *macro-* and *micro-accountability* processes.

2.4.4 Human Agency and the Accountability Process

Given the arguments for *human-centricity* and *micro-accountability* made above, the question of *human agency* and what influences or impacts said agency must be considered before we can situate it within the context of algorithmic accountability processes.

When characterizing algorithmic systems as socio-technical assemblages or actor-networks, power relations between human and non-human actors undergo a conceptual shift. As Müller [85] describes, though, **ANT** and assemblage thinking have differing views on *agency* and its origins. *ANT*’s radical insistence on relational agency as an “*exclusively mediated achievement*” [85, 30] tells us that—outside of the actors entering into associations with each other—no intrinsic agency beyond the relation can exist. By this view, we can see that the relation a human enters into with an algorithmic system would be the sole source of their capability for action as part of, for instance, a micro-accountability process. Beyond this assertion, however, **ANT** offers little to characterize this agency or allow particularly useful, generalizable deductions for the design of algorithmic systems as actor-networks aimed at maximising human agency. By comparison, turning to *assemblage thinking* allows for “*intrinsic qualities*” [85, 30] of the component parts of assemblages, which shape and add to the assemblage and its agency as a whole; in other words, components of an assemblage, be they social, technical or socio-technical, may have intrinsic *agency* outside of the assemblage, which then contributes to the assemblage in unpredictable ways and “*exceeds the properties of the component parts*” [85, 30]. Conceptualized this way, human intrinsic agency both adds to the agency of the

¹⁶Orig. “*Technik für Menschen*”

larger socio-technical assemblage, and is limited or enhanced, inhibited or promoted, and restricted or supported by the other components' own intrinsic attributes as well as the assemblage as a whole.

Following this assertion, it seems prudent to look to established theories of *human agency* from other disciplines that may offer a more structured perspective into the nature of human agency as an intrinsic quality. Various theoretical models exist, among them Giddens' *Structuration Theory (ST)* [213], Butler's post-structuralist discourse-theoretical approaches [214], and Bandura's *SCT* [215]. All of these offer some valuable insights that could be applied to the question of human agency in the context of algorithmic accountability to various degrees. However, of the three, Bandura's *SCT* offers a more human-centric and individual approach than Giddens' larger, sociological viewpoints, yet without Butler's complete deconstruction of agency as purely performative expression through discourse alone¹⁷. As the goal of describing human agency in the accountability relationship involving algorithmic systems is to find a mode of analysis that lends itself to improving said agency, Bandura's *SCT* offers the most versatile, yet concrete and individually applicable theoretical lens.

Albert Bandura's body of work [215, 217, 24, 218] on human agency culminates in what he terms *Social Cognitive Theory*. In this work, Bandura rejects the dualism between *autonomous* and *mechanical* human agency: humans are neither wholly independent agents of their own actions, nor are their actions the deterministic result of environmental influences. Instead, his social cognitive theory uses the model of *emergent interactive agency*. In Bandura's own words, humans

"[...] make causal contribution to their own motivation and action within a system of triadic reciprocal causation. In this model of reciprocal causation, action, cognitive, affective, and other personal factors, and environmental events all operate as interacting determinants."

[215, p.1]

The model of *triadic reciprocal causation* [219, 217] illustrated in Figure 2.7 describes the relation between *human behaviour* or *action*, *internal personal factors* including cognitive, affective and biological events, and the *environment* as "*interacting determinants that influence one another bidirectionally*" [217, p.6]. In other words, both internal personal factors and external, environmental factors influence human action as well as each other, and human action influences both the environment and the internal cognitive, affective and biological aspects of human existence. Although reciprocal, depending on the human

¹⁷Butler's post-structuralist approaches show some remarkable interconnections to Actor Network Theory, even beyond the obvious commonality of attributing agency to non-human entities through performativity. It might be an interesting avenue of thought to further pursue this approach, starting with some of the suggestions made by Maze [216], and subsequently apply it to an *ANT*-specific view on algorithmic systems. For the question at hand, turning to more structuralist and concrete models of human agency promises more practical applicability, hence the focus on Bandura's *SCT*

activity and circumstances, not all three aspects are necessarily equally strong in their influence.

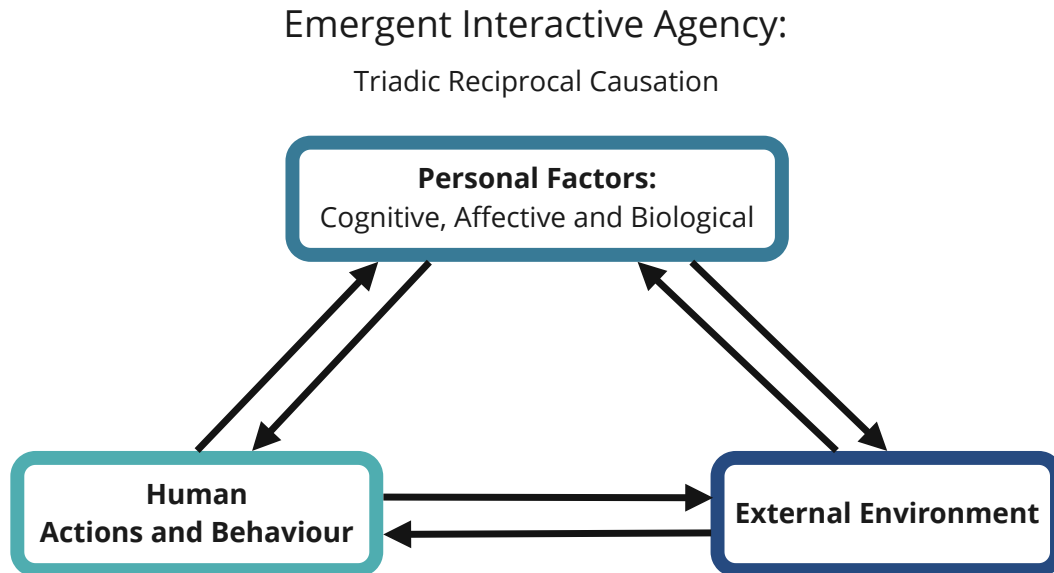


Figure 2.7: Bandura's [219, 217] *triadic reciprocal causation* as model of emergent interactive agency.

From this “*agentic perspective*” [24], human agency is characterized by four core features: *intentionality*, *forethought*, *self-reactiveness* and *self-reflectiveness*. In short, an action is considered an expression of *human agency* if it happens intentionally, in pursuit of a future goal or desired outcome, as a result of *self-regulation* and *motivation* to commit the action. The last feature, *self-reflectiveness*, describes the “*metacognitive capability to reflect upon oneself and the adequacy of one’s thoughts and actions*” [24, p.10], which may be exercised both before and after an action has been taken.

Human agency also extends beyond the *individual* or personal agency, and may also be expressed as *proxy* and *collective* agency [218]. Individual agency implies that humans influence their actions and their environment directly, but this level of control is not always attainable. In these cases, agents may exercise *socially mediated agency*, by influencing other agents to act on their behalf as proxies. Finally, complex goals and human endeavours require collaborative efforts to achieve, and give rise to interdependent actions in the form of *collective agency*, where groups of individuals “*pool their knowledge, skills, and resources, and act in concert to shape their future*” [218, p.165].

Applying these concepts to algorithmic accountability processes and modelling them as expressions of *emergent interactive agency* opens up the potential for positive change and provides a clearer perspective on what supporting or hindering influences on human action exist in specific instances of such a process. Seen through this lens, the accountability

process becomes a series of events, in which humans exercise their agency to fulfil the roles of *forum* and *actor*. The environment in which they do so, as I have described previously throughout this chapter, is a fundamentally socio-technical one: as part of the socio-technical assemblage that makes up the algorithmic system that forum and actor are, simultaneously, *interacting with* and *confined by*, they find their agency in fulfilling their roles influenced by both social and technical aspects. As social cognitive theory teaches us, however, this influence is not a unidirectional: both the forum's and actor's actions and personal, internal factors also influence the environmental structures, be they social, technical or hybrid in nature. Bandura sums up this reciprocity between personal agency and (social) structure succinctly:

“Personal agency and social structure operate interdependently. Social structures are created by human activity, and sociostructural practices, in turn, impose constraints and provide enabling resources and opportunity structures for personal development and functioning.”

[24, p.15]

When assessing an algorithmic system's potential in terms of algorithmic accountability with the ultimate goal of improving the accountability processes enacted upon it, making assumptions and predictions about human behaviour and the personal, internal factors affecting human agency would require extreme generalizations and reductionist perspectives that are doomed to be disproven. In light of human diversity, no two combinations of *forum* and *actor* could be guaranteed to follow these predictions, making any suggested improvements aimed at increasing human agency through these internal or behavioural determinants unpredictable in their results at best, or futile attempts at controlling human nature at worst. Conversely, however, a techno-determinist perspective akin to the *mechanistic* agency model refuted by Bandura would be equally futile: as the environment in the *triadic reciprocal causation* model does not constitute the single, deterministic influencing factor, simply changing the environment (e.g., implementing certain technical measures) does not guarantee a successful outcome either.

These limitations notwithstanding, a path forward emerges between them. While any attempt at predicting human behaviour is as futile as controlling the accountability process by purely technical, external means, the *socio-structural environment* in which the accountability process takes place can be assessed and analysed in regards to the *constraints* and *enabling resources* it provides, as Bandura [24, p.15] formulated it. While insights gathered through this analysis will not necessarily yield actionable results that can *guarantee* success, it still is possible to adapt the environment by lowering *constraints* and improving upon these *enabling resources*. By doing so, we can increase the *potential* for various, alternative expressions of human agency in the accountability process, thus overall improving the chances that such a successful process will occur.

The most central and pervasive mechanism of personal agency is the *belief in self-efficacy*; Bandura deems it so important, in fact, that he declares it to be nothing less than *the foundation of human agency* [24, p.10]. His line of argumentation in this aspect is clear:

“Unless people believe they can produce desired results and forestall detrimental ones by their actions, they have little incentive to act or to persevere in the face of difficulties. Whatever other factors may operate as guides and motivators, they are rooted in the core belief that one has the power to produce effects by one’s actions.”

[24, p.10]

This observation directly relates to the algorithmic accountability process: If the humans involved in these processes encounter constraints and limitations in this process that they deem prohibitive to a successful conclusion, they will be unlikely to pursue their expression of agency. Identifying and assessing these prohibitive limitations thus must be the first, arguably unavoidable step towards ensuring *actors* and *fora* can retain their belief in self-efficacy as part of an accountability process. Simply put: if an algorithmic system does not provide the social, technical and socio-technical tools to plausibly allow a successful accountability process to occur, and if the human *actor* and *forum* in this process do not *believe they have the agency to do so*, the process will fail *a priori*. Even more concisely: *enabling resources* must be *visible* and *demonstrably useful*—and not just *available*—in order to be have a positive impact on human agency.

Bandura, almost prophetically in hindsight, already diagnosed these issues with modern technology in 2001, long before the current boom of ubiquitous algorithmic technologies:

“Everyday life is increasingly regulated by complex technologies that most people neither understand nor believe they can do much to influence. The very technologies they create to control their life environment paradoxically can become a constraining force that, in turn, controls how they think and behave. The social machinery of society is no less challenging.”

[24, p.17]

Algorithmic systems, in all their complexity, fit this diagnosis rather well, as they simultaneously come with the promise of control, yet often limit our agency at the same time.

2.4.5 Accountability, Moral Responsibility and Computing

When considering algorithmic accountability as defined in Section 2.4, the question of whether or not the technical components of an algorithmic system should be considered *actors*—specifically ones that *can* and *should* be held to account—inevitably throws a

wrench into the metaphorical cogs of any well-formulated reasoning on the matter. After all, the technical system’s behaviour may well be causally responsible—through intended or unintended action or inaction, through functioning as designed or as a consequence of an error—for a negative outcome. Particularly in the case of *micro-accountability* as described in Section 2.4.3—the direct interactions between human and non-human actors with a human forum—excluding the machine from accountability can lead to the aforementioned issues of using it as a *talisman* to ward off criticism, or employing *intentional opacity* to relegate responsibility to the otherwise inscrutable non-human actor. On the other hand, and in order to determine whether an algorithmic system could be held accountable at all, we must ask the question: is it *morally responsible*?

Excursus: Subjectivity and Responsibility in Assemblages

Returning to the conceptualization of algorithmic systems as socio-technical assemblages offers surprisingly little help in answering this question. On the contrary, the assertion of radical *relationality* and *heterogeneity* of assemblages suggest even more fundamental and complex questions about the nature of *human* involvement as part of algorithmic assemblages, in particular related to the questions of *subjectivity* and *responsibility* that requires further clarification.

Assemblage thinking and ANT—together with other schools of thought such as *flat ontology* and *new materialism*—are often attributed to the larger context of *post-humanist* thinking, as Häkli [220] describes. While otherwise a quite heterogeneous set of approaches, they are united in a shared “*radical skepticism toward the notion of the subject and subjectivity in the context of human (political) agency [...]*” [220, p.2]. Häkli, as a scholar in *political geography*, discusses the contradictions resulting from this stance in the context of citizenship and the civil society; his arguments, however, are easily adaptable towards human agency and moral responsibility in the context of algorithmic assemblages as well. The central question he poses can be summed up as this: if reality indeed is “*composed of emergent, endlessly evolving, and interconnected assemblages that interact through cascading relations between and across their constituent parts, humans and non-humans alike*” [220, p.6, summarizing [221]], and if human *reflexivity* and *intentionality* as defining aspects of human agency are to be abandoned in lieu of the human subject’s dissolution into the larger assemblages it is embedded in, how can we assign *responsibility* for the consequences of *distributed agency* of assemblages? In other words: if the agency of an assemblage is so fundamentally spread out across humans and non-humans, tracing any outcome to singular causes, such as intentional human action, becomes impossible, thus negating the idea of moral responsibility for such an outcome by any single component or indeed the entire assemblage [220]. Thus, a strict adherence to post-humanist readings of distributed agency offers no applicable answers to the moral responsibility of *non-human* components of assemblages, besides negating such responsibility for the *human* components more or less categorically as well.

To resolve these contradictions, Krause [222] offers a more nuanced approach. In her arguments, she describes the “*complex quality of individual agency as simultaneously*

distributed and singular” [222, p.316]. Human agency, in her view, must remain closely connected to *reflexive, norm-sensitive subjectivity*, to retain a sense of moral responsibility, but can be detached from *sovereignty* as the boundaries of human agency are also shaped by the larger assemblages they are part of. In this way, agency is both *singular* to the individual, and simultaneously *distributed* as the larger assemblage influences its extent and expression. The consequence of this perspective, she argues, is a broader scope of responsibility:

“The close connection between agency and (reflexive, norm-sensitive but non-sovereign) subjectivity allows us to call people to account for a wider range of outcomes than is possible on a view that narrowly equates agency with conscious intention and control. In this sense, the distributed, material approach to agency poses no threat to the concept of responsibility but rather enhances it. So if, on the one hand, individual responsibility frequently will be partial—or may often approach a complicity model rather than a full culpability one—on the other hand, responsibility will be more widely distributed than we think.”

[222, p.316]

Following Krause’s arguments, a less stringent and more nuanced conceptualization of algorithmic systems as socio-technical assemblages regarding the existence of subjectivity of its human components offers a plausible and more useful way forward. First, let us assert that human subjects, as they are part of (and interact with) an algorithmic assemblage, are capable of a different level of autonomy than the non-human components therein: their choices are informed by their ability to reflect on their actions and to contextualize them normatively, as opposed to non-human components of the assemblage. At the same time, as conceptualized in the previous Section through Bandura’s **SCT**, the agency of these non-human components also depends on their environment—specifically, the algorithmic assemblage in question—and is thus *distributed*. Moral responsibility for the outcome of an algorithmic assemblage’s exercise of distributed agency thus may be, at least partially, attributed according to the individual human contribution to the outcomes, as difficult as this attribution may be in many cases. To assess this contribution and to consider whether or not the non-human components of an algorithmic assemblage may, indeed, also carry such responsibility, we can then return to earlier, more traditional philosophical concepts of (human) moral responsibility.

Preconditions for Moral Responsibility

Moral responsibility as a concept “[...] *is intertwined and sometimes overlaps with notions of accountability, liability, blameworthiness, role-responsibility and causality.*”, as noted by Noorman [223]. By this approach, accountability then would refer to the

ability to ascribe moral responsibility to an agent's¹⁸ actions - to hold them accountable for the consequences of their actions or their intentions. Based on this definition, a review of moral responsibility in the context of computing is the foundation on which an exploration of algorithms as accountable actors must rest.

Moral responsibility in philosophy and, more specifically, normative ethics links *human action* to intentions and consequences. *Agents*—individuals or groups of humans—which exercise their *agency* by performing an action that affects a *patient*¹⁹ can be ascribed moral responsibility for their actions under certain circumstances. Depending on the philosophical school, the moral value of that conduct may be based on the consequences of that action alone—the *consequentialist* view [224, 225]—or, in the *deontological* view [226], on the intentions of the agent performing the action²⁰. In order to perform this judgment, a set of conditions generally agreed upon by moral philosophers [223] must be met: Firstly, the agent's *causal contribution* to the consequences of their actions must be established; holding an agent responsible for consequences their actions had no influence on is clearly not appropriate. Secondly, the agent's *knowledge and ability to consider the consequences* of their actions—often summarized as *reflexivity*—is a necessary precondition to carry moral responsibility for their acts. Lastly and perhaps most importantly, the agents in question must have had the *freedom to act* according to their choices or *sovereignty* over their actions—we generally do not hold people accountable for actions taken due to coercion or if they did not have a choice to act otherwise [223]. The overlaps between moral responsibility in philosophy and Bandura's *agency* in *SCT*—including its preconditions of *intentionality*, *forethought*, *self-reactiveness* and *self-reflectiveness*—are no coincidence, and indeed reveal the interconnected nature of the two concepts.

In the context of computing in general, and algorithmic systems in particular, a number of challenges to these preconditions arise. Nissenbaum [136] identified the *many hands* problem as a challenge to establishing a *causal connection* between an agent's action and their consequences: modern computer systems involve such a large number of agents—from requirements engineers, software developers, project managers and decision makers to administrators and users—that linking a single agent's actions to a given, non-desirable outcome often becomes all but impossible [227]. When conceptualizing them as socio-technical assemblages, the situation gets worse, as *non-human* “hands” must now be accounted for as well. Furthermore, users of such systems as moral agents are often separated from the consequences of their actions both temporally and physically

¹⁸While the previous sections adhered to Bovens' diction of identifying *actors* within the accountability processes, in moral philosophy, *agents* is a more prevalent term. For the purpose of this analysis, the terms can be deemed interchangeable; however, as this section focuses mainly on moral philosophy, the domain-specific term of *agent* will be used.

¹⁹To avoid misunderstandings, '*patient*' in moral philosophy refers to the humans affected by a moral action, and must not be confused with the term patient in a medical sense, i.e., a person receiving medical treatment.

²⁰In the spirit of brevity, this summary omits other normative ethical positions such as virtue ethics; however, as noted in Section 2.4, accountability could be interpreted as a *virtue*, in which case virtue ethics would play a role in the analysis.

by great distances, obscuring their contribution to a morally undesirable outcome further. To name but one, albeit particularly placative, example: military personnel working as remote drone operators utilize a variety of technical and algorithmic systems to execute acts of (semi-)automated warfare (i.e., drone strikes), sometimes geographically separated by vast distances from across the world. The automated targeting systems that support them in doing so have, themselves, been designed and implemented by multiple humans often years in advance. A hypothetical loss of civilian life due to such an algorithmically supported drone strike opens up a Pandora's box of questions relating to moral responsibility of hundreds, if not thousands of moral agents whose individual contributions to the final outcome may have been miniscule.

Algorithmic technologies can both improve and limit a human's ability to *consider the consequences* of their actions. On the one hand, simulations, remote vision or data analysis might provide more context and information to a user to make a sound judgment, but on the other hand, hidden automated processes within a given algorithmic system as well as the level of complexity involved in their operation might make it nigh impossible for the user to predict the effects of a given action. In the current climate of renewed interest in AI, tracing the impact of an agent's decisions (such as adding or removing data from machine learning systems or adjusting the fitness functions of a supervised learning system) to their eventual, automated outcomes remains an important and unsolved problem reflected in the large number of urgent calls for 'explainable AI' (e.g., [228, 229]; see Section 2.3.2 on algorithmic transparency challenges for more details). Additionally, the advancement of new digital technologies and their broad application has brought forth a whole new array of human action hitherto unseen: For instance, the act of uploading a video to the Internet made possible by the rise of social media is a new type of human action enacting moral values that cannot be easily qualified through our previously established moral theories [230, 231].

Furthermore, the use of computing technology in different contexts may impact human *agency* (as described in the previous section) in multiple ways. On one end of the spectrum, ADM systems severely limit human interaction intentionally: fully automated systems may replace human decisions all together, giving rise to complex questions of moral responsibility and liability in relation to the causal impact of choices made by the actors that designed, developed, tested, deployed, or maintained that system [232]. Systems designed to augment human choices, such as the COMPAS criminal recidivism risk scoring system [233, 234], operate in a grey area in terms of human freedom of choice, allowing the users the binary choice to either heed or negate the system's output, but may not necessarily afford them any reasonable granularity of choice beyond that. While this technically *does* leave the choice with the human user, the system's design still controls the flow of information and influences the user's choice, which often manifests in a human tendency to trust machine decisions referred to as *automation bias* or *complacency* [235, 236, 237], and gives rise to the issues of meaningless human agency as described in Section 2.4.2.1 [177].

All of these arguments share a common assumption: that the agent of moral action and thus the entity to be held accountable is and remains a *human* exercising their agency. By and large, computers and algorithmic systems are not considered *moral* agents themselves, making it impossible to hold the system itself *morally* accountable for its actions, decisions and consequences of these actions. Even though post-humanist schools of thought such as assemblage thinking or ANT assert a distributed agency of human and non-human actants, they do not assert a moral quality to this agency, leading critics to point out the “*distinction between an agent and a mere cause*” [222, p.310]. The main arguments why computers cannot be considered moral agents are focused on a lack of abilities or certain human attributes of these systems. For one, they lack some of the things that make humans moral agents, such as *mental states*, *intentionality* of their actions, *common sense* or *emotions*. A different aspect of moral responsibility is the questions of punishment: since computers cannot suffer, they cannot be punished for their *actions* as part of being held accountable—yet Bovens’ and Wieringa’s established definition for the accountability process as explicated in sections 2.5 and 2.4.2 presupposes precisely that potential for consequences or sanctions, which can also fulfil the function of a punishment or deterrent. Lastly, they lack the ability to consider the consequences of their actions in a moral sense, since they lack the capabilities for moral reasoning or *reflexivity* [223, 222].

While few would contest these notions outright, bridging the gap between humans as the sole moral agents and algorithmic systems as purely amoral tools of human activity could offer some reasonable improvements to the process of ascribing moral responsibility and, subsequently, improving algorithmic accountability. One approach put forward by Floridi and Sanders introduces the concept of ‘*mind-less morality*’ and proposes to expand the concept of *moral agency* by extending the class of moral agents to include *artificial agents* [158]. Their argument is essentially to disconnect the concepts of *moral agency* and *accountability* from *moral responsibility*: while an artificial agent could be held *accountable* for its actions and decisions, it would not be held *responsible*. They compare this approach to the way we tend to treat certain animals as agents of morally charged actions (e.g., scenarios such as “an alpine avalanche rescue dog saves a human” or “a wolf kills some livestock”), yet generally don’t attribute moral responsibility to their actions. This approach depends on different *levels of abstraction* from which artificial agents can be conceived as performing moral action: while a low level of abstraction would only be suitable to describe a system through its components and biological, logical or mechanical processes, a higher level of abstraction would introduce thoughts, beliefs and desires into the description. To apply this to an artificial entity like a computer system, Floridi and Sanders contend such a system must be *interactive*, *autonomous* and *adaptive*: “*It, thus, does not require personhood or free will for an agent to be morally accountable; rather the agent has to act as if it had intentions and was able to make choices.*” [223]. Through disconnecting agency and accountability from responsibility, determining the exact human agent responsible for a given system’s behaviour, and attributing what partial or proportional share of moral responsibility they would have to bear, is no longer a necessity to attribute *moral agency* and accountability, thus

allowing direct intervention in the system itself as a moral response to its actions (in the form of modifying or even disassembling/deleting it). Remarkably, this ontological trick of separation between *moral agency* and *moral responsibility* also works well within assemblage thinking, seeing as both non-human components or the entire socio-technical assemblage may well exercise their agency in a morally charged way, but only the human facets of that assemblage may be held morally responsibly for the outcome.

Although Floridi and Sander's proposition resolves some of the issues regarding moral responsibility, moral agency and accountability, critics have pointed out that it shifts the attention away from the human creators and operators of artificial agents. Some scholars have proposed different views that do not attribute complete moral autonomy to artificial agents or computer systems, but rather connect them to their users and creators through *intentionality* and the *values* inscribed in them [238, 239]. Others include computer systems as *mediating artefacts* in the moral agency of human actors (both creators and users), maintaining that such mediation represents a different kind of moral agency: one that is, nevertheless, still an integral part of the greater moral agency that makes up this "[...] complex blend of humans and technologies" [223].

Applying this approach of separating moral responsibility from moral agency to the accountability process as defined by Bovens [22] yields some interesting consequences worthy of consideration. For instance, including algorithmic systems themselves or the technical components of algorithmic assemblages as potential *actors* that could be held to account would theoretically allow for an accountability process involving only one human (as the forum), posing questions and judging the response, and potentially imposing consequences on the system. For such an interaction to be plausible, the system in question must be able to respond to some forms of inquiry, be it through natural language processing and speech synthesis, or by providing specific interfaces for inquiries in a more structured form. Some of the solutions developed in the field of XAI already follow a similar pattern, where reasoning for a given output can be provided on request by a user, who may or may not deem the explanation satisfactory. Mechanisms for imposing consequences in such a relationship would remain solely in the realm of constructive feedback, for instance by formulating a request to delete certain data used in the calculation, or by escalating a decision to a human agent supervising the system. Sanctions as a form of consequences make little sense in this case, as explicated above: deterrence or the fear of punishment are meaningless concepts in the context of a non-human moral agent. Besides these limitations regarding potential consequences, an accountability process between human and non-human actor would be limited mostly by the systems' design. In other words, for an algorithmic system itself to be considered accountable to some extent, the potential for such an accountability process must be part of its design specification and implementation as well, and—given the limitless nature of potential human inquiry—could only ever cover the most common requests or inquiries. In other words, if the developers of the system did not conceive of a potential problem requiring an accountability process *a priori*, the system as an actor will not be able to be held to account for that problem either.

Following the classification in *macro-* and *micro-accountability* suggested previously, I propose the term *artificial accountability* to denote this limited capacity of algorithmic systems for accountability processes, as it is a type of accountability that must be created by humans to artificially replicate a natural accountability exchange. Given the arguments made above one may, at first glance, consider *artificial accountability* barely worth the effort. And yet, when applied to existing and real-world systems, we already see examples of human and non-human agent exchanges that could be seen as such a form of limited *artificial accountability*. Consider this case study of targeted advertising algorithms [240] and their explanations: Following calls for increased transparency of the ad selection process, online advertising platforms implemented functionalities allowing users to request explanations for the ads they were seeing, accessible through “Why am I seeing this ad?” buttons. The fact that these buttons are often intentionally obscure and hard to find through the use of *dark patterns*²¹ notwithstanding, the functionalities provided can be conceptualized as a form of *artificial accountability* process. The human user (as the *forum*) is posing a request to the Online Behavioural Advertising (OBA) system (the *actor*) to justify the decisions made for the ads the user is seeing (the *conduct*). Upon receiving the explanation, the user can impose (albeit very limited) *consequences* by reporting the ad as inappropriate, or accept the explanation as satisfactory. Until the point of imposing consequences, this process is entirely reliant on an automated mechanism and involves immediate human actors.

The example given above also serves as an illustration of the potential dangers of relegating accountability processes to artificial agents. As online platforms are struggling to respond to calls for better content moderation either due to unwillingness or technical limitations, it is entirely conceivable that conceptualizing these processes as a type of accountability serves as an excuse to not invest in proper *micro-* or *macro-accountability* processes. If an organisation can implement such an automated accountability process for an algorithmic system they developed, and subsequently claim to have fulfilled its responsibilities, they also could exert undue influence on the shape and form of the questions that *can* be asked by the *forum* as well as the *consequences* that forum can impose, and make this kind of *artificial accountability* entirely toothless by restricting its potential to effect meaningful change. In doing so, *artificial accountability* processes may well be misappropriated as a novel form of *ethics washing* [169]. That being said, there may still exist the potential for meaningful *artificial accountability*, given that it is part of a larger, compound process of *micro-* and *macro-accountability*. In order to fulfil that potential, such a process would need to be escalated to higher forms of accountability once it becomes clear that neither available options for *inquiry*, *account* or *consequences* are sufficient; the agency to trigger such an escalation would need to lie with the human forum.

Seeing as the current regulatory landscapes do not incorporate such requirements at all, current implementations of *artificial accountability* should not be considered sufficient replacements for *micro-* and *macro-accountability* processes.

²¹For a more detailed analysis of the use of dark patterns, see my work on human-centric perspectives on online consenting [9].

2.5 The Wicked Nature of Accountability and Transparency

The overview of algorithmic accountability given in the previous sections must, by definition of the problem, remain incomplete. The complexity of algorithmic system development, the variety of potential actors and fora involved, and the fluid nature of what constitutes plausible inquiries and accounts, as well as reasonable consequences with the potential to effect positive change, makes a complete and comprehensive guide to algorithmic accountability—one that is applicable to any and all situations—an impossibility. To reconcile these issues, it is helpful to consider the *nature* of the problem of algorithmic accountability—and, subsequently, algorithmic transparency as a *wicked* one.

Originally coined by Rittel and Webber [30] in 1973, the term *wicked problem* denote a class of problems mostly situated in the social realm that are particularly challenging to address. As Fitzpatrick [241] describes, the first major complication when addressing a wicked problem lies already in pinpointing *what exactly the problem is*, or, in other words, the fact that “[t]here is no definitive formulation of a wicked problem” [241, p.4]. Both algorithmic accountability and transparency show this characteristic: regardless of how many cases of lacking algorithmic accountability or transparency one might analyse and describe, another system, in another context, based on different technologies or developed for a different purpose, will inevitably face different variations or altogether new challenges. In fact, a mere analysis of accountability or transparency as a social problem may not even yield particularly valuable insights until potential solutions are being developed, tested and evaluated; leading to a progressively better, yet never complete understanding of the overall problem itself. Consequentially, as both problem definitions and (imperfect) solutions addressing them are fluid and evolving in parallel, a clear ‘stopping rule’ defining when the problem is ‘solved’ cannot be derived *a priori* either, and no enumeration of the solution space is possible [241, p.4].

As possible measures (be they technical, socio-technical, policy-based or other) are being developed and implemented, a new understanding of the nature of accountability and transparency for that specific instance will emerge that requires further adaptation and re-evaluation of the solutions previously derived. For instance, seeing as accountability processes for a given system may be multi-faceted depending on the forum or time within the software development cycle they are invoked, learnings for accountability requirements during the *ex post* phase of deployment can be derived during accountability processes in the prior *ex ante* and *in medias res* phases. Similarly, a general understanding that a system is lacking in terms of transparency may only give way to more specific requirements of transparency once the first solutions have been tried and evaluated as dissatisfactory.

This characteristic of wicked problems points to the cyclical, iterative nature any plausible approach to developing solutions must take, or, as Fitzpatrick formulates it:

“The process of solving a wicked problem is inherently non-linear. Progress is defined qualitatively in terms of how much more is understood about the problem rather than distance from the solution.”

[241, p.4]

It follows that solutions can never be evaluated in a binary fashion as either *true* or *false*, but at best qualitatively compared to previous solutions or the status quo in terms of how much they improved the situation. As there can be no *“immediate, ultimate, or definitive test”* [241, p.4] to do so, determining exactly when the problem should be considered as solved is impossible as well; at best, one can aim to find an acceptable solution to the problem that is “satisficing”, a clever portmanteau of the terms *satisfying* and *sufficient*. Even so, wicked problems are also characterized by the involvement of multiple stakeholders, who each bring their own perspectives to the table: Be that differing views on the problem formulation or on the merit of a given solution, conflicting qualitative judgments are to be expected, and what may be *satisficing* to one stakeholder may not be seen as such by another.

Given these considerations, one might conclude that addressing the wicked problems of algorithmic accountability and transparency is nigh impossible. At the core of this dissertation, however, stands the conviction to the opposite: given the right approaches, taking into account different perspectives of stakeholders, and not expecting *one-size-fits-all* solutions that often turn out to be *one-size-fails-all* solutions for wicked problems, it is possible to improve accountability processes and address issues of transparency in specific cases. To generalize these results into learnings for future solutions, I posit that, for wicked problems, the measures themselves are not necessarily the most relevant insights, but rather, the *ways by which the measures were developed*. In essence, studying and improving accountability processes may yield as many insights into methodology and analytical frames as it does into concrete solutions.

2.6 Chapter Summary

In this chapter, I have outlined the theoretical foundations that informed this dissertation, its general methodological approach, the design and analysis of the case studies, and the synthesis of the **A³ framework**. In an attempt to bridge the gap between the different, yet equally relevant disciplines for this dissertation, I addressed the *terminological anxiety* identified by Seaver [21] through presenting the different conceptualizations of the terms *algorithm* and *algorithmic system*. I traced the term from its historic roots and its technical meanings towards broader definitions involving the *social* as well as the *technical*, and finally conceptualized algorithmic systems through the eyes of **ANT** and assemblage thinking, as well as what *functions* they fulfil in the world. In doing so, algorithms were

revealed to be characterized by their *relationality*, *productivity* and *heterogeneity*, and perhaps above all, *multiplicity*: an algorithm or algorithmic system is never *singular*, and always constituted by *enactment* through human action and interpretation.

These abstract perspectives notwithstanding, algorithmic systems have real and tangible impacts on society at large and individual humans alike, as exemplified by their potential for *bias* and *discrimination* as a consequence of exercising their power in the world. To assess and control both their influence and impacts means facing significant challenges, in particular the questions of *algorithmic transparency*, *explainability*, and *(in-)scrutability*.

Finally, at the core of this chapter and, indeed, the central theme of this dissertation, lies the question of *algorithmic accountability*. By drawing on previous, conceptual work by scholars of *public accountability*, I discussed *algorithmic accountability* as the *relational* and *procedural* interaction between a *forum* and an *actor*, centred around the *obligation* to provide a *justification* for the actor's *conduct*, with the potential for the *forum* to impose consequences in response. I then introduced a new distinction for algorithmic accountability based on its *scope* by juxtaposing the under-researched dimension of *micro-accountability* against the larger backdrop of *macro-accountability*. As a more immediate form of accountability that occurs between individual humans rather than large organisations and abstract institutions, *micro-accountability* is certainly deserving of our attention if our commitment to *human-centricity* in the design and application of algorithmic systems is to be taken seriously.

Based on this commitment to human-centricity for algorithmic accountability, I introduced Bandura's Social Cognitive Theory (SCT) as theoretical model of human agency and the foundation for a deeper analysis of human agency in the context of algorithmic accountability processes. Human *emergent interactive agency*, as described by Bandura, manifests as the *reciprocal triadic causation* between *personal and cognitive factors*, *human action and behaviour*, and the *external environment*, all of which influence a human actor's potential and willingness to act in a given situation. Singling out the human actor's *belief in self-efficacy* as one of the most central mechanisms of personal agency, I exemplified the dire implications a lack of this belief may have on the accountability process *a priori*, and argued that *enabling resources* must be both *visible* and *demonstrably useful* in order to support and enhance human agency as part of the accountability process.

The discussion of *algorithmic accountability* in its procedural form and the subsequent arguments for a human-centric approach leads to the difficult philosophical considerations of moral responsibility in the context of non-human actors, including algorithmic systems. In a short excursion, I briefly presented the existing arguments for and against considering non-human actors as moral agents, and related the proposal by Floridi and Sanders [158] to separate *moral responsibility* as uniquely human capacity from *moral agency*. Incorporating this theoretical approach into the practical considerations necessary for resolving whether or not an algorithmic system alone could be considered an *actor* in an accountability process, I proposed the term *artificial accountability* for such a relationship. To illustrate the worth of this perspective, I described the example of automated explanations for targeted ad placements as an artificial accountability process

occurring between a human *forum* and a non-human *actor*. The short excursion concluded with the warning that, as a consequence of the limitations of such processes and the blatant lack of regulations that could enforce the requirements needed to protect human agency in such a process, *artificial accountability* should in no way be considered a sufficient replacement for human *micro-* and *macro-accountability* processes.

Finally, the chapter concluded with a ‘big-picture’ characterization of algorithmic accountability as a *wicked* problem, and the particular challenges that this class of problems poses. This discussion also served as a careful introduction of the limits both this chapter and this dissertation as a whole faces in terms of ‘solving’ the problem: as wicked problems lack both a *definitive formulation* and *definitive solution*, addressing them can only succeed through cyclical, iterative improvements that must be evaluated qualitatively against the previous solutions rather than against an unreachable and, indeed, undefinable ultimate goal. At best, *wicked* problems such as algorithmic accountability or transparency can be addressed with measures that prove to be, in the words of Fitzpatrick, “*satisficing*” [241, p.4].

2.7 Chapter Conclusions

At the onset of this chapter, I set out to address the “terminological anxiety” Seaver [21] diagnoses to be prevalent in CAS when it comes to defining “*algorithm*” and “*algorithmic system*”. Not just an exercise to appease the nagging discomfort of dealing with an amorphous and unclear term, answering SRQ1 yields some valuable insights that help guide further research. Discussing narrower technical definitions serves a purpose beyond just tracing the terms to their historic roots: it also makes it painfully obvious that humans and social aspects are entirely absent from these definitions, leaving these conceptualizations unsuited for studying how algorithms fit into the world and interact with it. Algorithms as *socio-technical systems* introduces the ‘social’, but much of the theoretical roots behind the concept are rather limiting due to its original scope of being part of an *enterprise*. As we can observe *algorithmic systems* flourish in so many more societal contexts, asking questions of power relations between humans and aspects of technologies requires expanding our view: hence the introduction of algorithms as *socio-technical assemblages* or *actor-networks*.

Concerning Trends in Automation

Considering algorithmic systems as socio-technical assemblages of ‘agentic’ human and non-human components reveals an underlying, rather concerning techno-deterministic quality of much of today’s technology development, namely the fundamental changes it introduces to the balance of power between human and non-human components. The larger trend towards automation in general exemplifies this shift particularly well: after all, what is automation if not a shift in power from human agency towards machine agency? Proponents of automation often argue that the ends justify the means here: handing over tasks to automation that humans do not enjoy (e.g., repetitive, mind-numbing tasks)

should be considered a boon for humanity, since it supposedly leaves us humans to do better, more fulfilling tasks ourselves. Some rather obvious flaws in this argumentation exist, mostly related to the existing inequalities of modern society and the logics of capitalist production, growth and labour: as a result of automating one sector, who is to say our newly won free time is not just shifted towards those *other* mind-numbing and repetitive tasks that the machines or algorithmic systems can not (yet) do? These flaws aside, the larger underlying problem with this argumentation is, I would argue, that we are not just ridding ourselves of onerous tasks, we also further reduce our agency as human beings.

With algorithmic systems in mind, this trend towards automation becomes particularly explicit. Developing new algorithmic systems—or making existing ones more sophisticated, more powerful and more capable—is implicitly a process of empowerment for the technological and, at best, socio-technical aspects of an assemblage, yet can be restricting and limiting for its social components. In other words, advancements that are aimed at improving the entire system’s agency—to make it more capable, to allow it to do more, and faster—often come at the cost of human agency, by shifting it towards the non-human components’ agency. This pattern notably conflicts with the common promise of technology to empower humans: while technology in general may enable us to do things we can not do on our own, *algorithmic* technologies also limit our choices of *what it is we can do*, and our understanding of *what exactly it is they do*. Technically opaque systems, be that as a consequence of the inherent qualities of the involved AI/ML technologies, or simply due to their size, complexity or distributed nature, exemplify how ever more sophisticated technological components leave humans struggling to comprehend their inner workings, and thus often leave them no further choice than to either trust the system or attempt to detach themselves from its socio-technical assemblage. Purposefully opaque systems exemplify a similar shift in power and agency, although the intentional nature of secrecy often serves to empower product owners rather than technology itself: black-boxed systems, after all, may also hide the banal simplicity of underlying technologies, or even deceptive practices of quasi-automation by outsourcing difficult-to-automate tasks to cheap and exploitative human labour.

This shift in power does not stop at how much agency humans working directly in concert with algorithmic technologies have. At the same time that *humans-in-the-loop* are relegated towards *humans-on-the-loop* or *humans-out-of-the-loop* altogether, we also struggle more and more to find the agency to hold these systems of automation to account. The need for accountability of algorithmic systems is easily demonstrated by the numerous examples of problematic case studies showcasing systems fatally flawed by bias, thus exercising their agency to discriminatory effects. Hence, the need to study and explore algorithmic accountability to find new ways of holding those socio-technical assemblages to account.

From Public to Algorithmic Accountability

As these answers to [SRQ1.1](#), [SRQ1.2](#), [SRQ1.4](#) and [SRQ1.5](#) show, accountability is a process that, like all processes, can succeed or fail depending on a myriad of factors. If we want to hold algorithmic systems to account successfully, we have to purposefully design and consider how a given system can support such processes when operationalized in a specific context. In other words, it is not enough to adapt the technical parts of an algorithmic system to be, in principal, capable of supporting such accountability processes, but we also must consider how the various assemblages resulting from the deployment of the system will enable them. To this end, turning to public accountability theory offers valuable learnings stemming from the long history of struggle to establish the rules, frameworks, guidelines, laws, regulation and fora that deal with accountability in other contexts. Integrating insights from those disciplines that have extensively studied such macro-accountability processes is a must, if we are to avoid re-inventing the wheel over and over again.

Adopting the concepts gleaned from public accountability to algorithmic systems, however, leads to the answers for [SRQ1.3](#) by introducing *micro-accountability* as a variant, and possible precursor, to *macro-accountability* processes. Beyond the more detailed discussion of *micro-accountability* as a concept and the forms, fora, actors, accounts and consequences it might take, a larger point about its necessity can be made. Considering the nature of algorithmic systems and the shifts in power relations they introduce as described above, a reliance on macro-accountability processes for algorithmic systems is simply not enough to ensure just and fair outcomes and avoid future harm. The problem with assuming only groups of people or abstract and impersonal organisations can plausibly fulfil the role of the forum in accountability processes is that it, implicitly, prevents direct action and immediate reaction by those affected. The ability to react to an algorithmic system exercising its agency more immediately, however, becomes crucial in our struggle to close the gap between the speed at which automation moves and the slow-moving pace of human society, bureaucracy and its institutions. Without automation, the acts of exercising power and agency of, e.g., bureaucratic systems are as slow (if not slower) than the humans affected by those actions, giving them more time for recourse or to organize into larger fora that could initiate macro-accountability processes. Case in point: the slow-moving pace of the Austrian administrative system was and still is, in many cases, the only chance vulnerable groups (e.g., asylum seekers) have to react to decisions they deem unfair, and seek help or organize into larger groups. A hypothetical system automating that process would—besides the horrendous outlook of bias and discrimination codified into algorithms—also significantly lower the chances of affected stakeholders to hold the government to account. Similarly, the Allegheny Family Screening Tool for child welfare risk assessments, as Eubanks [\[111\]](#) describes it, produces predictive risk scores that can trigger child abuse and neglect investigations. A lengthy macro-accountability process may well prove that parents were falsely marked as high-risk, but at that time, their children may already have been separated from them, suffering the psychological impact of that decision. Both hypothetical and real examples illustrate

how the speed at which automated systems work makes it necessary to, in addition to reliable *macro-accountability* processes, also strengthen individual stakeholders' agency as fora to demand *micro-accountability*, and react in a more immediate way than the *macro-accountability* processes would allow.

Towards Better Algorithmic Accountability

That is not to say that a focus on such a “*forum of one*” is the only way to adapt the accountability process to algorithmic systems: not all (systemic) problems introduced by automation in general or algorithmic systems in particular can be addressed by introducing or improving micro-accountability processes. Micro-accountability processes by that or any other name are, however, currently not given much attention in the mainstream of algorithmic accountability research, and thus offers significant potential for improvements. The reasons for this lack of attention may lie in the fact that, for many algorithmic systems, micro-accountability processes are difficult to implement, regulate, or—perhaps—even imagine. But at the same time, similar arguments could be made about the foundations of modern macro-accountability measures and processes, including labour laws or consumer protection regulation, before collective fora (e.g., unions) fought for them. Thus, *micro-accountability* as a *virtue* can also serve as an aspirational goal to initiate a broader societal debate of what powers and rights individuals working with or affected by algorithmic systems ought to be granted. Among others, a discussion of what constitutes *appropriate consequences* as part of algorithmic accountability processes, including both those levied by affected single individuals and those imposed by larger fora, will undoubtedly be necessary.

More immediately pressing, however, is the question of how to move forward given the complexity of the problems discussed in this chapter, and thus the overall answer to [SRQ1.6](#). As the framing of algorithmic accountability and transparency as wicked problems suggest, iterative improvements may be the most promising and, indeed, the only plausible approach to tackle these issues. To this end, empiric studies founded on these theoretical considerations—such as the case studies presented in this dissertation—are our best path forward. Considering the fundamental questions that remain unanswered, it is important to take an unflinching look not just at the shiny, overhyped and bleeding-edge *buzzword-du-jour* technologies, but specifically at those comparably banal, everyday and—perhaps—a little “*boring*” [\[242\]](#) systems that already permeate our world. In the spirit of “*learning to walk before attempting to run*”, it makes sense to first figure out what to do with our explanations before attempting to explain the “un-explainable” systems, and to learn how to hold the *simple* systems to account, before tackling the *complex*.

Methodology

The following chapter describes the methodologies of inquiry utilized for both case studies described in Chapters 4 and 5, as well as the comparative case study that is the foundation for the synthesis of the Algorithmic Accountability Agency Framework (A³ framework) described in Chapter 6. Since some methodologies are overlapping between the different empiric and theoretical forms of inquiry, and in order to avoid duplication and superfluous replication, this combined description is preferable to describing the methodologies in detail in their respective chapters.

As an additional methodological disclaimer, it should be noted that, for both of the case studies and otherwise throughout this dissertation, a number of sources, references, quotations or terms have been translated from their original German to English. Where such a translation occurred, the original German text is always referenced in a footnote, prefaced with the shorthand “*Orig.*”. Unless stated otherwise, all translations were performed by myself.

3.1 Overall Approach

As Chapter 2 explicates in some detail, for the purpose of analysis in this dissertation, algorithmic systems are conceptualized as complex socio-technical assemblages. To gain the holistic understanding necessary for this analysis, a variety of methodologies from the inter-disciplinary contexts covered by CAS have been employed in accordance with the arguments for such an approach presented in Section 1.2.2. As such, the selection of methodologies followed these guiding principles:

- Holistic capturing of social, technical and socio-technical aspects
- Complementary balance between approaches

- Plausibility of methodologies given research constraints

Primarily, the case studies require an analysis of the social, technical and socio-technical aspects of the algorithmic systems in question in order to address the research questions posed in Section 1.3. This includes gaining an understanding of *which* technologies are employed, *how* they are configured, *what informed* their design and *which stakeholders* were influential in these decisions, as well as how the resulting system is operationalized in practice. These questions are strongly interdependent: any analysis solely focusing on one of these aspects would have invariably fallen short of capturing the system as a whole and prohibit the unpacking of their socio-technical assemblage.

Second, the choice and intensity of the methodologies of inquiry employed must remain carefully balanced to complement their respective limitations. Burrell [127] highlights this precarious balance in her description of the approaches taken by law and social science scholars. While their critique of algorithmic systems with a strong focus on ‘algorithms in the wild’ from a broad socio-technical perspective may yield important insights on the big picture of these systems’ impacts on the contexts they are deployed in, Burrell also points out that such a contextually situated analysis may not help surface the broad patterns or risks related to certain classes of algorithms [127, p.3]. In other words, there exists a trade-off between the *specificity* of an analysis for a given context, and the *generalizability* for other contexts of the insights gained. Likewise, focusing strongly on the technical figuration of a given class of systems (e.g., *recommender systems* or *statistical profiling systems*) runs the risk of producing supposedly generalizable insights that, upon closer evaluation of a different case study in a different context, are easily contradicted again. To avoid this pitfall, the methodologies chosen for the case studies aim to balance both perspectives, considering both the specifics of their context and the general characteristics of the technologies employed. The comparative approach of juxtaposing the two case studies (which may, at first glance, seem almost incomparable in their different technological foundations) helps in keeping that balance and spotlighting the limitations by the methodologies used in the respective case studies.

Finally, any empiric inquiry involving fieldwork is implicitly limited by the opportunity of access to the system itself and the willingness to cooperate by the system’s stakeholders. Thus, the methodologies employed reflect these limitations and seek to circumvent them through alternative approaches where possible.

3.1.1 Case Study Selection

As the previous Chapter 2 describes in some detail, the landscape of algorithmic systems—from bleeding edge, prototypical, systems to current, state-of-the-art applications and legacy systems of the past—is vast and diverse, covering a wide range of technologies, fields, application contexts and scales. Given this diverse landscape, the *a priori* reasoning behind the choice for these two, specific case studies requires some explanation.

For *EnerCoach*, three primary reasons informed the decision to undertake the study as outlined below: the stakeholder's stated acknowledgement of deficiencies in transparency, the chance for virtually unlimited access to almost the entire socio-technical assemblage, and the potential for the use of non-traditional approaches to designing improvements through participatory design. First and foremost, identifying that there was a problem with this system regarding its transparency happening not through external critique, but by the users, developers (including myself), decision makers and other stakeholders of the system. These stakeholders knew something was wrong, but could neither identify what exactly was the problem, nor how to approach a solution. As a developer of the system, I had a personal history with the stakeholders, and saw that many of them had parts of the puzzle pieces, that a concerted research and development effort would be necessary to join them into a coherent picture. In contrast to other algorithmic systems, this involvement gave me a head start, both in terms of identifying the actors holding the puzzle pieces, and getting them on board to participate in this research project. Secondly, the opportunity to holistically analyse an algorithmic system of significant scale with this kind of unfettered access to all parts of its assemblage allowed for the kind of detail-oriented analysis researchers in [CAS](#) often can only dream of: finally, this case study would allow combining code, practice and intention to answer all the questions that, with many other, proprietary or otherwise inaccessible systems, often required guesswork to approach. Third, the stakeholder's apparent willingness to invest time and effort into tackling this complex problem they themselves identified also meant it was possible to employ non-traditional design methodologies in the form of participatory design. While the benefits of participation are well-documented, the restrictions many commercial enterprises introduce when it comes to developing new technology often limit the potential of these methodologies, or outright prohibit their use. Thus, *EnerCoach* presented a unique opportunity for study: the involved parties knew there was a problem, they were willing to open up the system completely as a researcher, and they were willing to engage with it through scientific methodologies and to invest time and effort to find solutions. Based on these factors, the decision to pursue the case study as promising was an easy one to make.

By the time the [AMAS](#) system was announced and my engagement with it in a research capacity started, the decision to pursue the *EnerCoach* case study had already been made, an in-person kick-off meeting with the *EnerCoach* working group as the primary point of contact for the study had already happened, and the initial interviews had been recorded. Compared to the rather positive outlook for *EnerCoach* and the measured pace of the case study progression, engaging with [AMAS](#) felt urgent due to its potential for negative consequences and the immediate nature of the public controversy and discourse that followed the initial announcement of the system. Thus, as a potential case study, [AMAS](#) presented with diametrically opposed characteristics compared to *EnerCoach*: the affected jobseekers were quickly identified as a vulnerable group that might be further disadvantaged by the use of this system; the system's developers and decision makers were, at best, defensive and at worst, hostile towards our initial attempts at dialogue and saw nothing wrong with the system; and access to the system, its underlying technology

and data, or access to the stakeholders for research purposes seemed out of the question. And yet, the **AMAS** system's context of application in public administration made it intriguing as a case study exemplifying a larger trend towards **ADM** or **ADS**, and the public controversy its announcement caused provided a short window to utilize public pressure to gain some further access and commence a research project. Finally, the announcement of such an obviously problematic system seemed like a serendipitous opportunity: having taught a seminar about the issues and challenges introduced by algorithmic systems for some years with numerous high-profile, but international examples, the chance to analyse a case study that was locally situated in Austria was too good to pass up.

As both the EnerCoach and **AMAS** case studies progressed—EnerCoach with a continued focus on algorithmic transparency as the core challenge, and **AMAS** with a broader perspective on various issues of bias, discrimination, policy, transparency in the context of automation in public administration, the common issue of algorithmic accountability arose. As it became clear that both systems would continue to produce questionable results regardless of how they might be developed further, the need for an algorithmic accountability for these results became clear as well. Implicitly, the fact that neither of these systems were particularly accountable to begin with was quite obvious, but to determine specifically *why* they were not, and ideally, *what* to do about it, remained unclear. Thus, the focus my research as a whole, as well as the focus of my dissertation shifted away from algorithmic transparency, bias, fairness and discrimination towards algorithmic accountability as an overarching topic. As I considered both research projects' suitability as case studies for algorithmic accountability, their differences became more of an intriguing challenge than just a hindrance. Analytic tools for algorithmic accountability which were versatile enough to cover these seemingly different cases also had a better chance of being applicable to other types of systems and interactions as well—certainly more so than tools created based on a set of very similar case studies. Alternatively, the case studies might have resulted in vastly or fundamentally different results describing completely different kinds of accountability challenges, which might have suggested the need for a more nuanced taxonomy of algorithmic accountability and more specific tools. In either way, there was more to be gained from studying different systems that I already had access to and knowledge about than attempting to find more similar case studies, cementing the choice to include and integrate both research projects as case studies in this dissertation.

3.2 Case Study Methodologies: EnerCoach

The case study of the EnerCoach system [1, 2] was conducted in two phases. Starting with an *analysis* in the form of a situated ethnography [21], the first phase yielded a thick description [27] of the system itself, its stakeholders, and the *status quo* of the system in regards to its transparency and accountability as identified by the stakeholders. The second phase consisted of an *interventionist* approach to address the issues and challenges identified in phase one through *participatory design* methodologies, with the

ultimate goal to implement and evaluate specific socio-technical measures that would improve the system's transparency and accountability.

The analytic phase is founded in Seaver's [21] conceptual approach of a situated ethnography of algorithmic systems *as culture*, as opposed to understanding algorithms as technical artefacts situated *in culture*. Seaver argues, essentially, that—in order to overcome CAS's defining feature of *terminological anxiety*—we currently approach the study of algorithmic systems by conceptualizing the term *algorithm* as an *emic* term: depending on who we ask, we get different answers describing the same (technical or socio-technical) artefact. This approach is thus defined by the social boundaries “*between algorithm people and other “technical people,” and between technical people and their non-technical others, who may not understand the definitions in play.*” [21, p.3]. In this sense, algorithms are not *culture* by themselves, but are situated within, and interact with, *culture*:

“Understood as such, algorithms themselves are not culture. They may shape culture (by altering the flows of cultural material), and they may be shaped by culture (by embodying the biases of their creators), but this relationship is like that between a rock and the stream it is sitting in: the rock is not part of the stream, though the stream may jostle and erode it and the rock may produce ripples and eddies in the stream. In this view, algorithms can affect culture and culture can affect algorithms because they are distinct.”

[21, p.4]

This view of algorithms as *distinct* from the cultural context surrounding them is problematic insofar as it shifts much of the focus onto definitorial issues of *what is and is not* part of the system in its quest to arrive at a final and stable demarcation of the artefact ‘algorithm’. To address this issue, Seaver suggests a different approach, founded in theories of anthropology, philosophy and ethnography. From his perspective, algorithms should be seen “[...] *not as stable objects interacted with from many perspectives, but as the manifold consequences of a variety of human practices.*” [21, p.4]. They are never singular, always *multiples*, and fundamentally unstable as they constitute themselves by being *enacted* as a consequence of human practices. This description also fits well with the conceptualization of algorithms as socio-technical assemblages introduced in Section 2.1.3, giving further merit to Seaver's approaches. Characterized as such, the study of specific algorithms *as culture* takes on the character of a situated ethnography.

The following sections detail the practical implementation of this ethnographic approach, starting with the disclosure of my auto-ethnographic role as a researcher and co-developer of the EnerCoach system. Based on this appropriation of ethnographic methods, the subsequent sections outline the data sources utilized, their modes of analysis, and finally, the theoretical and practical implementation of the interventionist parts of this case study in the form of *participatory design* methodologies.

3.2.1 Auto-ethnographic Considerations & Disclosure

As a researcher in [CAS](#), I have been engaged in the study of algorithmic systems and the related issues of transparency and accountability since the fall of 2016 as part of my work at the Centre for Informatics and Society¹ (C!S) at the Vienna University of Technology (TU Wien). Prior to my engagement with these topics as a scientist, I had been (and continue to be) working as a senior back-end developer at WIENFLUSS², a software development firm and web design agency located in Vienna, Austria. As such, I was part of the initial team of software engineers at WIENFLUSS tasked with implementing the EnerCoach energy accounting system in the spring of 2014, and have been the lead developer for that team since early 2015. *EnerCoach Online*, as it was initially called, was a collaborative, online energy accounting system for communities in Switzerland, and the next iteration of the original and widely used EnerCoach Tool, previously implemented as non-collaborative Microsoft Excel document. Initially tasked with only implementing the system according to (at that point yet-to-be developed) specifications, the focus of my work quickly transitioned to supporting the EnerCoach Working Group (as a client of WIENFLUSS) to produce these viable specifications, and subsequently implement them. After the initial release of the system in mid 2016, I continued to work as the lead developer for EnerCoach, closely working with the project management and software development teams within WIENFLUSS on the one hand, and the EnerCoach Working Group on the other. This project included extending, troubleshooting and maintaining the system—a role I am fulfilling to this day.

As a researcher at the C!S, I initiated the research project culminating in this case study in 2018, whereby I formalized a research cooperation between WIENFLUSS, the EnerCoach Working Group and the C!S shortly thereafter. As part of this agreement, I gained unfettered access to the codebase and all related documents for the purpose of this study, as well as a commitment for cooperation with the various stakeholders of the system to partake in interviews and allow in-situ observations of training workshops for new users. Figure [3.1](#) illustrates the timeline of this research project in the larger context of my work as an employee at WIENFLUSS and lead developer of the EnerCoach system there; having worked on the EnerCoach project as a developer for roughly 8 years by now, my engagement with the project in a research capacity started in early 2018 and lasted approximately 3 years until the first full publication of results [\[1\]](#). Further engagement with the system as part of the comparative case study (see Section [3.4](#)) is not shown in the timeline, as it was performed as a separate study (albeit relying on the results of the case study, but commenced only after its conclusion).

Given this arrangement, the approach to the EnerCoach case study was intrinsically *auto-ethnographic* in nature, extending and appropriating the ethnographic methodologies suggested by Seaver that were outlined in the previous section. Auto-ethnography is characterized as “[a] form of self-narrative that places the self within a social context.” [\[25, p.15\]](#). Reed-Danahey [\[25, 26\]](#) points to the roots of auto-ethnography in qualitative

¹<https://cisvienna.com>

²<https://www.wienfluss.net>

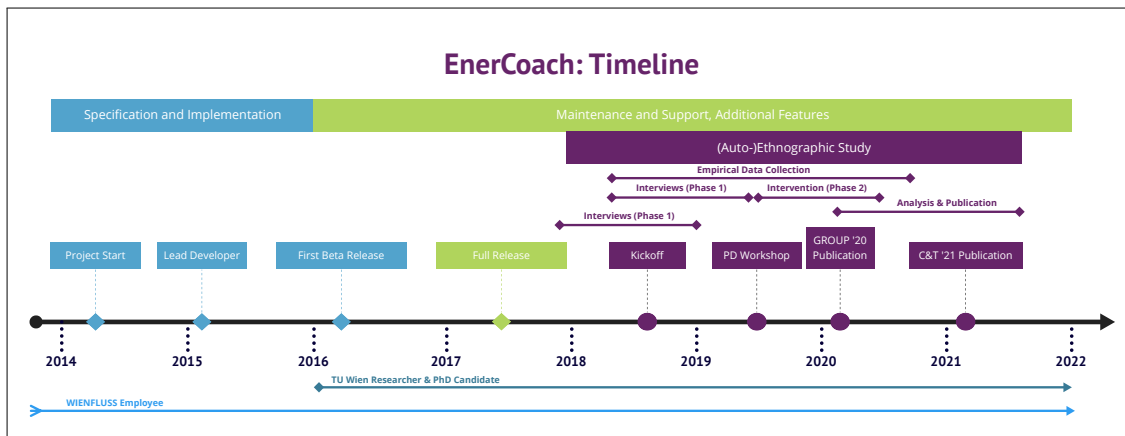


Figure 3.1: EnerCoach research project timeline showing the relative length of engagement in a research capacity compared to my overall employment at WIENFLUSS.

research in the social sciences on the one hand, and literary critics “[...] *mainly concerned with the voices of ethnic autobiographers*” [25, p.15] on the other. Between those two readings of the term, this case study adheres to the former, rather than the latter, tradition. Specifically, I employ *algorithmic auto-ethnography* as the reflexive approach to studying an algorithmic system by synthesizing insights from two points of view: the perspective of a researcher, and the perspective of a software engineer that is part of the socio-technical assemblage himself. Outgrowing its social science roots, auto-ethnographic methods have been applied within computer science in general and the **HCI** community in particular, e.g., as an educational tool [243], to investigate discrimination of black people in IT through their own experiences [244], or as part of design studies [245, 246, 247].

Responding to critique from proponents of the more positivist paradigms of science, I side with advocates for (auto-)ethnography by arguing that, rather than limiting the validity of results of such a mode of inquiry, this reflexivity and acknowledgment of one’s position can be enriching and add texture to the resulting insights. Likening the work of a critical qualitative researcher to that of a *bricoleur*, a quilt-maker, Denzin and Lincoln make much of the same points in their metaphorical summary:

“The interpretive bricoleur understands that research is an interactive process shaped by one’s personal history, biography, gender, social class, race, and ethnicity and those of the people in the setting. [...] The product of the interpretive bricoleur’s labor is a complex, quilt-like bricolage, a reflexive collage or montage; a set of fluid, interconnected images and representations. This interpretive structure is like a quilt, a performance text, or a sequence of representations connecting the parts to the whole.”

[248, p.5-6]

Criticism of auto-ethnographic methods and approaches far exceed these claims of limited validity of results, as Ploder and Stadlbauer [249] explicate, although *published* criticism is surprisingly scarce. The reason for this gap may be a type of ostracization in and of itself, as they postulate:

“Critique implies a certain level of attention to an approach, and one of the most powerful weapons in academic discourse is to ignore it.”

[249, p.8]

In their reflection on various arguments levied against auto-ethnographic approaches, they identify a number of critical points auto-ethnographic researchers in the cultural and social sciences³ find themselves confronted with, sometimes openly, but more commonly implicitly and behind closed doors. Among them, they describe how researchers find themselves accused of solipsistic or even narcissistic tendencies, of threatening disciplinary identities, and of making a “*strategic mistake*” [249, p.7] in the larger struggle for acceptance and recognition of qualitative research. As far as it pertains to this case study, the auto-ethnographic elements are to be seen as the starting point, rather than the end; claims of solipsistic epistemology are thus rather easy to dismiss in light of the number of sources extending beyond my own experience. As for the second criticism, I wholeheartedly embrace the notion of contributing to the weakening of disciplinary identities: As I have argued in previous sections, the fundamentally inter-disciplinary nature of this case study and, indeed, this dissertation as a whole is founded on the idea that studying algorithmic systems from a critical perspective requires methodological and disciplinary transgression. Lastly, even if qualitative research were truly threatened by the inclusion of auto-ethnography as an acceptable methodology (which I doubt), I would argue that my experience as a software engineer and the insights I can contribute hardly fit the bill of introducing an “*affective immediacy*” that “*creates an emotional closeness to the writer that precludes criticism.*” [249, p.6], as Ploder and Stadlbauer formulate on of the core criticism they encountered.

In light of these considerations regarding potential points of criticism, and including the arguments made in the previous Chapter 2 on the complexity of algorithmic systems as socio-technical assemblages and the stated goal of unpacking and disentangling said assemblages, the auto-ethnographic approach is indeed a viable, and in some respects even uniquely suited methodological approach to achieving the goal of a rich, critical, reflective and meaningful analysis of the EnerCoach system.

For this case study, this means that—as a researcher—I am acknowledging and explicating the role I have had in the cultural context and socio-technical assemblage that is the

³Ploder and Stadlbauer initiate their analysis based on their own auto-biographical experiences in the German-speaking cultural and social science communities; as much of their observations and the critical arguments they list are founded in and referencing international sources, their list of critiques clearly applies beyond just the D-A-CH regions of academia.

EnerCoach accounting system, and work to critically reflect on both my subjective impact on the system's socio-technical figuration as well as the impact my role may have had on the analysis of the system as part the research study. Furthermore, this approach demands a disciplined reflection on the way other members of that socio-cultural context (e.g., stakeholders, users, and interviewees) perceived me as I interacted with them in my role as a researcher. Interrogating to what extent issues of power and knowledge have played a role in the material gathered through such fieldwork is, thus, an integral part of the analysis as well. To do so, great care was taken during fieldwork to ensure that the study participants could freely express their points of view or share their knowledge, without taking a stance of authority as a developer on the technical aspects of the system, unless specifically prompted by direct questions the participants asked.

The auto-ethnographic method requires a critical, reflexive and sensitive approach in order to be considered as a valid scientific form of inquiry. A carefully balanced synthesis between 'objective' and 'subjective', or as Reed-Danahey [25] puts it, 'outsider' and 'insider' points of view, can help avoid giving the impression of positivist 'objectivity' unfitting for an analysis of the EnerCoach system that puts an equally strong emphasis on the social and socio-technical aspects of the system as it does on the purely technical ones. To this end, I explicate clearly whenever insights stem from my own experience as a developer, as opposed to insights gained through empiric data collected from external sources. In addition, these insights are rarely seen as the endpoint of inquiry, and rather as the starting point for further validation and verification. To clarify: where I, as a developer, may have had knowledge into the reasoning behind specific design decisions, this insight was cross-verified through interviews, observation or document analysis where possible. In the few cases where this was not possible, but where such insights were still relevant to the analysis, these observations will be clearly expressed as speculative or subjective views.

As a consequence of being both an employee at WIENFLUSS and a researcher studying the system, particular care needed to be taken to avoid conflicts of interest, including those where economic interests of my employer may have biased the results in some way. In a bid to avoid such pitfalls, the research agreement included the common understanding that any and all implementation efforts would only be performed as part of the research project only and could not be considered billable hours under any circumstances. Given the very limited budgetary resources available to the EnerCoach Working Group for the implementation of additional features during the time of the study, this did not present any significant economic disadvantage for WIENFLUSS, since whatever concrete measures or other prototypical implementations would result from the study and workshop would otherwise not have been possible to implement as paid service by WIENFLUSS due to these budgetary constraints. Likewise, the EnerCoach Working Group members as primary agreement partners on the other hand entered into the research project with the understanding that any and all implementation measures would be limited by what was reasonably possible to do as part of my research project and during my working hours as a researcher. These agreements meant that I also needed to maintain a detailed record

of working hours to be able to, upon request, document my efforts. Finally, in the first clarifying negotiation with the owners of WIENFLUSS, they made it absolutely clear that my employment status would not be influenced by the study itself or its results in any way, and that a disclosure of results would only be necessary to the extent required to discuss whether sensitive information could be published; such disclosure was to happen by my own initiative should I deem certain sensitive materials worthy of consideration for publication. At the end of the study, I did not feel the need to discuss any of the things I published or otherwise disclosed as part of this dissertation.

Besides the caveats of auto-ethnographic inquiry and the strategies to address them outlined above, the dual role as researcher and developer of the EnerCoach system provides significant benefits that made this case study into a unique opportunity to study an algorithmic system ‘in the wild’. First and foremost, as widely recognized by scholars utilizing ethnographic methods in general and in the literature on algorithmic transparency in particular, gaining access to the subject of study can be a challenge; a lack of such access may even prohibit gaining viable insights into certain crucial aspects of an algorithmic system (see Section 2.3.2 for a more detailed discussion of these issues in relation to algorithmic systems). As such, the EnerCoach case study represents one of the rare cases of a complex algorithmic system with a rich community of users and stakeholders that I had the privilege to study in detail and with virtually no restrictions in terms of access. Secondly, my personal involvement in the development process from the start of the project provided invaluable insights into both the overall strategic goals of the system and the minute details of value-laden choices made as part of the conceptualization and implementation process. The willingness of both the EnerCoach Working Group and WIENFLUSS to cooperate with the research project made an in-depth, holistic analysis of all aspects of the system possible.

Finally, working closely together with the EnerCoach Working Group during the implementation phase until the first release of the system afforded me the chance to gain the necessary domain knowledge that would, later on, allow me to effectively communicate with the study participants on the highly specialized subjects of energy accounting and sustainable energy technologies. This prior knowledge also reduced much of the terminological friction, translation work and misunderstandings that normally occur when ethnographic researchers first enter a domain they know little about.

3.2.2 Phase 1: Data Sources and Analytic Methodology

Seaver [21] suggests some ethnographic tactics for the study of algorithms that influenced the choice of methods used to analyse the EnerCoach energy accounting system. Amongst others, he emphasizes the value of heterogeneous sources of information and an, at times, “*apparently undisciplined approach to ethnographic data collection*” [21, p.6] by encouraging scholars to ‘scavenge’ for ethnographic data. To arrive at what Clifford Gertz [27] called a *thick description* of the EnerCoach system—including its technical figurations, its stakeholders and the practices they are engaged in, as well as the larger world of energy accounting and sustainability initiatives it is embedded in—a variety of sources were

utilized. The following list outlines these sources as well as the methodology of their analysis.

Coded interview transcripts *A total of 8 semi-structured qualitative interviews ranging from approximately 60 to 120 minutes each were conducted including the members of the EnerCoach Working Group, EnerCoach Hotline staffers, energy consultants / auditors and other end-users of the system.*

The interviews were conducted between September 2018 and October 2019 and based on predefined interview guidelines⁴, which roughly focused on three core topics: *Stakeholders, Project Structure and Attributes and Transparency, Accountability Literacy*. In the first section, (1) the interviewee’s history with the tool, (2) their knowledge about other stakeholders of the system, (3) its users and their common tasks, as well as (4) the larger organizational context the tool is embedded in, were at the focus of the questions. In the second section, (1) the interviewee’s knowledge and assessment of the overall goals of the system, (2) the different functionalities and their utility for various stakeholders, as well as (3) their knowledge of the history of the system (from conception to implementation and current use), was collected. For the last section, the interviewee’s gave their impressions of the current state of the EnerCoach tool in relation to the issues of transparency, accountability, explainability, as well as their assessment the corresponding levels of literacy of the various stakeholder groups, including a self-assessment of their own levels of understanding and expertise.

To protect the study participants’ anonymity as per the general research agreement and in accordance with the requisite consent forms they signed, any references to their person (including for direct quotes) throughout this dissertation are blinded through the letters “A” through “H”. Of the eight interviewees, three were active members of the EC Working Group, two were active hotline staffers, and the remaining two were energy consultants/auditors as well as end-users for their own communities. While these were the primary roles of these interviewees, significant overlaps in expertise and roles either existed before the interviews or manifested as the interviewees transitioned to new roles throughout the research project duration. To explicate: All EC Working Group members (interviewees “A”, “B”, “C”) and both hotline staffers (“E”, “F”) also had previous working experience as energy consultants; one energy consultant (“D”) was also frequently working as an energy auditor; another energy consultant (“G”) would later informally join the EnerCoach Working Group; one hotline staffer (“H”) had previously been purely a community/end-user, and one EC Working Group member (“C”) was also the lead developer responsible for the original Excel version of the EnerCoach tool.

The resulting interviews were transcribed in part by the interviewer (5 instances) and in part by a professional transcription service (3 instances). In the case of the latter, an additional verification step of the quality of transcription was performed

⁴See Appendix [A.1](#) for the original German version of those guidelines

after a cursory examination showed some minor discrepancies between the original audio recordings and the transcription, which were subsequently corrected.

After the completion of each interview transcript, the resulting text was analysed following Mayring's [250] methodology for qualitative content analysis. Mayring's approach considers three strategies for the analysis of large corpora of qualitative data: *Summarization*, *Explication* and *Structuring*. Of these three, the analysis mostly relied on *Summarization* as an *inductive* form of interpretation. *Structuring* or deductive application of categories was not applied given the exploratory character of the interviews and their content, as well as to avoid introducing preconceived notions of expected categories into the material. *Explication* was employed insofar as the resulting annotated material was situated within the larger context of the EnerCoach system; this was done by supplementing codes with references to other data sources (including references to code sections, emails or field notes) to clarify meaning or validate claims made by the interviewees.

Considering the sceptical stance of some proponents of (auto-)ethnographic methodologies towards structured coding schemes (e.g., Denzin and Lincoln [248]), the choice to follow Mayring's approach deserves some closer attention. Structured coding schemes are sometimes seen as a (misguided) attempt at creating a more *objective* reading and analysis of the source material to lend its results more credibility and validity [251]. Such a stance, naturally, would stand in stark contrast to the assertion of auto-ethnographic approaches that value the interpretative sovereignty of the researcher. Thus, the choice for Mayring's rather structured approach but with a specific focus on *inductive* reasoning follows much more pragmatic reasons. First, following a structured process helped enforce a critical interrogation of interpretative assumptions as part of the coding process, which helped put a stronger focus on where the auto-ethnographic experiences informed specific choices of codes, themes or focus. Second, the structured nature of the process helped embed the results within the larger corpus of ethnographic data through cross-referencing terms between interviews and other materials, such as the code samples, emails or field notes about the training session. Finally, in the most practical sense, the structured nature of Mayring's approach helped tackling the significant amount of data to be analysed by dividing the process into more manageable steps. In particular, the definition of analytic units was helpful to accommodate for the rather stark differences in dialect and speech resulting from Switzerland's diverse demographics and the various levels of Swiss German dialect used by the participants. In summary, the adherence to this particular type of qualitative analysis should not be seen as a repudiation of the underlying concepts of situated (auto-)ethnography, but rather as a pragmatic and practical approach to better integrate the various kinds of empiric data and make their analysis more manageable. As such, the structure and rules of analysis were not particularly restrictive or prescriptive in terms of the interpretative process: besides the definition of analytic units, the iterative process (including, e.g., the number of repetitions) itself was not strictly defined before the start, but adapted as new requirements and data emerged from

both the following interviews and other empiric sources.

The *summarization* technique itself was an iterative process roughly organized around the following steps as outlined by Mayring [250, p.70]:

1. Definition of analytic units: Unit of coding, context and evaluation
2. Paraphrasing, generalizing and reduction of text passages by either omission or selection
3. Collection of results as codes
4. Re-evaluation of attributed codes with the source material

For step one, the following analytic units [250, p.61] were defined a priori:

- **Unit of Coding:** The smallest unit of text a code would be applied to were full sentences or sentence fragments where the speaker did not formulate a (grammatically) complete sentence
- **Unit of Context:** The largest unit of text a code would be applied to were series of consecutive sentences responding to a specific question by the interviewer. Interjections splitting up such consecutive answers would also require new codes to be applied to the following sections as a general rule; exceptions were made only for non-lexical conversation sounds by the interviewer (utterances of understanding or encouragement otherwise devoid of topical meaning).
- **Unit of Evaluation:** Transcripts were evaluated in *chronological* order, both within a single (group) interview and between different interviews.

Steps 2 through 4 were performed iteratively within every interview transcript at least twice to (re-)evaluate and generalize existing codes. Additionally, a third iteration of generalization and reduction of the codes assigned after all interviews had been analysed was performed across all interview transcripts to align the results. All of these steps were performed manually via the ATLAS.ti⁵ software suite for qualitative content analysis.

Resulting from this process was a set of 20 distinct codes with a total number of 303 code occurrences in the texts. Table 3.1 gives a quantitative overview of codes and occurrences.

Email communication *To explicate and validate insights gained through the qualitative content analysis outlined above, selected email communication was included in the analysis.*

This included communication between users, members of the EnerCoach Working Group, members of the EnerCoach hotline, and members of the project management

⁵<https://atlasti.com/>

| Code | Number of occurrences |
|--------------------------------|-----------------------|
| Accountability | 21 |
| Alternatives | 4 |
| Autoethnography | 1 |
| Case Study | 3 |
| Goals | 9 |
| History | 20 |
| Influence of Stakeholders | 10 |
| Known Issues | 39 |
| Literacy | 27 |
| Measures | 10 |
| Real-World Impacts | 12 |
| Specifications | 14 |
| Stakeholder | 41 |
| Tensions | 4 |
| Time Investment | 3 |
| Tracing Input <> Output | 5 |
| Training/Information Resources | 15 |
| Transparency | 41 |
| Trust | 9 |
| Usage | 15 |
| Total | 303 |

Table 3.1: Overview of codes and their number of occurrences after the qualitative content analysis following Mayring [250]

and development team at WIENFLUSS, as well as automated emails generated by the EnerCoach logging system and sent to the developers of the system. Emails were only considered as part of the analysis if all authors and recipients had previously agreed to contribute to the study and signed a disclosure agreement. Given the fact that communication with these stakeholders almost always happened via email in addition to sporadic preceding phone calls, these emails were preferable over recording and transcribing verbal or phone conversations.

EnerCoach Training Session Observation *Two training sessions offered by the EnerCoach Working Group to current and future users of the system were observed.*

Each of these two consecutive training sessions was attended by 15 participants and recorded through field notes. The resulting data was augmented with the presentation slides used by the lecturer. A summary interview focused on questions arising from the observation was conducted with the lecturer immediately after the second training session in late October 2019; due to the short and focused nature of the interview, and in line with Seaver’s recommendation to “[t]reat interviews

| Language | Files | Comment Lines | Code Lines |
|------------|-------|---------------|------------|
| CSV | 581 | 0 | 184123 |
| SQL | 4 | 415 | 62250 |
| Python | 157 | 7525 | 34102 |
| JavaScript | 71 | 9976 | 24994 |
| PO File | 9 | 8081 | 10195 |
| HTML/XML | 83 | 328 | 8491 |
| CSS | 8 | 231 | 766 |
| SVG | 21 | 7 | 660 |
| Total | 934 | 26563 | 325581 |

Table 3.2: Overview of number of files and lines of code in the EnerCoach code base as provided by *cloc* [252].

as fieldwork” [21, p.7-8], this interview was transcribed, but not included in the coding process outlined above, but rather considered as an addendum to the field notes gathered during the observation.

Complete code review *A complete code review of the system as it was deployed in the spring of 2020 was performed.*

Starting with the functional code of the application itself, but extending into inline comments, commit messages into the version-controlled codebase, as well as the ticketing systems used to track open issues and the development of new features. Table 3.2 shows a simple tally of files and lines of code⁶ in the EnerCoach system as a broad indication of the scope of the project. Finally, monitoring tools used as part of the regular maintenance of the tool (i.e., an Elastic⁷ stack including Logstash⁸ and APM⁹ tracing) allowed quantitative insights into usage patterns and the utilization of certain features, and direct production database access was used for certain specialized queries and the collection of usage statistics both before and after the implementation of transparency-improving measures.

The code review outlined above should not be confused with the long-standing practice of code reviews in the software development industry (c.f. [183, 184, 185]). As discussed in Section 2.4.2.2, code reviews in this sense can be considered a type of *forum* for *professional accountability*, and find wide spread use as a measure to improve code quality and reduce errors and bugs (albeit with certain limitations, as the critique by Doğan et al. [186] shows). By contrast, the code review performed as part of the EnerCoach case study was less concerned with a normative evaluation of code quality than with a critical analysis of *functionality* and *embedded values*. In

⁶For the code review, only Python, SQL, JavaScript and XML files were analysed.

⁷<https://www.elastic.co/elastic-stack/>

⁸<https://www.elastic.co/logstash/>

⁹<https://www.elastic.co/observability/application-performance-monitoring>

line with this, a critical interrogation of the impact that design and implementation decisions had on the transparency and accountability of the system—as captured in the stakeholder interviews—was the central approach. To that end, I analysed the EnerCoach codebase as a specialized form of *document analysis* as outlined by Bowen [253]. The five specific uses of documents Bowen identifies are (1) *providing context*, (2) *identifying questions to be asked*, (3) *providing supplementary research data*, (4) *tracking change and development* and (5) *verifying findings and corroborating evidence* [253, pp.29-30]. The EnerCoach code review fulfilled most of these functions, but was particularly helpful in providing insight into the progression of the tool over time, yielding questions to be asked as part of the interviews and verifying the findings produced by the analysis of technical and specification documents.

In practice, the code review adhered to the following procedure:

1. Identification and categorization of relevant files
2. Cross-referencing of files and code excerpts with specifications
3. Identifying relevant code samples referenced as problematic by the stakeholder interviews in terms of transparency and accountability
4. Formulating supplementary questions for further interviews

The steps outlined above were iterated multiple times before, during and after the qualitative interviews.

Technical & specification documents *Certain internal technical documents (including specification documents, excel tables outlining calculation procedures) were included in the analysis.*

Conceptually, these specification documents allowed tracing the development and decision-making processes, juxtaposing the functionalities between the predecessor of the EnerCoach tool (a non-collaborative implementation based on Microsoft Excel) and the current, web-based implementation. Where possible, different iterations or subsequent versions of these documents were compared to spot progression and differences across time as the specifications were refined. Similar to the code review outlined above, the document analysis followed in the tradition of constructionist research [253, 254] and was performed in parallel to both the code review and the interview process.

While the list of data sources and mode of analysis listed above represents the core set of sources included in the analytic phase of the case study, *field notes* (taken as part of the observation of the training session, as well as during the interactions with various stakeholders preceding or following the interviews and workshops) represent important supplementary qualitative data. Utilized as annotations, they documented my insights and observations during the fieldwork and connected the various other data sources to

each other. Favours timeliness over structure, I did not follow specific formats when jotting them down (with the exception of timestamping). Consequently, I did not consider them part of the core data sources as part of the analysis, and rather used them as a starting point for follow-up questions and inquiry; any insights documented through field notes were not included in the case study itself unless they were cross-validated through other sources, such as interview transcripts, specification documents or the code review.

3.2.3 Phase 2: Intervention through Participatory Design Workshops

The second phase of the case study following the analysis outlined above was conducted in the tradition of *participatory design* [255, 256, 257]. Having identified the specific deficiencies of the EnerCoach system in relation to transparency and accountability, the insights gathered suggested that human-centric approaches that did not involve relevant stakeholders in the design processes themselves would not suffice to tackle the wicked nature of the problem as outlined in Section 2.5. Instead, a *participatory* approach offered the chance to provide the people affected by this system the agency to “[...] *make decisions about how they do their work and, indeed, how they perform any other activities where they might be supported by an IT artifact.*” [255, p.251].

Participatory design can be traced back to its historic roots as interwoven strands of inquiry and action-based research in the early 1970s and 1980s. Kensing and Greenbaum [258] provide an excellent historic overview of these roots, reflecting on both the political context in which the methodology first emerged and the theoretical foundations this methodology was built upon. Most relevant to this case study, they list a number of guiding principles of participation, some of which provided the arguments for choosing this approach. First and foremost, a core tenet of participatory design is the *equalisation of power relations* in the work context [258, p.33]. Organisational and hierarchical power structures can systematically disenfranchise certain groups of workers, who often remain invisible to those designing and implementing the very tools they must subsequently use in their work. In the context of the EnerCoach system, the various groups of users and stakeholders showed similar characteristics in agency to influence the system’s socio-technical figuration. Including them in the design process for improvements to the system offered a means to increase their agency in that regard. Considering Bandura’s model of emergent human agency introduced in Section 2.4.4, participatory design also offers the benefit of supporting the formation of a greater *belief in self-efficacy* for the participants, as the inclusion in these design processes also demonstrates to participants how valuable the contribution of their expertise can be, and how their participation directly affects the outcome of the design process. Secondly, participatory design approaches recognize the value of “*situation-based actions*” [258, ibd.] as a means to move away from abstract specifications and towards designing with and for the people in the working environment they are used to, fostering a process of “*mutual learning*” [258, ibd.] and leading to the discovery of different, more applicable “*tools and techniques*” [258, ibd.]. Finally, the shift in perspective from technical experts designing technology *for* users towards a process of designing *with* users encourages the emergence of “[a]lternative visions about technology”

[258, p.34] that may be difficult to imagine for technical experts working in a different context from the one they design for. All of these principles apply to the case study at hand, making the choice for participatory design as the principle mode and methodology of intervention evidently obvious.

In practical terms, the second phase case study consisted of a one-day workshop held on site at the office of one of the companies¹⁰ maintaining and providing support to the users of the EnerCoach system for the German-speaking part of Switzerland. A total of five participants, including three members of the EnerCoach Working Group as well as a staffer of the EnerCoach hotline and an expert user, were included. The workshop concept consisted of two parts: first, the participants completed a collaborative exercise to explain the inner workings of the algorithms calculating one of the reports provided by the system, to their best knowledge and understanding. For this exercise, I only took the role of an observer, neither correcting nor contributing in other ways to the result. This exercise led to a proposal for a visualization for said algorithmic process, which was then contrasted with the actual implementation in terms of its accuracy and expressiveness. For this second part, I assumed the role of a technical expert, contributing my expertise as a developer of the system, and clarifying questions and misunderstandings that emerged during the prior exercise.

For the second part of the workshop, the participants collected a number of concrete issues relating to transparency and their proposed solutions, by creating mockups [259, 260] and non-functional, collaborative paper prototypes [261, 262] of potential technical measures. These designs were then discussed in terms of their feasibility for implementation and potential for resolving the issues brought up by the participants. During this part, I answered the participants' technical questions relating to feasibility of implementation when prompted, but otherwise did not proactively intervene in the process. As the final result of the workshop, the participants ranked the mockups of technical measures they designed by importance, which I used as an input to implement the measures as specified.

After implementing the proposed measures, they were rolled out to the live production instance of the system and evaluated both qualitatively through gathering feedback from the staffers of the EnerCoach hotline (via phone calls and subsequent emails), and quantitatively through database queries and server logs detailing the frequency of use. These two approaches were designed to complement each other; while the hotline staffers provided their subjective impression of how well the measures were received by the users and how useful they were for their own work, the quantitative analysis of actual usage patterns for these new features was needed to validate the subjective claims made by the staffers.

The workshops themselves were recorded both as video and audio recordings, which served as an additional data source to evaluate the feasibility and potential of the participatory design methodology itself in the context of algorithmic systems and their related issues of transparency and accountability. A written transcript of the workshop was analysed

¹⁰Nova Energie GmbH, <https://novaenergie.ch/>

and coded following the same procedure as outlined above in Section 3.2.2, and the video recordings were used to extract still images documenting the process.

3.3 Case Study Methodologies: AMAS / AMS Algorithm

The second case study focused on the **Arbeitsmarkt-Assistenz-System (AMAS)** system [3, 4], developed by the **Public Employment Service Austria (AMS)** in cooperation with the private research firm **Synthesis Research GesmbH**. The research project for this study was conducted as a cooperation between the **Centre for Informatics and Society (CIS)** and the **Institute of Technology Assessment (ITA)** at the **Austrian Academy of Sciences (ÖAW)**, co-financed by the **Austrian Chamber of Labour for Upper Austria (AKOÖ)**.

Contrary to the EnerCoach case study methodology documented above, access to the system was quite limited before and throughout the project. In the fall of 2018, the **AMS** announced their plans to test and deploy the **AMAS** system, followed by the publication of the first document “**The AMS Labor Market Chances Model: Documentation of Methods**”¹¹ [DOK_1] purporting to describe the system’s inner workings. A strong public interest and numerous critical voices from a variety of academic and non-academic sources arose in response. Amongst them, a research group interested in an in-depth investigation of the system formed including Doris Allhutter (**ITA**), Astrid Mager (**ITA**), Fabian Fischer (**CIS**), Gabriel Grill (University of Michigan at Ann Arbor) and myself.

Our subsequent requests for information from the **AMS** only yielded very limited results, until the **AKOÖ** voiced interest in financing a joint research project and help facilitate access to allow a more in-depth, scientific analysis and evaluation of the system. After an agreement was reached, the **AKOÖ**—itself a member of the **AMSs** Administrative Board—negotiated access to further documents, which should detail the technical functionalities as well as shed light on the process of how the **AMAS** system was conceived and developed. This agreement to disclose non-public information for the purpose of this study also compelled the **Synthesis Research GesmbH** to participate in the project by answering additional questions and providing further specification documents. No direct access to the system itself, the underlying data sources, or users (e.g., **AMS** caseworkers or affected jobseekers) was granted; consequently, the case study would rely first and foremost on the detailed analysis of the documents provided, as well as on the transcripts of meetings with representatives of the **AMS** and **Synthesis Research GesmbH**, and other publicly available documents (e.g., parliamentary inquiries, audit reports by the **Austrian Court of Audit (ACA)**, open letters and responses by interest groups such as the **Austrian Ombud for Equal Treatment (GBA)**).

Document analysis as a qualitative research methodology follows the constructionist tradition of treating textual sources as a manifestation of social reality, or at least a depiction thereof. Atkinson and Coffey summarize this stance succinctly:

¹¹Orig. “*Das AMS-Arbeitsmarktchancen-Modell: Dokumentation zur Methode*”

“Documents are ‘social facts’, in that they are produced, shared and used in socially organized ways. They are not, however, transparent representations of organizational routines, decision-making processes or professional diagnoses. They construct particular kinds of representations using their own conventions. Documentary sources are not surrogates for other kinds of data. [...] This recognition or reservation does not mean that we should ignore or downgrade documentary data. On the contrary, our recognition of their existence as social facts (or constructions) alerts us to the necessity to treat them very seriously indeed. We have to approach documents for what they are and what they are used to accomplish. We should examine their place in organizational settings, the cultural values attached to them, their distinctive types and forms.”

[263, p.58]

While organizational documents should not be seen as a “surrogate” for other data, as Atkinson and Coffey put it, they nevertheless offer a unique perspective into the organizational processes and structure of the actors involved in the development of the **AMAS** system. Given that the documents we received were a mix between publicly available and internal documents, they are a prime example for how the **AMS** chooses to “*represen[t] themselves collectively both to themselves and to others.*” [263, p.56] in regards to the **AMAS** system. Heeding Atkinson and Coffey’s warning as quoted above, the information gleaned from them was considered from a critical perspective and not assumed to be either factually truthful or factually false in a positivist sense. Even within the corpus of documents, contradictions and diverging descriptions occur, alerting us to the fact that the sum of all documents is neither a coherent nor a polished representation, but rather a collection of single points of view, diverse in structure, purpose, intended audience and contextually embedded within a time and place of origin within the organisations. This point notwithstanding, taken collectively they do create a social reality, which can be shown to conflict with other social realities constructed through oral interviews with the stakeholders of the system (in this case, the **AMS** and Synthesis Research GesmbH). The juxtaposition of these realities helps shed light on the process of co-production that resulted in the design, development and implementation of the **AMAS** system, and contributes to unpacking its socio-technical assemblage.

More specifically, Bowen’s [253, pp.29-30] list of uses for document analysis also apply to varying degrees:

- Providing context
- Identifying questions to be asked
- Tracking change and development
- Providing supplementary research data
- Verifying findings and corroborating evidence

First and foremost, the documents provide *contextual information* about the system and its genesis, e.g., through outlining discussions in preliminary meetings or protocols of formal agreements of the [AMSs](#) board. They help distil intention and meaning, the motivations behind certain decisions, and situate them temporally in the development process. Even the absence of decisions or attention given to certain aspects of the development of the [AMAS](#) system may indicate intentional obfuscation or lack of prioritization of these topics (e.g., the fact that independent scientific evaluation of the system was never mentioned in the documents). These observations provided the starting point for *further questions* and lines of inquiry posed to the stakeholders we would meet with.

Some types of documents, particularly recurring meeting transcripts between the [AMS](#) and Synthesis Research GesmbH, make a progressive *tracking of changes* possible. Understanding how the decision processes played out over the course of development provides invaluable insights into the day-to-day processes of co-production that influenced the final outcome. While not all documents are primarily relevant to the core topics of the case study, they nevertheless *supplemented our insight* on precursors for the system and the overall context.

The final use of documents as suggested by Bowen [\[253\]](#), *verification* and *corroboration*, contradicts Atkinson and Coffey’s [\[263\]](#) stance that documentary data should be regarded as primary qualitative data “*in their own right*” [\[253\]](#), p.59], and not simply as validation for other data. As the wealth of documents provided to us represented the primary focus and core material we worked on, we elected to take the stance of Atkinson and Coffey over the approach proposed Bowen, and regarded both oral qualitative data and document analysis as different social realities—with overlaps and contradictions towards each other—that nonetheless represent different aspects of the larger socio-technical assemblage of the [AMAS](#) system.

3.3.1 Practical Implementation

The case study of the [AMAS](#) system took place between late 2018 and late 2020. After the initial announcement of the system by AMS CEO Johannes Kopf through the parallel publication of an interview and a more detailed news article at the Austrian newspaper *derStandard* [\[264\]](#), [\[265\]](#), the Synthesis Research GmbH published the first document [\[DOK_1\]](#) in an effort to show the [AMS’s](#) commitment to transparency about the system; the document was widely circulated and used as a primary resource by both academic and non-academic commentators and critics of the proposed system. Following this period of focused attention on the system, our research group published the first in-depth critical analysis [\[3\]](#) of the system in early 2019 based on the documents that were publicly available at the time [\[DOK_1\]](#), [\[BER_15\]](#), [\[PARL_1\]](#), [\[PARL_2\]](#), [\[BER_13\]](#), [\[BER_14\]](#), [\[NOTES_7\]](#), [\[NOTES_6\]](#), [\[NOTES_10\]](#). Given the limited amount and vagueness of information available in the documents published by Synthesis Research GmbH, many aspects of the [AMAS](#) system were still shrouded in secrecy, making the need for further study evident. During the summer months of 2019, the formal research agreement between the [ITA](#) / [C!S](#) and the [AKOÖ](#) was formed, leading to the first meeting and group interview with

representatives of the [AMS](#) and Synthesis Research GmbH in October 2019. By January 2020, the complete set of documents used in this case study and subsequent publications was delivered to us, and a second meeting to discuss further questions resulting from the analysis of the material was scheduled for March 2020. As a consequence of the COVID-19 pandemic, this meeting never occurred in person, but was substituted with an exchange of questions and answers via email between our research team and members of the Synthesis Research GmbH. The final and most detailed report on the system titled “The AMS Algorithm: A socio-technical analysis of the Labour-Market-Assistance-System (AMAS)”¹² [\[4\]](#) was published in November 2020. Figure [3.2](#) provides an overview of the timeline for this case study and the AMAS system as a whole.

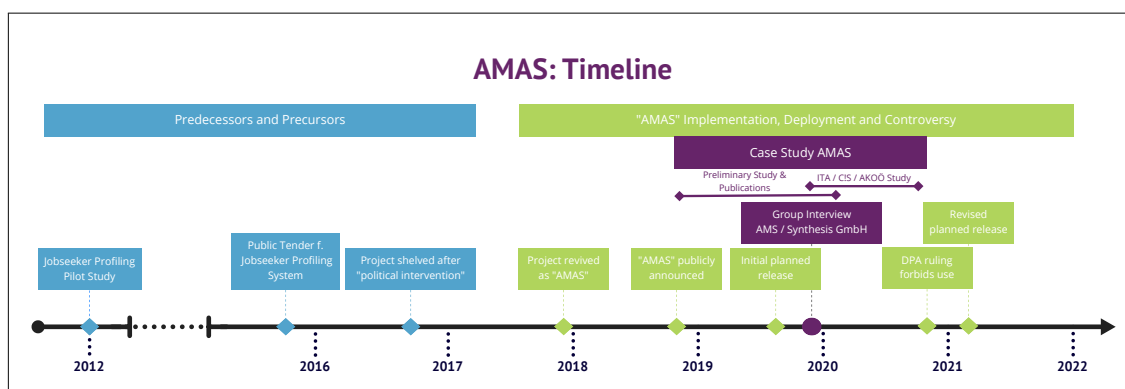


Figure 3.2: AMAS research project timeline outlining precursors to the system, AMAS implementation and planned release dates, as well as case study research project duration and milestones.

After receipt of the trove of documents, the following steps were taken to analyse the corpus:

Creation of document index *A document index was created and shared between the research group for collaboration.*

Within the index, each document was initially assigned a unique ID and referenced through its filename and path on a shared document folder.

Extraction of bibliographic information *After the initial index was created and all documents were entered into it, a set of bibliographic data was extracted from each document.*

Bibliographic data extracted included:

- Author(s) / Issuing organisation
- Date of receipt

¹²Translated by the author, original title in German: “Der AMS Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)”

- Date of publication
- Number of pages (where applicable)

Authorship via personal reference was treated as preferable over listing the issuing organisation. Furthermore, the date of publication was determined primarily through dates listed in the document content; where this information was absent within the document, file metadata was consulted as a contingency.

Prioritisation, assessment of relevance and short description *As the third step, a cursory reading of the documents was performed.*

The goal of this preliminary reading was to determine a rough estimate of relevance (as discrete values of ‘low’, ‘medium’ and ‘high’ relevance) and subsequent prioritisation of the documents for detailed reading. Furthermore, this reading included entering a short one-line description of the document content into the index for ease of navigating the trove of documents.

Document classification *As a fourth step, the documents were iteratively and inductively classified into one or more of 33 document types.*

To structure the process of analysis and clarify the contextual relevance and meaning of the documents, each document was assigned with an initial category. After all documents had been coded in this way, a second step of unification of codes merged similar document categories into a parent category, resulting in the final count of 33 document categories as listed in Appendix [A.3](#). Table [3.3](#) shows an overview of assigned document types and a tally of the number of occurrences of each type. The total sum of assigned document types exceeds the total number of documents because some documents were classified in more than one category.

Reading and annotation *The final step at the core of the analysis included multiple in-depth readings by at least one member of the research team, as well as direct annotation with codes and comments within the documents.*

The current status of which documents were read/annotated to which extent and by whom was also tracked in the document index, which allowed us to orchestrate the exchange of specific documents that required the various perspectives on the material contributed by different members of the research team.

The document index in its final form is reproduced in Appendix [A.3](#), albeit limited to the fields *Document ID*, *Title*, *Author(s) / Affiliation*, *Type* and *Publication Date* in compliance with the research agreement underlying this case study. Given the fact that most of the documents are not publicly available, and to avoid confusion, references to the index are not listed as part of the general bibliography of this dissertation, but rather in the following form based on the document identifier and hyperlinked to the index:

[\[DOC_1\]](#) or [\[DOC_1\]](#), p. N]

3. METHODOLOGY

| Category Identifier | Document Type | Number of documents |
|---------------------|----------------------------|---------------------|
| AG | Meeting Agenda | 3 |
| CALC | Calculation Tool | 1 |
| CHECK | Checklist | 1 |
| CONS | Consent Form | 1 |
| DEX | Data Excerpt | 3 |
| DI | Discussion Input | 1 |
| DOC | Documentation | 5 |
| GD | Guidelines | 4 |
| GDPR | GDPR-Information | 1 |
| HB | Handbook | 2 |
| IG | Interview Guidelines | 2 |
| INFO | Handout / Info Sheet | 3 |
| IQ | Inquiry | 3 |
| IQR | Response to Inquiry | 6 |
| L | Letter | 2 |
| NO | Notes | 1 |
| POL | Policy Brief | 1 |
| PP | Position Paper | 1 |
| PRES | Presentation / Slides | 10 |
| PROT | (Meeting) Protocol | 50 |
| Q | Questionnaire | 4 |
| QA | Questionnaire (annotated) | 2 |
| R | Report | 14 |
| RA | Rant | 1 |
| REQ | Requirements Specification | 2 |
| RU | Ruling | 5 |
| SC | Screenshots | 2 |
| SPEC | Specification | 2 |
| SUP | Supplement | 2 |
| T | Tender | 2 |
| TC | Terms and Conditions | 1 |
| TD | Target Definitions | 1 |
| WP | Working Paper | 1 |

Table 3.3: AMS document categories and number of related documents

As a consequence of the internal nature of many of these documents, reproducing large swaths of content, longer direct quotations or inclusion of graphics or other visual excerpts are not always possible due to the conditions of disclosure. Some aspects have been reproduced through the creation of our own graphics based on data from the documents, and in select cases short direct quotations were translated and included in the analysis, if the content was not deemed immediately secret or was available in other forms in public documents or statements.

In sum, the complete corpus of documents analysed contained *134 separate files* with a total number of *2491 pages* (excluding the two Excel documents [DAT_1, DAT_2]). All documents were published between 2008 and 2020, with 17 documents whose publication date could not be determined. The majority of the documents originated either with the AMS itself or with the research firm developing the statistical models for the system (Synthesis Research GmbH), with a few notable exceptions, such as documents published by Austrian Federal Ministries, the Austrian Parliament, the Organisation for Economic Co-operation and Development (OECD) or the European Commission.

3.3.1.1 Qualitative Interviews

As stated in the outline of the practical implementation of analysis for this case study above, a single in-person meeting between the AMS, Synthesis Research GmbH and members of our research team occurred in October 2019, before the complete corpus of documents was provided to us. Originally, we planned on utilizing the recording of this meeting as an additional qualitative data source, and transcribe and analyse it following Mayring’s approach [250] to qualitative content analysis. While we did obtain the consent to record, transcribe and utilize the meeting in this way, the topics discussed and questions answered deviated significantly from our initial plan for the meeting, leaving very little substantial information that would have improved our understanding of the AMAS system beyond the insights gained through the document analysis conducted later on in the process. Furthermore, we expected to get another chance at a more in-depth interview focused on the questions we would undoubtedly have after analysing the documents; however, these expectations were foiled by the onset of the COVID-19 pandemic, relegating the second interview to an email exchange of questions and answers. Consequentially, we elected to forgo the more traditional analysis of the interview recording and only transcribed and coded it in a single iteration, marking specific claims or points of interest to be corroborated through and compared against the document corpus later on.

3.3.1.2 Critical Data Studies

Two documents were treated with a different approach in addition to the one outlined above due to their specific content and meaning: The “Interface Definition Model 2020”¹³

¹³Orig. “Schnittstellenbeschreibung Modell 2020”

[DAT_1] and “Regional Agency Types / Model incl. Classification”¹⁴ [DAT_2]. The former contained the definition of all variables utilized by the system, as well as their potential values and the rules required to derive said values for a given jobseeker based on the data available about them. The latter explicated the assignment of one specific variable—the performance of the regional [AMS] offices—to one of the five types/values.

Both documents were core resources for a critical interrogation of the underlying models of the [AMAS] system. While not immediately considered as datasets of actual, live data, they nevertheless contained the meta-description of the data model the system operates on. Following the approach outlined by Poirier [266], these documents were *critically read* through three different modes: a *denotative*, *connotative* and *deconstructive* reading.

First, a denotative reading “*extrapolat[es] the literal meaning*” [266, p.2] of the variables and potential values in a given dataset. The goal of this reading is to determine, on a technical level, which variables were considered by the data producers (i.e., the [AMS] and Synthesis Research GmbH) and what values they could potentially take, as precisely as possible. Through this “*strategically reductionist*” approach, the critical reader “*momentarily assumes a neutral position, not pursuing a neutrality ideal, but instead accounting for the formal semantics that enforce what is understood for “count” in data.*” [266, p.3].

The second reading—the *connotative* mode—contextualizes the dataset by inferring which cultural, political or other contextual influences affected its definition and creation through referencing supplementary data sources. As Poirier summarizes, the aim of this mode of reading is to “*situate data semantics historically and culturally in order to interpret how implied meanings are derived from data.*” [266, p.4], thus looking beyond the literal meaning of the variables and possible values of the data dictionary to determine *why* they were chosen in this way. In the case of the [AMAS] system, the wealth of policy, documentation and meeting protocols available served well to elucidate this contextual information.

The last form of analysis was the *deconstructive* reading: Here, the focus lies on what literal and contextual information is not covered by the dataset—in other words, to determine where the representational limits of its grammar lie in regards to the various realities it purports to cover. Through this reading, the trade-offs made by the data producers are foregrounded to highlight “*absent meanings and unacknowledged tensions that are always already haunting data-based representations*” [266, p.4].

As Poirier concedes in her own case studies utilizing these strategies together with students of [STS], these three modes of reading are not as exclusively distinct as they may appear at first glance; for example, a *deconstructive* reading also requires a *connotative* reading to determine absences and presence of meanings [266, p.4]. Consequently, our analysis also combined the different readings where necessary, with the ultimate goal to “*enmesh numerical representations in powerladen semiotic systems, helping elucidate the assumptions and political commitments on which data rest.*” [266, p.4].

¹⁴Orig. “RGS Typen / Modell inkl. Einteilung”

3.4 Comparative Case Study Methodologies

Building on the two case studies as previously described, Chapter 6 synthesizes the learnings derived from the case studies towards a generalizable evaluation and assessment methodology in the form of the [Algorithmic Accountability Agency Framework \(A³ framework\)](#). In doing so, I appropriate the approach of a *qualitative comparative case study* as described by Bartlett and Vavrus [\[267\]](#) for the fields of Comparative and International Education to the field of [CAS](#). In the following paragraphs, I outline the theoretical foundation for this approach and connect it to the previous methodologies employed for the separate case studies.

Bartlett and Vavrus [\[267\]](#) reject more traditional approaches to comparative case studies following a “*compare and contrast*” logic of comparison in favour of a “*tracing*” approach on *horizontal, vertical and transversal axes* across cases. Traditionally, comparison between cases is founded on a “*variable-oriented notion of comparison*” [\[267, p.7\]](#), tracing back to positivist logics of inquiry in earlier social science research by requiring pre-emptively defined “*units of analysis*” to demarcate the boundary of cases, and thus the subjects of study. In Bartlett and Vavrus’ words:

“Unit of analysis isolates the entity being analysed, the what or who that is being studied, and typically refers to individuals, groups, or organizations [...].”

[\[267, p.7\]](#), summarizing Babbie [\[268\]](#)

By doing so, the bounding units of analysis are considered the constant, while varying other factors are used to test hypotheses; cases are being shown to be comparable precisely because the pre-emptively defined ‘units of analysis’ are the same. While such an approach may well be applicable and useful to argue for the validity of results, Bartlett and Vavrus point to the fact that such a static approach conflicts with the *emergent* nature of qualitative research [\[267, p.9\]](#), which often does not pre-emptively specify “*methods, theory or data*” [\[269, p.548\]](#). Hence, they agree with Becker’s [\[269\]](#) characterization that “[n]ot fully pre-specifying these ideas and procedures, as well as being ready to change them when their findings require it, are not flaws, but rather two of the great strengths of qualitative research.” [\[269, p.548\]](#).

As an alternative to this traditional approach, Bartlett and Vavrus [\[267\]](#) propose a *process-oriented* and *heuristic* comparative case study design that embraces the *iterative* nature of emergent qualitative research by “*divorc[ing] the phenomenon of interest from the context in order to gain analytical purchase*” [\[267, p.10\]](#). Furthermore, they also draw parallels between *ethnographic* methods and comparative case studies, as both share a focus on the perspectives of social actors, and stress the importance of a *critical theory* stance as informative for their approach to examine issues of *power* and *inequality*. Finally, they also interrogate and redefine concepts of *culture* as “*an everchanging, active,*

productive process of sense-making [267, p.11] as well as *context* as “*relational and spatial*” and characterized by mutually influential connections between actors [267, p.12].

Through these arguments, Bartlett and Vavrus derive a methodological concept for comparing case studies along three axes [267, p.14]. First, they suggest a *horizontal* axis that contrasts and traces entities (including, but not limited to) social actors, documents and other influences across cases. Secondly, they encourage the comparison of a *vertical* axes along various scales and scopes applicable to the case studies. Finally, they add a *transversal* dimension to compare and trace commonalities and differences over time, as a reminder that both horizontal and vertical axes “*should be considered historically, but often are not; hence the need for the third axis.*” [267, p.14].

3.4.1 Application to CAS

In applying this concept of a processual comparison to this case study, I set out to trace relevant *social actors* and *influences* both *horizontally* across the EnerCoach and AMAS cases, as well as *vertically* across the various applicable scopes, and additionally consider a temporal dimension as *transversal* axis. Through this comparison, a number of observations emerged, the most relevant of which are detailed in the sections 6.1.1 and 6.1.2 on the differences and similarities between the cases. From a larger perspective, the search for generalizable insights into the nature and operationalization of accountability processes lead to the application of Bovens’ taxonomy [22] for accountability as a mechanism to each of the cases, yielding *horizontal* dimensions of comparison between different combinations of *fora* and *actors* as instances of *social actors*, making them comparable across both system’s assemblages.

Building on the same taxonomy, but including the scoping perspective of scale introduced in Section 2.4.3 as *macro-* vs. *micro-accountability*, a *vertical* dimension of comparability emerged. Considering how these two perspectives compared within each case study, and subsequently, across both of them as well, supports the arguments that (1) both perspectives are relevant and applicable to each case, and that (2) they are indeed interrelated and interdependent instances of similar processes playing out at difference scales.

In addition to these *horizontal* and *vertical* perspectives, I incorporated a *transversal* perspective by considering the case studies not just as static, momentary snapshots of *in medias res* algorithmic systems, but as a fluid and multiplicitous assemblage developing over the time of their lifecycles. In doing so, important comparative observations about the timeliness and nature of socio-technical measures that support potential accountability processes in these systems emerged. In particular, contrasting the differences between the *ex-post* application of participatory design methodologies for EnerCoach, and the *a priori* design and implementation of the explanation text functionality of the AMAS system revealed the consequences for each of those approaches at their respective point in the system’s lifecycles, and the resulting limitations placed upon it.

Finally, by synthesising these three dimensions of comparison and the observations resulting from them, the common influence of *human agency* emerged as a theme applicable to each of the dimensions, and with it the respective questions of hindering or enabling factors. By building upon the established theoretical model of human agency in the form of Bandura's *Social Cognitive Theory* [24] as discussed in Section 2.4.4, and integrating this perspective into the procedural conceptualization of accountability, a coherent guiding framework using emergent human agency as the analytic lens and capable of addressing the variety of social, technical and socio-technical influences shaping the accountability process took form. To refine the initial, primary guiding questions, I applied them to a number of different micro- and macro-accountability processes and scenarios situated within each case study, and iteratively distilled the set of most relevant secondary guiding questions presented in the final framework.

3.5 Chapter Summary

In this chapter, I outlined the general methodological approach underlying this dissertation, as well as the specific methodologies used for each of the separate case studies and the comparative case study.

Starting with the overall approach, I defined the three guiding principles of reasoning behind the subsequent methodological choices as *holistic capturing*, *complementary balance* and *plausibility in light of constraints*. In the description of these principles, I emphasised the importance of covering algorithmic systems as relational socio-technical assemblages in their entirety as opposed to limiting the analysis to only social, technical or socio-technical aspects. Furthermore, I declared my strategy to arrive at a balanced representation of these various aspects, including the value of the comparative case study in ensuring said balance.

For the case study of the EnerCoach energy accounting system, I discussed the methodological foundations of my approach of *algorithmic ethnography* based on Seaver's [21] suggestion to consider *algorithms as culture*. Before detailing the specific methodologies and their concrete implementation across the two phases of the case study research project, I declared the *auto-ethnographic* nature of the study, discussed the value and common criticisms of such an approach, and disclosed, in detail, my involvement and role with the EnerCoach system, and presented the strategies employed to ensure validity of the results. For the first, *analytic* phase of the study, I gave a detailed list of data sources and the qualitative and quantitative methodologies used to incorporate them in the analysis, as well as supplied some key metrics to outline the scope of analysis. For the *intervention* phase of the study, I introduced the theoretical foundations and reasoning behind the use of participatory design methodologies, as well as their concrete implementation as part of the case study.

For the case study of the **AMAS** system, I first disclosed the collaborative nature of the study and introduced my collaborators. I then gave a quick overview of the timeline of how the study came to be, and provided the theoretical foundation informing

the methodological implementation of a *constructionist document analysis*, provided arguments for the value and outlined the limitations of this approach. Next, I detailed the practical implementation of this analysis, including the *creation and curation* of the document index, process of *extraction of bibliographical information*, *assessment* and *classification*, and finally, *reading and annotation*. I also explained the process of qualitative document analysis as analogous to the EnerCoach case study, and detailed the methodologies founded in *CDS* that were used in select cases of document analysis pertaining to the data model of the system.

Closing the chapter, I described the methodology for the qualitative comparative case study resulting in the creation of the [A³ framework](#), and related the theoretical foundations provided by Bartlett and Vavrus [\[267\]](#). I also gave a procedural description of how I appropriated and applied this conceptual approach cross-disciplinarily to the field of [CAS](#) and the two case studies at hand.

As a final note on the structuring of this dissertation in regards to these methodological considerations, in the spirit of brevity, the choice to aggregate the methodological description for both case studies and the comparative synthesis within one chapter was based on the fact of overlapping methodologies (e.g., qualitative text analysis and interviews). Where applicable, the following chapters reference the requisite methodologies and their descriptive sections in this chapter to allow the reader to quickly cross-reference them. Otherwise, the separation of methodology from content and results also aims to preserve the readability and comprehensibility of these following chapters.

3.6 Chapter Conclusions

Considering the range of methodological approaches within and across the case studies, as well as the comparative case study, some conclusions can be drawn about this eclectic selection and combination of, at first glance, rather incompatible or at best unrelated approaches. As I have described in Section [1.2.2](#) on the need for interdisciplinarity in [CAS](#), however, the benefits of transgressing disciplinary boundaries—both theoretically and methodologically—are manifold and ultimately worth the challenges introduced by such eclectic approaches. In line with Danaher et al.'s [\[51\]](#) arguments, mono-disciplinary approaches have, in the past, been insufficient in addressing the complexity and variety of algorithmic systems, both during their development and when evaluating their performance and impacts. Considering the arguments made in the previous Chapter [2](#) on the nature of algorithmic systems and their socio-technical assemblages, the methodologies chosen for these case studies and the approaches of how to compare them are a reflection of that broad, inter-disciplinary view on algorithms as well.

Diverse Contexts Demand Diverse Methods

The EnerCoach case study exemplifies well how ethnomethodologies as an overarching strategy can combine both qualitative and quantitative data sources without denying the interpretative nature of results in order to cover various the heterogeneous facets of that system, including the different stakeholders and their needs, intentions and limitations, the technical components of the assemblage and its impacts on the stakeholders, as well as larger considerations of policy and politics of sustainability in Switzerland and beyond. Precisely because the ‘big picture’ of EnerCoach as a case study involves such diverse facets, incorporating all of them and unravelling the socio-technical assemblage requires a combination of different approaches, often seemingly contradictory or unrelated. The overarching, primary methodological approach of an (auto-)ethnography, however, also informs the way these contradictions can be addressed: As an auto-ethnographic researcher, it is an interpretative act in and of itself to carefully weigh the importance and relevance of the results that these various empiric and theoretical approaches can deliver in order to decide upon inclusion or exclusion of source material. Beyond the analysis, the same logic applies to the choice of interventionist strategies—both in the sense of choosing to intervene at all, and which methodologies for intervention to apply. In the case of EnerCoach, the choice for participatory approaches was the result of such an interpretative act as well. The possibility of employing such an approach due to the levels of access and willingness of the participants played as much a role in this decision as the potential of participatory design methodologies in light of the previous theoretical considerations human agency in socio-technical assemblages. Considering the evaluation of this participative methodology as part of the EnerCoach case study presented in the following Chapter 4, the value of auto-ethnographic approaches also serves as an implicit answer to [SRQ2.1](#) and [SRQ2.2](#).

The case study of [AMAS](#), on the other hand, teaches a more pragmatic lesson in terms of methodological power. Here, an analysis of a complex algorithmic system had to be performed in an adversarial environment fraught with tensions, particularly due to the controversial nature of the system and the public, often contentious, discourse surrounding its announcement. The effect of these tensions was twofold: on the one hand, it meant that the stakeholders of the system, specifically the [AMS](#) and Synthesis Research GmbH, were extremely guarded when it came to allowing access to the system for research purposes, limiting the choice of possible methodologies from the onset. On the other hand, the public discourse surrounding the system brought forward other groups (such as the [AKOÖ](#)) interested in an independent analysis, which ultimately helped pressure the [AMS](#) into agreeing to a collaboration, hesitant as it may have been. Regardless of these limitations, the primary source of information about the system in the form of a trove of documents proved versatile enough when analysed with appropriate methodologies to provide similar insights as those gleaned of the EnerCoach system with a wholly different level of access: a history of the tool and its inception, technical aspects and details, a stakeholder analysis, and the tentative operationalization of the system could all be extracted from those documents.

Implicit and Explicit Comparability

Finally, when comparing these two case studies it seems equally difficult to reconcile their different underlying methodologies as it is to compare the cases themselves. While I address the question of comparability in terms of the nature of these systems in more detail in Section [6.1](#), the question of comparability based on the underlying methodologies must be discussed at a different level. Fundamentally, many comparative studies draw their sense of validity and rigour from the precise application of the same methodologies, implicitly providing an argument for the overall comparability of the involved cases [267](#), [268](#), [269](#). After all, if the same variables, the same methodologies, and the same types of results can be derived from different cases, they must share some intrinsic similarities making them comparable. In the specific case studies presented in this dissertation (as well algorithms in general as subjects of inquiry in [CAS](#)), such similarity in methodologies may often be impossible, or, arguably, at least ill-advised. For these two specific case studies, limits of access simply made applying the same methodologies impossible; neither were the documents available for the EnerCoach system sufficient to characterize the whole system satisfactorily, nor was the kind of deep and unhindered access to code, internal communications, and interview partners for virtually all stakeholder groups even a remote possibility for [AMAS](#). Given the variety of research contexts in which a case study of an algorithmic system may be undertaken, these limitations to applying the same methodologies are to be expected for many other combinations of case studies as well. Beyond these rather practical issues of access, however, the fundamental attributes of algorithmic systems, particularly when considered as socio-technical assemblages, suggest such an *a priori* rigid methodological approach as ill-advised. As I have detailed in Section [2.1.3](#), the *ontogenetic* nature, *relationality* and *heterogeneity* of algorithmic assemblages also means that applying the same methodology repeatedly, even within a single case study or system, are likely to produce the same empiric results. Implicitly, a researcher's engagement with an algorithmic system and its stakeholders also transforms the overall assemblage of that system, as it now includes the researcher and their agenda, the modes of inquiry, questions asked, and—implicitly or explicitly—normative assumptions carried by their research design and implementation. In other words, the fluidity of the assemblage thinking approach also muddies the boundaries of what is the object or subject of research, making any assumptions of comparability based on methodological precision and similarity rather difficult to maintain.

Consequentially, attempts to draw conclusions from the *comparison* of such different case studies that either involve different methodologies by necessity, or similar methodologies applied differently due to the fluid nature of socio-technical assemblages, will require less *algorithmic* (i.e., in the sense of *formulaic* or *prescriptive*) strategies for comparison, and more *heuristic* ones [267](#), p.6]. Thus, in lieu of intrinsic arguments towards comparability, part of the research process must be the crafting of a rationale of comparability similar to the one presented in Section [6.1](#) for the case studies of this dissertation. Such rationale may only emerge during the comparison itself, but also may, as an analytic exercise,

offer further insights into the nature of the systems as well. Bartlett and Vavrus' [267] methodological framework for comparative case studies offers, I would argue, exactly the right compromise between abstract generality and concrete dimensions to inform research strategies beyond their original disciplinary context of *Comparative and International Education*, including, but not limited to research in CAS.

In summary, both case studies embody Seaver's suggestion to "scavenge" [21, p.6] for data where possible, and show that the methodology of analysis must be determined by the (potentially) available source material first and foremost. Given the nature of that material, making well-reflected methodological decisions should mean that researchers should not limit themselves to narrow disciplinary confines, but rather make bold choices to incorporate the knowledge and experience provided by disciplines with a longer history of working with such materials, even if such decisions come at the cost of having to venture into unfamiliar methodological terrain: in the end, when trying to resolve the contradictions and tensions introduced by such methodological eclecticism, researchers might find further insights precisely *because* and not *despite* integrating new perspectives instead of following well-trodden paths. Similarly, comparative inquiry must take into account the nature of algorithmic socio-technical assemblages instead of forcing potential cases into a narrower corset of methodological similarity, and forge its own path towards arguing for the plausibility and validity of its results.

Case Study: EnerCoach

In this chapter, I present the case study of the EnerCoach energy accounting system, a collaborative eco-feedback and assessment tool used by Swiss communities, their energy auditors and consultants to collect data about their energy and water expenditures and assess the sustainability of their energy practices. The case study focuses primarily on the issues of *algorithmic transparency* and *algorithmic literacy* endemic in the system, as well as detailing and evaluating the interventions taken to improve these issues through participatory design methodologies. Starting with an introductory vignette in Section 4.1 to set the stage and highlight the motivation for this case study, I then give an outline of prior research in the domains of eco-feedback tools, civic technologies and energy accounting systems in Section 4.2. At the centre of the case study lies the thick description of the EnerCoach system from a socio-technical perspective in Section 4.3, the results of the evaluation of the system in terms of its transparency deficiencies in Section 4.4, as well as the results of the interventions implemented through the use of participatory design methodologies in Section 4.5. Since the case study's original and primary focus was an investigation into algorithmic transparency, algorithmic accountability in the EnerCoach system will be discussed later in Chapter 6.

The core content of this chapter was previously published as contributions to the GROUP '20 conference [2]¹ and Communities & Technologies '21 [1]² conferences respectively. While the primary results are unchanged, the limitations of conference proceedings only allowed for an abridged accounting of the case study, justifying this in-depth reproduction as part of this dissertation.

¹Publication title: “*Beyond Transparency: Exploring Algorithmic Accountability*”

²Publication title: “*Tackling Algorithmic Transparency in Communal Energy Accounting through Participatory Design*”

4.1 Exploratory Vignette

As outlined in detail in Section 3.2.1 on auto-ethnography, I found myself in the privileged dual role of both a researcher in CAS, and a software developer and support specialist for the EnerCoach Energy Accounting system. From this position, I present the following exploratory composite vignette [60] of my experience as a software developer at WIENFLUSS, which prompted my initial interest in the system and the subsequent research project from the perspective of CAS.

In the spring of 2014, the co-owner of WIENFLUSS and my boss, approached me to ask for an assessment of whether or not WIENFLUSS could take up a project to implement the new, online, and collaborative version of the EnerCoach energy accounting system. The original EnerCoach tool (shown in Figure 4.1) was a rather byzantine set of Excel sheets relying heavily on VisualBasic macros and was still in use by hundreds of communities in Switzerland to monitor and assess their building’s energy and water consumption, and provide reporting for energy audit processes. WIENFLUSS had agreed to implement the frontend parts of the new system (interfaces and design, as well as information architecture) and liaison with the client, but the main, back-end implementation was to be done through a subcontractor. However, at this time, the subcontractor was not able to complete the project, and WIENFLUSS was considering taking on the project on its own. I was given access to the original tool, as well as what was available in terms of specifications for the new tool, and tasked with giving my honest assessment of whether or not we would have the requisite skills and resources to complete this project for our client. After taking a look at the materials provided, I concluded that it would be a challenge, but nothing we could not handle, based on our proven expertise with web applications in the domains of sustainability and environmental policy, such as the online tool for the European Energy Awards (EEA). While the available specifications were rudimentary at best, I nevertheless looked forward to the challenge to develop and implement proper specifications together with our client, the Nova Energie GmbH located in Aarau, and the EnerCoach Working Group of the EnergyCity program for all of Switzerland.

Fast forwarding to 2018, I found myself brooding over an implausibly large Excel table listing the detailed intermediary steps of a complex multidimensional matrix required to generate one of the system’s more complicated reports—the Energy Certificate Report—for a small community in Switzerland. An EnerCoach hotline staffer had written a support request email, asking for clarification of why the results showed an abysmal performance for the community in question in terms of its energy classification: they had double- and triple-checked their data and could not explain the results to the energy consultants that were working with the community on their yearly reports. I was completely stumped at this point—according to my debug outputs, the system was calculating everything to specification, and yet the final and primary energy consumption ratings showed a glaring “G” classification, which meant a strong suggestion the buildings in question were “[i]n need of redevelopment”. After hours of looking through the community’s data, the definitions of its buildings and building zones, electricity meters, thermal production systems and the energy mixes they were feeding into their

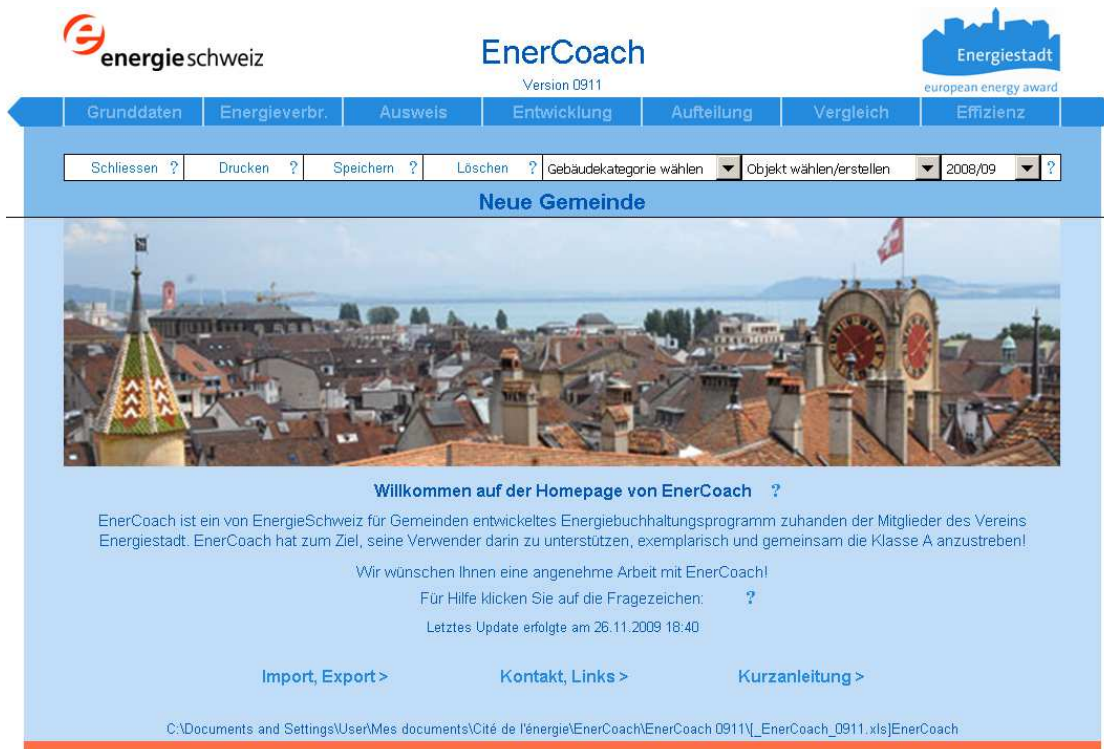


Figure 4.1: “Homepage” of the original EnerCoach Tool implemented in Microsoft Excel, the predecessor to the current EnerCoach Online Tool.

systems, a potential culprit finally emerged: the community was using *direct electrical heating* (e.g., radiators providing heat directly through electricity) for a number of their buildings. As part of the specifications—and dutifully implemented by WIENFLUSS as specified—energy consumption for this type of thermal production system was being counted *twice* for all reports, leading to the surprisingly high energy consumption per m² of the affected buildings and the community as a whole. This calculation was one of many hidden exceptions to the general design of the system, which otherwise mandated a bit-by-bit aggregation of energy consumptions and reference areas exactly as provided by the community. Digging into the code of the current implementation, the tool’s specifications, as well as the documentation and specifications for the original Excel Tool, I was able to find both the code snippet responsible as well as a reference in a footnote of the specifications (see Figure 4.2) describing this behaviour as *a feature and not a bug*. Following a long line of breadcrumbs finally led me to the underlying reasoning: At the time of specification for the predecessor of the current system, it was a stated policy of the EnergyCity program to discourage the use of such heating systems given their inefficiency, as well as the fact they were often powered by electricity from non-renewable sources, and thus were not considered to be a sustainable way of heating a building. As a consequence, those energy consumptions would be “[...] counted double for the energy

key figures in accordance with EnergyCity policy, but only counted once for the absolute values (e.g., evolution of consumptions)”³ [270, p.2]. I put my findings into an email several pages long, sent it off and closed the support ticket as solved.

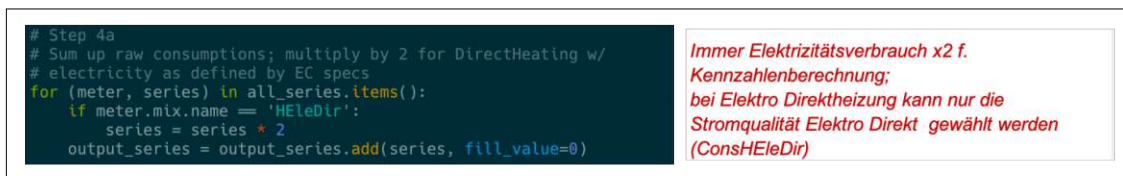


Figure 4.2: The original code snippet responsible for doubling energy consumptions produced by direct electrical heating, and the respective reference in the specification documents roughly translating to “Always electricity consumption $\times 2$ for keyfigure calculations; for direct electrical heating only allow *Elektro Direkt* (*ConsHEleDir*) as electricity qualifier.”

Having completed the support request, the issue nevertheless stayed with me. Far from the only support request of this kind, I spent a significant amount of my work time digging through system logs and debug calculation outputs to answer similar questions. Sometimes these observations ended up pointing to actual bugs in the system, but more often the system was performing exactly *as specified*, yet clearly not *as expected* by its users. Finding the issue was a laborious task, often requiring hours of tracing calculations through code that I had personally written, based on specifications I had created in close cooperation with the client. Why was this such a challenge? Did we not follow best practices of software development in our implementation, incurring tech debts that now made answering these questions so complicated? Was the system intentionally opaque towards its users in an effort to discourage falsifying data or “gaming the system”? And if not, how were the users or even the hotline staffers—experts in the domain of energy accounting—ever supposed to be able to figure these things out for themselves if I, as the lead developer of the system, needed hours of debugging and access to the innermost parts of the application otherwise hidden from users to find the answers myself?

As a researcher in [CAS](#), my interest was piqued. At first glance, I realized there were multiple issues at play, none of which had an obvious or simple solution. First, it occurred to my how much translation work was required to resolve even just a single instance of a request as the one outlined above: between the users formulating their inquiry to the hotline, the hotline investigating and adding their own input, and the final request to WIENFLUSS to investigate, the simple act of formulating the question already required multiple steps of formulating and re-formulating the problem through various levels of technological, algorithmic and energy accounting domain knowledge. Providing an answer meant one had to go through the same process in reverse. Miscommunications and conflicts were par for the course in such a process, and yet none of the usual approaches recommended as part of standard software development best practices seemed promising

³Orig. “Der Verbrauch für die Energiekennzahl wird doppelt gerechnet gemäss Energiestadt, jedoch nur einfach für die Absoluten Werte (z.B. Verbrauchsentwicklung) verwendet.”

enough to resolve the issue on their own. Neither would it be plausible to document every possible combination of data entered to explain the result, nor would *automated reasoning* be capable of including the cultural and contextual meanings and reasoning embedded in the system's behaviour.

Only one thing was blatantly obvious: The burden placed on the users, administrators and even the developers of this system was unsustainable due to its lack of explainability and transparency. Consequently, an arguably crucial tool meant to empower communities in their fight against climate change was held back from fulfilling its potential and might even be shut down due to a lack of accountability! How could users understand and trust a system whose outputs frequently confounded even the experts? Would anyone keep using a system that was as complex, opaque and unaccountable, or would they look for simpler, but also less comprehensive, accurate and effective, alternatives?

At this point, I realized the unique and privileged position I found myself in. Here, for once, was a system that I could take apart and analyse bit by bit, including its technical, socio-technical and social figurations. I could account for more meaning and reasoning about certain implementation details than any researcher studying a system from the outside ever could—since I had actually written the code and knew why I made the decisions that shaped the system's technical side. What I needed to investigate further were the socio-technical and social aspects: What reasoning determined these specifications? Who was actually using the system, and in which ways were they using it? And—finally—what measures, be they technical or otherwise, would the users and stakeholders themselves need to improve the situation? Armed with these questions, I approached the EnerCoach Working Group, WIENFLUSS and the [C!S](#), and proposed this case study to figure out if the transparency of the system was truly as much of a problem as the frequent issues suggested, and if so, what we could do about it—both from a scientific standpoint and a practical one.

4.2 Prior Research

As the world gradually learned to accept the grim reality of catastrophic, human-made climate change over the past decades, scientific interest in technologies addressing climate change through mitigation has been substantial. In their latest report on climate change *mitigation* [271], the [Intergovernmental Panel on Climate Change \(IPCC\)](#) confirms the potential of digital technologies towards achieving several of the [Sustainable Development Goals \(SDG\)](#) put forward in the UN 2030 Agenda for Sustainable Development [272]. They specifically mention sensor technologies, the [Internet of Things \(IoT\)](#) and [AI](#) as contributors to improving energy management and efficiency, but also warn of losing part of these gains to increased demand for goods and services due to the use of digital devices [271, p.14].

Digital tools supporting the *quantitative measurement* of energy consumption/CO₂ emissions and a *normative assessment* of sustainability performance rely on these and similar digital technologies as well, as the wealth of available research shows. Generally speaking,

research on these technologies in the scientific disciplines relevant to this dissertation can be categorized by the intended target audience as individuals or organizations either with or without expert domain knowledge. Starting with the former, research into *eco-feedback tools* tends to focus on creating tools that help mitigate climate change by inducing individual behavioural changes resulting in more sustainable energy practices. Studies target a range of application domains, including mobility behaviour [273, 274, 275, 276], household energy consumption [53, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286], sustainable workplace practices [283, 287], and water consumption [288]. Other related aspects of eco-feedback tools include research into energy literacy [281] or the impact of visualizations [289, 290] on user understanding and behaviour.

While a complete literature survey of the available research into eco-feedback tools transcends the scope of this case study, some common themes across the research landscape can be identified. First and foremost, the majority of eco-feedback tools is founded on the understanding that users can be ‘nudged’ [291] towards more sustainable behaviours by simply using the tools provided. Subsequently, much attention has been paid to the way information is presented to the target audience, drawing on decades of research in HCI on user interface and interaction design, and visualizations. For instance, Meurer et al. [292] report on their case study for design requirements of an app for the elderly aimed at “*fostering the adoption of sustainable mobile behavior*” [292, p.1], in which they used low-mid fidelity prototypes in interviews to gain insights into the requirements of their target audience. Similarly, Spagnoli et al. [278] present a mobile interface called ‘EnergyLife’ for giving eco-feedback on household electricity consumption based on insights from environmental psychology and feedback intervention. As a final example, one of our own previous research projects ‘*eRollin’ on green*’ [8] focused on visualizing CO₂ expenditures for users of eScooters, comparing actual emissions with hypothetical emissions of alternative modes of transport (e.g., by car, train or bus). In our study, we were particularly interested in how well our online mapping and visualization tools were received in terms of user trust and understanding, and the overall impact this perceived transparency had on the participant’s motivation to change their mobility behaviour.

With all this focus on persuasive technologies, it is crucial not to overlook the limitations of such an individualist and prescriptive approach. As Brynjarsdottir et al. [293] point out, “*persuasive sustainability can [be] understood as a modernist technology that works by narrowing its vision to define sustainability as resource optimization pursued by individual rational actors conceptualized apart from the messy realities of everyday life*” [293, p.8]. Eco-feedback tools may well encourage behavioural change in specific individuals, but do so on a very limited scale, and their efficacy is inherently limited by the individual’s agency for change. Beyond any factual impact single individuals can have, their willingness to exercise their agency may also be impacted by their limited *belief in self-efficacy* [217] of their actions: after all, it is difficult to grasp the impact individual decisions can have on the global scale, and the personal cost of behavioural changes may well dwarf their perceived larger impact, thus limiting individual efforts to enact such changes. A strong

focus on individual responsibility to address climate change also may serve as a “straw man” argument to avoid systemic change on larger scale, e.g., by enacting sustainable policies for industries such as global shipping and logistics, or industrial manufacturing. Addressing this problem requires different tools, built for different audiences, which leads to other types of digital climate change mitigation technologies aimed at a more adept target audience with greater domain knowledge: *energy accounting tools*. While many of the above-mentioned eco-feedback tools provide some accounting functionalities, *energy accounting tools* are doing so on a larger scale and moving beyond an individual’s area of influence, ranging from small, medium and industrial enterprises, communities and municipalities, to regions, states and countries as a whole.

Research into energy accounting goes back as far as the early 1980s (e.g., [294, 295]), albeit less focused on the environmental impact of energy consumption than on cutting energy costs. As a notable exception, Cornwall’s “*Guide to Energy Accounting*”, published in 1984, explicitly declares “*motivating staff and students to save energy*” [296, p.1] as a primary goal for their system. The advent of the digital transformation allowed much more sophisticated methods for collecting data (often in real time), analysing and visualizing it. The majority of attention in this field is focused on the various methodologies for calculating and estimating energy consumption and CO₂ expenditures, rather than being concerned with the way this information is presented to users [297, 298, 299]. Other approaches are tackling domain-specific issues related to energy consumption, including energy accounting in mobile devices, embedded systems or other IoT hardware such as wireless sensor networks [300, 301]. Energy Management Systems (EMSs) are related to energy accounting systems and have received some attention from researchers in recent years [302, 303]. The goals of EMSs, however, differ from energy accounting, including (semi-)automated energy optimizations for buildings and energy grids, and—as Schminke [302] points out—are as of yet mostly focused on the theoretical potential of these systems rather than their real-world implementation. Subsequently, the existing literature on EMSs is hardly applicable to energy accounting systems.

Given the importance of energy accounting systems for both economical and ecological reasons, and the complexity of methodologies related to these systems, it is surprising that no significant research efforts seem to have focused on issues of transparency for specific systems to date. While the need for transparency in energy consumption and management practices is well established—exemplified by the existence of the ISO 20001:2011 standard describing requirements and approaches for sustainable energy management practices [304]—research into how well these existing, concrete systems perform in terms of transparency is scarce. McGlenn et al.’s usability study of their Building Energy Management System (BEMS) *BuildVis* [305] comes closest: while not exactly an energy accounting tool, *BuildVis* provides some similar functionalities and faces comparable challenges in terms of visualizing complex data for heterogeneous audiences and users (e.g., facility managers or energy auditors). This is as close as it gets to a human-centric analysis of energy accounting tools; while numerous commercial products exist that offer energy accounting functionalities (e.g., [306, 307, 308]), no

studies evaluating such systems seem to exist to date. This may well be due to the inherent issues of intentional opacity outlined in Section 2.3.2, as companies offering such services may have a strong interest in keeping the internal processes and algorithms of their products secret. This lack of transparency, however, is even more frustrating given the fact that energy accounting practices are often celebrated as a means for improving transparency of energy practices themselves [304], and—*nomen est omen*—are supposed to provide a form of *accountability*.

This lack of prior research into issues of transparency in energy accounting systems notwithstanding, some comparative insights can be gleaned from the attention individual eco-feedback tools have received. Understanding the outputs of such systems in the form of various reports and visualizations can be framed as *sense-making* processes, or, as Lebiere et al. put it, trying to deduce “[...] a meaningful and functional representation of some aspects of the world.” [309, p.1]. Sense-making is the cognitive activity involving “*framing, elaborating and reframing data*” in an iterative manner (Tellioglu et al. [10, p.2], citing Klein et al. [310]). Wood et al. [311], for instance, study an eco-feedback tool for households in the UK through the analytic lens of sense-making, and evaluate the transparency and explainability of the system’s ‘Energy Dashboard’. Their results show how contextual information such as indoor environmental conditions and advice for energy-saving behaviours can help user’s sense-making activities, but also point out how these processes may yield very different outcomes depending on the households or persons involved. When applied to energy accounting tools, these insights suggest the challenge of facilitating successful sense-making processes and subsequently providing better algorithmic transparency may well be even more substantial for the more complex, distributed and abstract outputs in energy accounting systems on a larger scale.

4.3 Socio-Technical Description of the EnerCoach System

The EnerCoach system is a web-based collaborative energy accounting tool. It is primarily in use in Switzerland⁴ and, as of 2022, over 700 communities, municipalities and other communal organisations are using the tool regularly to both collect and visualize data about the energy consumption for electricity and heating, as well as water consumption, of their buildings. The EnerCoach tool occupies a peculiar space between the more individual-focused eco-feedback tools, and the more expert- and industry-focused energy management systems: while its capabilities certainly compare to the reporting and accounting aspects of energy management tools, its design also takes the normative characteristics of eco-feedback tools: not just a descriptive accounting tool, the EnerCoach system was always intended to nudge communities towards more sustainable energy practices beyond the legal requirements mandated by Swiss federal law, pointing them towards buildings and facilities in need of renovation or refurbishment, and rating their performance in terms of energy efficiency and CO₂ emissions.

⁴Since 2019, an extension to a pilot community in Ukraine exists, but is not seeing much use, particularly in light of the current geo-political situation due to Russia’s invasion in early 2022.

Beyond an intrinsic motivation to improve their sustainability in terms of energy consumption, the tool also fulfils an additional function for the Swiss communities that are part of the EnergyCity⁵ program. EnergyCity functions as the Swiss branch of the EEA⁶, alongside other other national branch organisations such as the French *Cit'ergie*⁷ program run by Agence de la transition écologique (ADEME), Austria's *e5*⁸ program supported by the Austrian Energy Agency (AEA), and Luxembourg's *PacteClimat*⁹. The EEA program and their national implementations are, at their core, a certification program that requires at least some kind of energy accounting to be done by the participating communities. One EC Working Group member describes the connection between the EnergyCity certification and energy accounting as follows:

*“The concrete connection between EnergyCity and energy accounting was and still is today that EnergyCity, naturally, has certain criteria that need to be evaluated: What is the state of community buildings? How good are the energy key figures for electricity and heating? What energy carriers are being utilized by the community, are they completely made up from fossile fuels or sustainable and renewable?”*¹⁰

EC Working Group Member “A”

As such, the EEA program and EnergyCity as Switzerland's national implementation offer participating communities a catalogue of potential measures they can take to improve their performance in sustainable practices, including (but not limited to) energy consumption and CO₂ emissions. Communities participating in the program go through a mandatory audit cycle, in which their performance in one of six overarching categories is being assessed in terms of their *implementation quality*. Communities reaching certain milestones of improvements will receive the EEA and EEA Gold labels, which serve as an international benchmark and allow the comparison of different local authorities across Europe [312]. As part of this catalogue of measures, communities must fulfil the requirement of energy accounting for communal buildings and installations. The point system weighing the impact of certain measures in this catalogue is designed in such a way that a community has virtually no chance of reaching an EEA Gold certification without fulfilling this requirement, making an established, certified energy accounting system—such as the EnerCoach tool—a requirement for almost all communities taking part in the Swiss EnergyCity program and thus, the EEA as a whole. Given the fact

⁵Orig. “EnergieStadt”, see <https://energiestadt.ch>

⁶See <https://www.european-energy-award.org>

⁷See <https://territoireengagetransitionecologique.ademe.fr/>

⁸See <https://www.e5-gemeinden.at/>

⁹See <https://www.pacteclimat.lu/fr/acteur-engage>

¹⁰Orig. “[D]er konkrete Bezug zur Energiebuchhaltung war oder ist auch heute noch, dass Energiestadt natürlich entsprechende Kriterien hat, die es zu bewerten gilt: Wie gut ist der Gebäudebestand einer Gemeinde? Wie gut sind die Energiekennzahlen Wärme und Strom? Mit welchen Energieträgern arbeitet die Gemeinde, ist das völlig fossilbasiert oder eben nachhaltig und erneuerbar?”

that EnerCoach was provided to Swiss communities free of charge until recently, and that competing tools for the Swiss market can be both costly and may lack features required for the EEA certification, the fact that the EnerCoach tool was the first choice for many (particularly smaller) communities in Switzerland comes as no surprise.

Early predecessors of the tool can be traced back to at least 1995, as one of the original members of the EnerCoach Working Group explained: a simple Microsoft Excel table, relating yearly energy consumption for electricity and heating to the *energy utilization reference area*¹¹ to calculate the *energy key figures* in kWh/m² for a single building. Over the following decades, the Excel document used to track this information evolved into a highly complex and sophisticated energy accounting tool for the Swiss communities. By the early 2010s, this tool had stretched the boundaries of what could be achieved in a single Microsoft Excel sheet, and plans were made for a complete re-implementation of the tool on a modern platform. The primary aim of this re-implementation was to centralise energy accounting into a collaborative platform, reducing the mess of exchanging single Excel files for each community between the various stakeholders of the energy accounting and EnergieStadt auditing/certification process. To this end, complete compatibility with the original Excel document was a hard requirement, and import routines transferring data from previous years were the entry point for many of the communities using EnerCoach today. Other benefits of the re-implementation included multi-lingual capabilities to reflect Switzerland's unique cultural landscape of the four prevalent languages (Swiss) German, Italian, French and Rhaeto-Romanic¹².

To arrive at a *thick description*²⁷ of the EnerCoach algorithmic system and to unpack its socio-technical assemblage, the following sections outline the results of the *stakeholder analysis* and give a technical and functional description of the system and its constituent parts.

4.3.1 Stakeholder Analysis

As described above, the widespread use of the EnerCoach system and its embeddedness within the larger contexts of the Swiss EnergyCity and EEA programs leads to a heterogeneous set of stakeholders, each with their own needs and requirements towards the system in terms of functionality and particularly, algorithmic transparency. To avoid terminological confusions, the term 'stakeholders' refers to the definition [313, p.451] commonly accepted within the HCI and CSCW communities, i.e., any individual or group of people affected by or interacting with the system.

The stakeholder analysis performed through the use of coding the qualitative interviews described in Section 3.2.2 revealed the following groups of people, illustrated with their various relationships and interactions in Figure 4.3.

¹¹Orig. "Energiebezugsfläche"

¹²Of those four, only German, Italian and French were part of the initial I18N localisation in addition to the English language baseline implementation; in 2019, Ukrainian was added as an additional language to support the Ukrainian pilot community of Zhytomyr.

EnerCoach Stakeholder and User Groups

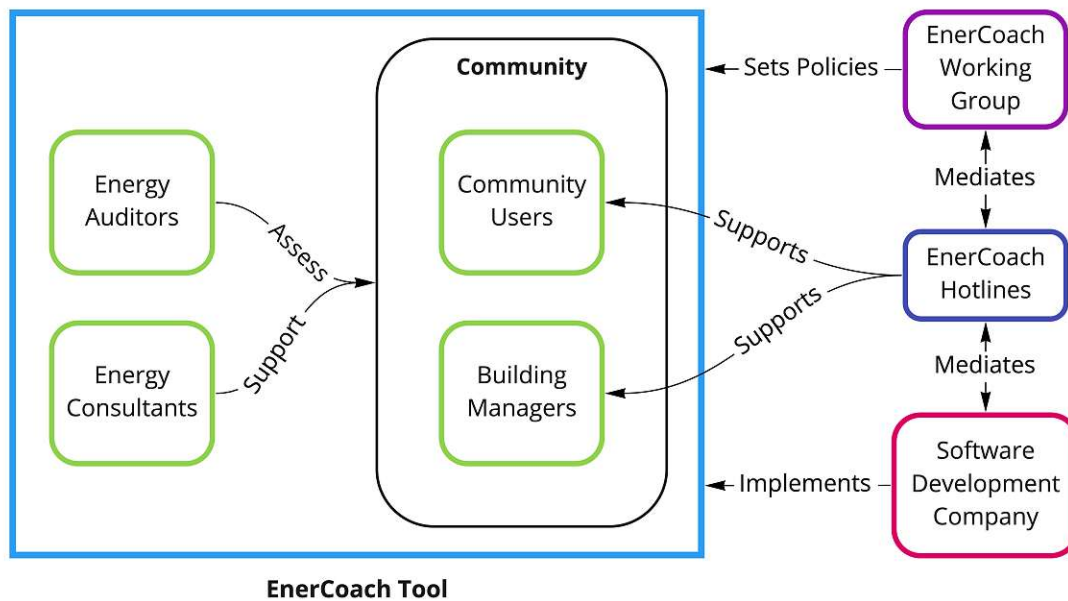


Figure 4.3: EnerCoach stakeholder groups and their interactions.

First and foremost, the *EnerCoach Working Group* consists of various energy accounting and Swiss energy and sustainability policy experts. They serve as a board or steering committee for the system, defining both the high-level goals of the EnerCoach system and their concrete implementation in the form of calculations needed for the reporting system. Until recently, they were also the entity negotiating primary funding for the development and maintenance of the system with the Swiss EnergyCity program and its (political) funding partners, and in turn, respond to the wishes of the EnergyCity stakeholders for future developments of the EnerCoach system.

As a consequence, both past and current iterations of the system were heavily influenced by the EnergyCity program as a primary financier:

“EnergyCity had and still has a certain influence, since certain measures from their catalog must be evaluated using EnerCoach’s reports. With every catalog extension, new evaluation criteria were added. [...] So EnergyCity added a requirement, they want to evaluate the electricity consumption, but also evaluate the electricity mix. Electricity was, in the original Excel tool, just measured as kilo-watt-hours, and now they want to know how much of that is from renewable sources, which quality, and ideally certified as Nature Made Star.”¹³

EC Working Group Member “A”

The EnerCoach Working Group was also responsible for the initial design, iterative development and maintenance of the previous, Excel-based EnerCoach tool; its members have extensive policy and energy accounting experience, having worked closely with communities as energy consultants, and, in some cases, still work for or own energy consultancy agencies themselves.

WIENFLUSS is the software development company implementing the EnerCoach tool, maintaining the system on technical level, and providing technical support. The technical implementation is based on specifications co-developed between representatives of the *EnerCoach Working Group* on the one hand, and *WIENFLUSS* on the other, with the working group members providing non-technical policy specifications, and *WIENFLUSS* transforming them into viable, implementable, and technical specifications. *WIENFLUSS* also provides the UI/UX design expertise, and is responsible for the overall look-and-feel of the system and its information architecture.

The last group of stakeholders working adjacent to—but not usually in an executive capacity with—the EnerCoach tool are the staffers of the *EnerCoach Hotlines*. Since Switzerland has a unique, multi-lingual population, with regions primarily using the German, French, Italian and Rhaeto-Romanic languages respectively, different hotlines exist to provide language-specific support to these communities in German, French and Italian. They are often the first point of contact for questions raised by users of the system, helping to troubleshoot problems or giving advice on how to represent various regional building types, energy mixes and other local idiosyncrasies of Swiss communities. Situated between community users, the EnerCoach Working Group, and *WIENFLUSS*,

¹³Orig. “Energistadt hat da nach wie vor Einfluss indem gewisse Maßnahmenpunkte aus dem Katalog eben anhand der Auswertungen von EnerCoach bewertet und beurteilt werden. Durch die Erweiterung des Katalogs kamen auch neue Bewertungskriterien dazu. [...] Es kam der Anspruch natürlich von Energistadt, man möchte den Stromkonsum bewerten, aber man möchte auch den Strommix bewerten. Strom war in der ursprünglichen Excel Versionen einfach Strom in Kilowattstunden und jetzt möchte man wissen wie viel davon erneuerbar, in welcher Qualität, im besten Fall noch Nature Made Star zertifiziert.”

they also mediate policy questions and relay technical questions and bug reports between those stakeholders. Additionally, they also provide regular training sessions introducing the basic functionalities and reporting capabilities to new users of the system. Finally, they publish an email newsletter with the latest announcements, changes to the system, and other noteworthy information for interested users. As such, the staffers of the *EnerCoach Hotlines* generally have a solid grasp of energy accounting methodologies as well as most user-facing aspects of the system, but lack the technical insight into the inner workings of the system as well as the ‘big picture’ policy and decision-making processes of the EnerCoach Working Group. They are, however, the pre-eminent experts on user experience and concerns, having first-hand knowledge of the kind of questions and issues users commonly face when using the system, and influence the continuing development of the EnerCoach tool by relaying these in the form of change requests to WIENFLUSS and the EnerCoach Working Group.

Finally, the *end-users* of the system are a diverse group of individuals with varying degrees of expertise in energy accounting in general, and the EnerCoach tool specifically. *Community representatives* are members of the local authorities such as the mayors office or a department of the local administration tasked with energy accounting, sustainability efforts or the EnergyCity certification process. Their expertise can vary greatly, ranging from building or facility managers tasked with energy data entry only, to administrative employees entering energy costs and energy mixes and preparing the reports, up to domain experts responsible for the whole energy accounting process from start to finish. Given the complexity of energy accounting on a community level, many of those users have little to no knowledge of the underlying methodologies prior to their first contact with the system. As mid-level employees of the local administration, they are seldom the ones making the decision to use the EnerCoach tool in lieu of other energy accounting systems either, having no choice but to adapt to the tool’s workflows. For those users, the training sessions provided by the EnerCoach Hotlines are vital sources of information, as they may be the first and only introduction to energy accounting and the EnerCoach tool itself they receive.

While the fact that such end-users make up a significant portion of the EnerCoach tool’s user base remained undisputed by the interview partners, whether or not this was intentional and to be seen as a good thing was the subject of some controversial discussion between the EC Working Group members in the first group interview.

One such member explicated the problem as follows:

“Most of those communities, when they first come in contact with energy accounting, they take a look around and then they surely see someone, often from the financial accounting offices—because they have to work with the [energy] bills anyway—and then they say ‘You could just collect that data!’ [...] [But] at the same time we don’t task an employee from the building authority with doing the community chart of accounts? So why shouldn’t we have similar expectations for our tool - it seems an important question to me, [...] where do we say ‘Stop! This is really basic [knowledge], and we won’t spend a single minute to explain this anymore!’¹⁴

EC Working Group Member “A”

At the same time, both the EC Working Group members and the energy consultants interviewed recognize the fact that the tool was (at the time of the study) still available to communities free of charge as a crucial asset in promoting energy accounting as a methodology for more sustainable energy practices, as the lead developer of the original Excel tool describes:

“For us it was and remains a low-threshold introductory tool for communities to explore this topics of ‘energy’ in the broadest sense, because it has, as I said, a very low threshold, you can get an overview of the energy state of affairs of community buildings and, building on these findings, deduce further measures.”¹⁵

EC Working Group Member “C”

For many communities, the required expertise and resources to effectively use the EnerCoach tool can be prohibitive in terms of an autonomous implementation of energy accounting. While some may try to use the tool on their own despite their lack of expertise and resources, as illustrated above, many communities recognise the need

¹⁴Orig. “[D]ie meisten Gemeinden, wenn sie sich mit dem Thema Energiebuchhaltung auseinandersetzen, dann schauen sie so in die Runde und dann gibt’s sicher jemanden, meistens in der Finanzverwaltung—weil er da die Rechnungen so oder so durchgehen muss—, und dann heisst es: ‘Du könntest doch die Daten erfassen!’ [...] [Aber] gleichzeitig stellen wir ja auch nicht irgendeinen Mitarbeiter vom Bauamt hin um den Kontenplan in der Finanzverwaltung zu führen, oder? Und warum sollen wir für unser Instrument nicht auch ähnliche Ansprüche haben - und mir scheint’s eine wichtige Frage zu sein, [...] wo sagen wir irgendwann mal ‘Stop! Also das sind Basics, und da verwenden wir keine Minute um irgendeine Erklärung abzugeben!’”

¹⁵Orig. “Für uns war und ist es auch ein niederschwelliges Einstiegs-Instrument für Gemeinden für diese Thematik ‘Energie’ im weitesten Sinne, weil es wie gesagt sehr niederschwellig ist, man kann sich mal einen Überblick verschaffen zum energetischen Zustand der kommunalen Gebäude und aufbauend auf dieser Erkenntnis weitere Maßnahmen ableiten.”

for expert support and often rely on *Energy Consultants*¹⁶, to whom they outsource their energy accounting processes either in parts or as a whole. An *Energy Consultant* “[...] executes the [certification] process with the community through the entire catalog of measures, and describes the status quo. Then they work out the planned energy policy measures that result from this status quo for the following four-year-period.”¹⁷, as EC Working Group member “B”—also working as an energy consultant for numerous Swiss communities—put it. Energy consultants thus are domain experts in energy accounting, and often work with a number of local communities. As such, they may be contracted by a community to (1) perform data entry based on documents provided by the community itself or the energy companies providing electricity to the community, (2) perform data checks and plausibilisation of reports generated by the system, (3) provide their own interpretation in the form of energy reports including suggestions for improvements, (4) manage the audit processes necessary for EnergyCity / EEA certification, or (5) all of the above. *Energy Consultants* usually have more than a passing knowledge of the EnerCoach tool (as well as other energy accounting systems available for Switzerland), and are knowledgeable of the requirements for a certification as part of the EnergyCity and EEA programs. Depending on how long they have worked with the communities in question, their knowledge of the local community’s buildings and facilities, energy accounting history, and local stakeholders involved in the process may vary.

Finally, for those communities taking part in the EnergyCity / EEA certification process, *Energy Auditors* may be given access to the EnerCoach tool as part of the audit process. Often working closely with *Energy Consultants* or the community users, they verify the correctness of the data entered and subsequent reports generated by the system, and ratify the community’s implementation performance of the measures suggested by the EEA program. Auditors are auditing the community as much as they do their energy consultants, as one EC Working Group member put it:

*“[T]he auditor is also an EnergyCity consultant, but has the additional role to ensure, in so-called audit- or re-audit-sessions, that the evaluation benchmarks are applied roughly equally across communities and that the consultants evaluate these measures based on consistent criteria.”*¹⁸

EC Working Group Member “A”

As external audit cycles for the EEA program are happening in four-year-increments, they may not have recent knowledge of a community and their buildings, and must rely

¹⁶Often, these consultants are directly affiliated with the EnergyCity program itself, and are, subsequently, called EnergyCity consultants or “Energierstadt-Berater” in German.

¹⁷Orig. “[...] vollzieht den [Zertifizierungs-]Prozess mit der Gemeinde über den ganzen Maßnahmen-Katalog, er legt den Ist-Zustand dar. Er erarbeitet mit der Gemeinde dann die energiepolitischen Maßnahmen, die aus dem Ist-Stand abzuleiten sind für die nächste 4 Jahres-Periode.”

¹⁸Orig. “[D]er Auditor ist auch Energierstadt-Berater, hat aber die zusätzliche Rolle in einer sogenannten Audit- oder Re-Audit Sitzung dafür zu sorgen, dass die Beurteilungsmaßstäbe in allen Gemeinden etwa gleich angewendet werden und dass die Berater:innen nach einheitlichen Kriterien die Maßnahmen bewerten.”

either on the EnerCoach tool itself or the additional information provided by their point of contact within the community.

The stakeholder analysis presented above highlights the strong disparities in various levels of domain knowledge, knowledge about the local community, and expertise in the use of the EnerCoach tool. Interactions between those stakeholders are thus often characterized by a significant translation effort. Consider the following hypothetical support request scenario¹⁹:

“A building manager for the community of Châtel-St-Denis has entered the electricity, gas and water consumption data for a given year into the system for the buildings they are responsible for. The administrator of this community has generated the reports for these buildings and notices a significant improvement in the energy rating generated by one of the reports. Since they have no obvious explanation for this improvement, they set out to investigate cause and plausibility of this result.”

To verify this report, they have to confirm the data was entered correctly with the building manager, and subsequently contact their community’s contracted energy consultant to check the report for errors. The consultant notices the gas consumption for the heating of one of the buildings has decreased in comparison to the previous years; they suspect this to be the reason for the better rating. At this point, they may try to contact the building manager directly to find out what may have prompted such a decrease; they may also contact the EnerCoach hotline to confirm that the decrease in gas consumption would correspond to the improvements in the building’s rating. If the hotline suspects a fault in the system, they would file a support request to WIENFLUSS to verify that the calculations are correct. Depending on the response, the hotline would then relay that feedback to the energy consultant, who, in turn, would do the same for their community contact. At the end of this process, the reason for the improved rating may have been a fault in the system, a mistake in the manual data entry, or—as was the case in one of the actual examples provided by a hotline staffer—simply the fact that the building in question was part of a school complex that was shut down for a semester due to renovations, during which parts of the building were not being heated any more. As benign as this example may seem at first glance, it illustrates well how each and every interaction between the affected stakeholders requires various levels of translation work to combine the expertise on the local community, energy accounting methodologies as well as knowledge about the EnerCoach tool in order to resolve the issue.

As a final observation, it is worth noting that the stakeholders of the system may not always be cleanly attributable to one or the other group, and there may, in fact, be

¹⁹This scenario is based on various anecdotes gathered through the interviews with EnerCoach Hotline staffers, energy consultants and the EnerCoach Working Group. While all separate interactions are based on real occurrences, the combined scenario is a hypothetical one to illustrate the complexity of the process.

significant overlaps between the groups in some cases. *Energy Consultants* may work as *Energy Auditors* for other communities; *EnerCoach Working Group* members may have been *EnerCoach Hotline* staffers previously; and community administrators may take courses or continue their education in energy accounting to further their career opportunities, or even join the hotline staffers. These characteristic of overlap were also reflected in the selection of interview partners, most of which also had experience in more than one of these roles (see Section 3.2.2 for more details on the interviewee’s expertise). Consequently, this stakeholder analysis can only serve as an overview indicative of the various groups and their needs in general, and requires further in-depth analysis when looking at specific interactions or issues, as one and the same person may act as representative of different stakeholder groups across different contexts.

4.3.2 Technical Implementation

The EnerCoach system is a web-based, collaborative online application available publicly at its main domain name <https://enercoach.energiestadt.ch>, as well as language-specific domain aliases²⁰. The application is hosted on a server in an Austrian data centre; a staging instance is provided by WIENFLUSS on one of their own servers for testing purposes.

The system is implemented as a single, monolithic web application, including both frontend and backend parts. The technology stack in use includes [Python 3.7](#) as the underlying programming language, [Redis 6.0.6](#) as in-memory caching solution, [NGINX 1.21.5](#) as a web server and [UWSGI 2.0.18](#) as the application server. The requisite data is stored in a [Postgresql 12.9](#) database; analytics, APM tracing and log file management is performed by an off-site instance of the [Elastic Stack \(ELK\)](#).

The python web application is based on the [kotti](#) and [pyramid](#) frameworks respectively. Various other python libraries are included in the implementation, most notably [SQLAlchemy 1.2.15](#) as a database middleware, [Pandas 1.0.1](#) as data analysis framework, and [Highcharts JS 9.0.1](#) as graphing and visualization library.

The following sections provide a short introduction to the EnerCoach data model, the core functionalities available to users of the system, and an overview of the various reports the system can generate. A full and complete description of all models, features and reports—while possible after the code review—would transcend the scope of this case study and, arguably, this dissertation. Furthermore, some detailed information is sensitive and cannot be made available to the general public for security and privacy reasons. Thus, this introduction aims to establish the necessary basic terminology used in the system, provide a sense of scope, size and complexity of the system, and allow the reader to gain enough insight to follow the results of the case study presented in the subsequent sections.

²⁰These include <https://enercoach.citedelenergie.ch>, <https://enercoach.cittadellenergia.ch> and <https://enercoach.enefcities.org.ua> for the French, Italian and Ukrainian versions respectively.

4.3.2.1 Data Model

The EnerCoach data model represents the core assumptions underlying established concepts of energy accounting for buildings and communities, albeit including some specific adaptations for the Swiss context.

The database objects are organised in a hierarchical *N-ary tree* structure of nodes; except for the root node, each node has exactly one parent node, and can have multiple child nodes. Conceptually, the system differentiates between *countries*, *organisations*, *energy mixes*, *organisational units*, *objects*, *meters*, and *object zones*. The following paragraphs describe each entity and their most relevant attributes. Figure 4.4 shows a simplified representation of this data model from the community level downwards.

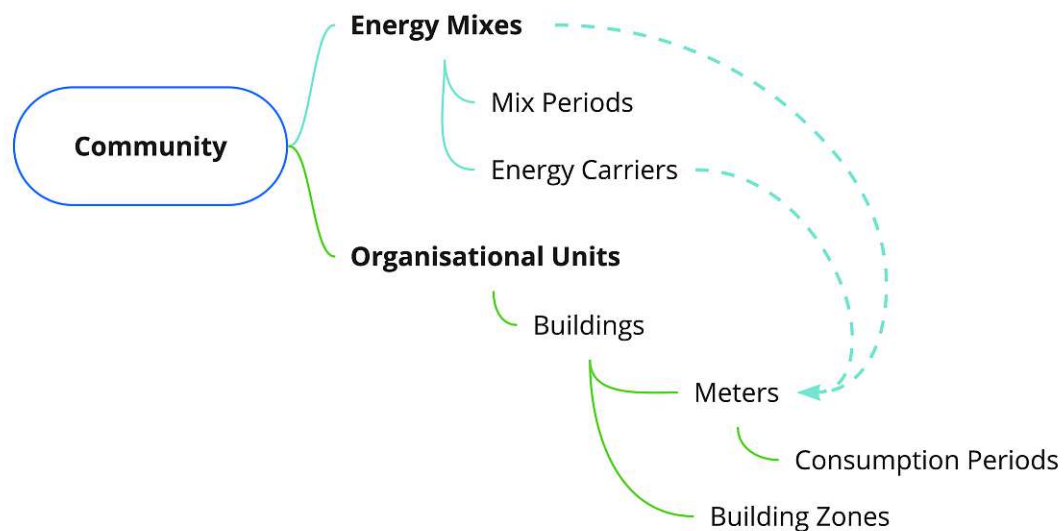


Figure 4.4: Simplified data model of a community in the EnerCoach energy accounting system.

Countries and organisations Starting with the children of the root node, country objects are the primary collection nodes for the only two countries currently represented in the system - Switzerland and the Ukraine.

Each country contains a list of organisations which represent either single communities or, in some cases, groups of geographically adjacent communities or communal regions sharing their energy accounting within the same node. Some other organisations types include property management firms tasked with energy accounting for the buildings they manage, church communities, geographic regions, political communities or even states²¹. Table 4.1 gives a quantitative overview of community types currently in use.

²¹While the system internals refer to these entities as ‘organisations’, most of the stakeholders colloquially

| Community Type | Number of occurrences |
|---------------------|-----------------------|
| Church Community | 2 |
| Community | 686 |
| Confederation | 2 |
| Other | 37 |
| Politic Community | 10 |
| Property Management | 5 |
| Region | 4 |
| State | 1 |
| Total | 747 |

Table 4.1: EnerCoach community types currently in use in Switzerland.

Organisation attributes beyond their type include *location data* (address, postal code), *contact information*, *climate-related data* (altitude, nearest weather station) as well as self-reported *basic demographics* (number of inhabitants).

Energy Mixes *Defined per community, energy mixes characterize the composition of energy used by the various community buildings in terms of their primary energy carriers.*

In energy accounting methodologies, the absolute energy consumption of a community or building (measured in **Kilowatt-hours** or less commonly, **Megajoules**) is never able to tell the complete story. The corresponding CO₂ emissions for a given energy consumption may vary significantly depending on the *energy carrier(s)* used. *Energy carriers*, in this sense, are primary sources of energy (e.g., coal, nuclear energy, solar energy, or wood chips), and are characterized by their *primary energy factor greenhouse gas emission factor, share of renewables, density, and heating value*, among others. These attributes allow various assessments of the climate impact the energy consumption may have: 100 kWh of electricity sourced solely from renewable sources such as wind energy has a significantly lower impact than 100 kWh generated in a coal plant, for instance.

Since the energy used by buildings for electricity and heating only rarely stems from a single source, though, the EnerCoach system supports the creation of so-called *energy mixes*, which specify, for a given period of time, the constituent energy carriers used to generate the energy used by a community. Each mix can be used by multiple *meters*, but each meter can only be assigned to one energy mix at a time. EnerCoach distinguishes between three types of mixes:

Electricity mixes *Electricity mixes are the most commonly used energy mixes in the EnerCoach tool and specify the energy carriers used to generate electricity.*

refer to them as ‘communities’. Going forward, both terms are used interchangeably in the following chapters depending on the context.

They may include more emission-friendly energy carriers such as water, wind and solar power, biomass energy plants, or geothermal power generation, waste incineration plants, as well as less sustainable power sources such as nuclear, oil and gas power plants. Electricity mixes can be used by *electricity meters*, as well as by a subset of the available *thermal production systems*.

District heating mixes *District heating* mixes specify the energy carriers used by district heating plants. These may include, for instance, gas or oil boilers, geothermal heat, various heat pumps, industrial heat recover and renewable sources such as water, wind and solar power. District heating mixes can only be used by *thermal production systems*.

Gas mixes *Gas* mixes only define the shares of natural and biogas for their mix periods.

Gas mixes can only be used by *thermal production systems*.

Data about energy mixes is provided by the energy supplier; for instance, electricity companies are obliged to provide a detailed list of the shares of the energy carriers used to generate the energy supplied in a given billing period. Energy mixes in EnerCoach can be defined for arbitrary periods of time with with a granularity of one day. As an illustrative example, Figure 4.5 shows the form interface for editing an electricity mix period in EnerCoach; the mix period in question describes the electricity mix for the period between 01/01/2020 until 12/31/2020 as 60% Water power, 30% Wind power and 10% nuclear energy, with an average, weighted primary energy factor of 1.53, a greenhouse gas emission factor of 5 g/MJ of energy (equal to 18 g/kWh) and a renewable share of 90%.

Organisational units *Organisational units cluster a community's objects into user-defined groups.*

Communities use this functionality to separate building groups, vehicle fleets or street lighting into their own organisational units. Most commonly, this functionality is used for larger communities, or if a given organisation in EnerCoach represents various districts or other financially independent and separately administrated parts of a community. Every community has at least one default organisational unit that contains all of its *objects*; larger communities have multiple organisational units grouping its buildings.

Objects *Objects refer to buildings or building complexes in the EnerCoach tool.*

Each object belongs to exactly one organisational unit as its parent node in the hierarchy. *Object* attributes, similar to organisation attributes, include location data (e.g., address, postal code, city) and contact data (responsible person, phone and email contact data), as well as optional climate-relevant data (altitude and, if different from its parent organisation, the nearest climate station). Finally, some historical and categorical data about an object is kept on file, including its date of construction and/or renovation, certification, Eidgenössischer Gebäudeidentifikator,

Edit Mix
×

Validity period

Electricity Labeling Owned facilities / Purchased certificates

| Energy carrier | Supply | Share | Primary energy factor | Greenhouse gases | Share of renewables | |
|--|--|--|-----------------------|------------------|---------------------|----------------------------------|
| Water power <input type="button" value="v"/> | | <input style="width: 50px;" type="text" value="60.000"/> % | 1.20 | 3 g/MJ | 100 % | <input type="button" value="x"/> |
| Wind power <input type="button" value="v"/> | | <input style="width: 50px;" type="text" value="30"/> % | 1.29 | 7 g/MJ | 100 % | <input type="button" value="x"/> |
| Nuclear power <input type="button" value="v"/> | | <input style="width: 50px;" type="text" value="10"/> % | 4.22 | 7 g/MJ | 0 % | <input type="button" value="x"/> |
| <input type="button" value="v"/> | | <input style="width: 50px;" type="text" value=""/> % | | g/MJ | % | |
| Energy mixture definition | <input style="width: 50px;" type="text" value="1000"/> MWh | 100.000 % | 1.53 | 5 g/MJ | 90 % | |

Figure 4.5: Example of the EnerCoach online form for an electricity mix period.

and main building category (i.e., the category of its first object zone). Objects are the parent nodes for both *meters* and *object zones*.

Object zones *Object zones characterize objects in terms of their use and utilization reference area in m².*

An *object zone* represents a part of a building or building complex. Each zone is only valid for a given period of time, or ‘service life’; zones that are currently in use have no end date for their service life set. Zones are defined by a set of attributes, including the *utilization reference area* or floor size of the zone in m², their purpose, building type and structural form, as well as the availability of infrastructure (e.g., light and appliances, electric ventilation, cooling, and use of warm water production).

Zones are further distinguished by their building *category* according to one of the 12 building categories as defined by the [Swiss Society of Engineers and Architects \(SIA\)](#)²² [314]. In addition, three non-SIA categories for *vehicles*, *public lighting* and *other* are available. Table 4.2 gives a quantitative overview of these categories and their frequency of use.

Each of these building categories is used for various calculations as part of the reporting system, particularly in regards to calculating expected target threshold energy and water consumption, as well as target and threshold CO₂ emissions. Combined with the other attributes of a given zone, these categories influence

²²Orig. “*Schweizerischer Ingerieur und Architektenverein (SIA)*”

| Zone Category | Number of occurrences |
|--------------------|-----------------------|
| Administration | 2500 |
| Apartment building | 2745 |
| Commercial | 247 |
| Family house | 680 |
| Hospitals | 249 |
| Industry | 782 |
| Meeting places | 2024 |
| Public baths | 182 |
| Restaurants | 559 |
| Schools | 5898 |
| Sports buildings | 1964 |
| Warehouses | 1426 |
| Vehicles | 235 |
| Public lighting | 441 |
| Other | 569 |
| Total | 20501 |

Table 4.2: EnerCoach zone categories in use.

said normative target consumptions and emissions to account for different building types and usage: for instance, it is to be expected that the energy and water consumption of a public bathhouse or pool would be magnitudes larger than for an apartment building, or that the heating costs for a school’s gym far exceed those of an industrial building. Each *object* can contain one or more *object zones*, and the sum of all zones should adequately represent the different use of a given building or building complex and allow an accurate calculation of expected energy consumption and greenhouse gas emissions for the building as a whole.

Meters *Meters are the primary point of data entry for consumption data of a given building.*

Each building can have one or more meters, which allow collecting energy consumption and costs for given *consumption periods*. The EnerCoach system distinguishes three types of meters:

Electricity meters Electricity meters track the electricity consumption of a building. Each electricity meter can either be assigned to one of the community’s electricity mixes, or directly use renewable energy carriers such as hydro, solar or wind power.

Through this option, it is possible to track not just energy consumption, but also energy production, in order to accurately model buildings that rely on producing their own electricity (e.g., through solar panels).

Thermal Production Systems Thermal Production System (TPS) are a type of meters tracking energy consumption for the purpose of creating heat or cooling.

A variety of different types of TPS can be defined in the EnerCoach tool, including water boilers, heat pumps, electric heating, thermal solar collectors, air conditioning systems as well as district heating systems. Depending on the type of TPS, the energy carrier(s) used can be qualified through either energy mixes or direct energy carriers. For instance, a district heating system requires a district heating mix to be defined on the community level; a heat pump may use either a gas mix, an electricity mix or environmental heat (as a direct energy carrier); and a central heating boiler could be using a gas mix, coal, various types of fuels, oil, or renewable energy carriers such as wood chips, logs or pellets.

Water meters Water meters track the water consumption for a given building. They are the only meters not assigned to an energy carrier, as they only track water consumption, but do not include information on energy costs related to the production, transport or disposal of water and wastewater.

As a noteworthy peculiarity of implementation, the meters for a given object are not correlated in any way with the object's zones in the EnerCoach system. This means that, for example, the energy consumption and costs tracked for a specific electricity meter cannot be attributed directly to a specific building zone, even if the physical electricity meter is located in a separate building zone (such as an electricity meter tracking only the energy consumption of a school's caretaker's residence). The reasons for this apparent disconnect are, as so often in complex software projects, due to the complexity of transforming import data from the previous iteration of the EnerCoach system into the new system.

Besides these core entities represented in this (simplified) data model, a few additional entities exist in the EnerCoach system, but are not part of the hierarchical tree of nodes described above:

Weather stations *Weather stations are used to model different geo-climatic conditions throughout Switzerland.*

They track mean temperatures for their region / location on a monthly basis as *climate data*, and are used to account for increased heating energy costs at higher altitudes or colder parts of Switzerland. The unique geological conditions in Switzerland make it necessary to calculate the average temperature difference between a constant indoor temperature and the outside, and apply an aliquot factor to energy consumptions of building dependent on where they are located. The climate data is updated on a monthly basis for all 41 weather stations known to the system.

User Data *Users of the system are authenticated against principals in the database, and assigned certain rights and roles dependent on their permissions.*

As the stakeholder analysis above illustrated, a variety of users with different access requirements to the system exist. Each user gets assigned one or more roles for certain nodes in the system, allowing them to edit or view data, manage general attributes of objects, meters, object zones and mixes, or generate reports. For security reasons, a more detailed description of these permissions is omitted at this point.

Constants *Various constant data entities are defined throughout the system.*

For instance, a total of 65 energy carriers used as part of the energy mix functionalities or to directly qualify the energy used by a given meter are implemented as constants, with each energy carrier defined by their primary energy factor, greenhouse gas emission factor, the share of renewable energy, density, heating value and unit.

Another set of constants declare the attributes of the 12 [SIA](#) building categories, including their various purposes, standard room temperature, average availability throughout a given year, as well as building-related factors such as intercept, gradient, building envelope factors and average consumption demands for water, heating and electricity.

These constants are updated very rarely in case policy changes necessitate an adaptation.

As a final note on the EnerCoach data model as presented above, it should be noted that the conceptual model as outlined here differs from the internal database model significantly, as the database middleware and web framework in use (SQLAlchemy and kotti) transform these fundamental models into a database scheme. For reasons of clarity, relevance, brevity, and operational security of the system, an entity-relationship level diagram and description of the database must be omitted.

4.3.2.2 Core Functionalities

Based on the data model as described in the previous Section [4.3.2.1](#), the EnerCoach system provides the following core functionalities (assuming the user in question has the requisite permissions):

Administrative Functionalities *Administrators can engage in various administrative tasks supported by the system.*

Administrators can invite new users to the system, manage their roles and permissions, create, edit and delete organisations, and perform caching-related operations (e.g., clear an organisation's cache and regenerate certain pre-calculated data.)

Community Management *Community managers can perform various [Create, Read, Update, Delete \(CRUD\)](#) operations in their assigned communities.*

The system provides functions to (1) edit community attributes; (2) create, update or delete organisational units, objects, zones and meters, and edit their attributes; (3) create, update and delete energy mixes and mix periods, and assign mixes or energy carriers to different meters; and (4) manage user roles and permissions within their assigned community.

Data Entry *Building managers can perform [CRUD](#) operations for their objects, zones and meters.*

These tasks include (1) adding consumption periods and logging energy consumption and generation as well as costs, (2) creating, updating and deleting object zones.

Reporting System *Various reports are available on both the organisational and building level.*

At the core of the EnerCoach energy accounting process resides the reporting system, which generates a number of graphs, aggregate table reports and custom visualisations based on the data provided for either an organisation as a whole or single buildings within that organisation. A more in-depth look at these functionalities can be found in the following Section [4.3.2.3](#).

Data Exports *Based on the report output, a number of exports can be triggered to download PDF or Excel files representing the reporting output.*

These include a rather large data export for an entire community, including consumption, mix and building data, as well as the relevant energy carrier constants, and some of the report graphics transformed into Microsoft Excel graphics. For the PDF outputs, an alternative to the *Energy Certificate Report* (see Figure [A.1](#)) produces a the building-specific plaque depicted in Figure [A.2](#).

The outline of functionalities provided above is neither particularly detailed nor exhaustive in the spirit of brevity and relevance. Where necessary for the subsequent analysis of transparency and accountability deficiencies, and the results of the PD workshop, these functionalities will be described in more detail as they are referenced.

4.3.2.3 Reporting System

An essential, core feature, and arguably the source of most complexity in the EnerCoach algorithmic system, is the reporting system. A total of 13 separate reports are available on the organisation level, of which 9 can also be generated for a single object or building. Each of these reports accepts a number of parameters that significantly shape the output generated for the user. The following list details the various reports available in the system, their parameters as well as significance for the energy accounting process.

Since the illustration of these descriptions with screenshots from the actual system are, on the one hand, highly relevant to the case study, but, on the other hand, also quite disruptive for the flow and comprehensibility of this section, the referenced figures can be found in Appendix [A.2](#).

Energy Certificate *Available on both organisation and object level.*

Based on the quantitative evaluation of usage data (see Section [3.2.2](#)), the most relevant and widely used report in the EnerCoach tool is the *energy certificate report* shown in Figure [A.1](#). This report is fundamentally normative in nature, as it directly evaluates the community's actual final and primary energy consumption, CO₂ emissions and water consumptions against their respective target values. As described in the previous Section [4.3.2.1](#), its calculations are based on object zone categories and their reference areas. The report delivers a rating of either the entire community or single objects on an 8-point scale from A to G, with A and B ratings denoting optimum and standard targets for sustainable performance, C through E representing adequate to sub-par performance, and F and G showing an immediate need for redevelopment of buildings and energy systems. These target values depend on numerous factors, including the composition and categories and building zones, their size, as well as climate data adjustments to reflect that energy expenditures for heating in cold winters is expected to be higher.

Parameters for this report include the reporting period in a 12-year sliding window, filtering by certain building categories or organisational units, and the option to either generate yearly or 3-year-average reporting periods.

Poster Report *Available only on the object level.*

A variation of the *energy certificate report*, the poster report shown in Figure [A.2](#) provides a PDF-exportable building plaque to visualize a single building's rating on the same normative scale from A to G.

Parameters for this report include the year it should be generated for, as well as the option to show a comparison year in the final output.

Evolution of Energy Key Figures *Available on both organisation and object level.*

The key figure report provides a relative look at the energy or water consumption per m² and year for a given community or building. It is available in three variations, depending on the meter type, and shows key figures for electricity, thermal/heat and water (see Figure [A.3](#)) consumptions respectively. Additionally, the report shows the number of objects taken into account, and—in the case of thermal/heat key figures report—the climate correction factor included in the calculation for each year.

Parameters for this report include the the reporting period in a 12-year sliding window, cropping the results based on the completeness of the available data, and filtering by certain building categories. For the latter case, the report also includes

horizontal bars illustrating the target and threshold values for the requisite key figures, as illustrated in Figure [A.4](#).

Evolution of Energy Consumption *Available on both organisation and object level.*

The basic energy consumption report shows a stacked area chart of absolute final energy consumptions grouped by energy carrier or energy carrier groups (see Figure [A.5](#)). Energy consumptions include both data from electricity meters and [TPS](#), and separates electricity consumptions based on mixes in renewable and non-renewable shares. Additionally, the report shows the number of objects taken into account.

Parameters for this report include the the reporting period in a 12-year sliding window, cropping the results based on the completeness of the available data, and filtering by certain building categories.

Evolution of Greenhouse Gas Emissions *Available on both organisation and object level.*

The basic greenhouse gas emissions report shows a stacked area chart of greenhouse gas emissions based on the emissions factors of the involved energy carrier or energy carrier groups (see Figure [A.6](#)). Greenhouse gas emissions include both data from electricity meters and [TPS](#), and separates emissions generated by electricity consumptions based on mixes in renewable and non-renewable shares. Additionally, the report shows the number of objects taken into account.

Parameters for this report include the the reporting period in a 12-year sliding window, cropping the results based on the completeness of the available data, and filtering by certain building categories.

Evolution of Energy Costs *Available on both organisation and object level.*

The basic energy costs report shows a stacked area chart of costs grouped by energy carrier or energy carrier groups, and—contrary to the consumption and greenhouse gas reports described above—includes costs for water use as well (see Figure [A.7](#)). Costs include both data from electricity meters and [TPS](#), and separates electricity costs based on mixes in renewable and non-renewable shares. Additionally, the report shows the number of objects taken into account.

Parameters for this report include the the reporting period in a 12-year sliding window, cropping the results based on the completeness of the available data, and filtering by certain building categories.

Energy Carrier Shares *Available on both organisation and object level.*

Similar to the evolution reports for consumption, greenhouse gas emissions and costs, this report visualizes the relative shares of various energy carriers or energy carrier groups for a given year in the form of three pie charts (see Figure [A.8](#)). Additionally, the report provides a data table with the absolute numbers visualized above.

Parameters for this report include an optional filter by building category, and the requested year to aggregate.

Comparison of Energy Indicators *Available only on the organisation level.*

On the organisation level, the evolution of energy indicators or key figures only shows a weighted average of key figures for the entire organisation or an organisational unit, depending on the parameters. To visualize the relative performance of the various buildings, the Comparison of Energy Indicators report lists those key figures separately for each of the buildings and a given year. Three variants for electricity, heat and water (see Figures [A.10](#), [A.11](#), and [A.12](#) respectively) exist.

Parameters include the year, type (electricity, heat or water), and optional building category filter.

Energy Indicator vs. Reference Areas *Available only on the organisation level.*

Since visualizing the performance of a given building based on its key figures alone often falls short in describing the building's overall impact on the performance of a whole community, the Energy Indicator vs. Reference Area Report provides a custom chart to juxtapose both data points for each building. Through a series of rectangular areas, with the key figures for heat and heat plus electricity charted on the y-axis, and the floor space (utilization reference area) charted on the x-axis, the report lists all buildings sorted by their key figure performance from worst to best. The resulting visualization is interactive by necessity, since bigger communities would produce unreadable charts due to their large number of buildings. A pager at the bottom allows the user to scroll through the objects, and a counter allows controlling the number of buildings to be shown on each page (see Figure [A.9](#)).

Parameters include the year and an optional building category filter.

EnergyCity Report: Renewable Energy *Available only on the organisation level.*

To address the requirements of the certification process mandated by the EnergyCity / [EEA](#) programs, the Renewable Energy report (see Figures [A.13](#) and [A.14](#)) provides an aggregated data table comparing the various renewable energy consumptions and the share of certified renewable energy thereof. The simplicity of the output is deceiving; this report is one of the most complex and multi-faceted reports in terms of its calculation algorithms, with multiple fallback calculation methods in case of missing data. For instance, the report was originally not intended to include actual consumption data entered at the meter level, but only consider absolute, organisation-wide consumption numbers based on energy supplier billing documents. To this end, the report data should be entered as absolute consumption numbers into the various energy mixes the community uses, and consumption data provided to meters is only used under certain conditions as a fallback. This report is also the only report visualizing a community's use of purchased energy certificates (to offset non-renewable energy consumption) and certified energy (as generated by renewable, local micro-energy generation systems such as hydro-power plants,

windmills and solar panels). Two variants of the report exist for electricity and heating respectively.

Parameters include the year and the type of report (electricity or heat) requested.

EnergyCity Report: Greenhouse Gas Intensity *Available only on the organisation level.*

Similar to the Renewable Energy report described above, the Greenhouse Gas Intensity report is part of the suite of reports targeting the EnergyCity / [EEA](#) certification process specifically. The report provides two aggregated data tables (for heat and electricity, respectively; see Figure [A.15](#)), listing various data points related to the greenhouse gas emission performance of the community grouped by the twelve [SIA](#) building categories. Data points include the number of objects per category, the sum of floor space/reference value, absolute and relative emissions, as well as the target and threshold values for emission per building category. The report also calculates the overall rating in percent based on the ratio of key figures and target values, and translates it directly into a point grade used in the EnergyCity certification process.

The only parameter for this report is the requested year.

EnergyCity Report: Energy Efficiency *Available only on the organisation level.*

The final report targeting the EnergyCity / [EEA](#) certification process is the Energy Efficiency report. Available in the three variants for electricity, heat and water (see Figures [A.16](#), [A.17](#) and [A.18](#) respectively), this report provides the same information as the Greenhouse Gas Intensity report for absolute and relative energy consumptions instead of greenhouse gas emissions, grouped by [SIA](#) building category. This report also calculates the overall rating in percent and points used in the EnergyCity certification process.

Parameters include the requested year and type of report.

Overview Report *Available on both organisation and object level.*

As a convenience, an overview report collects the previously described reports into a single, long dashboard, adding only marginal details about community attributes or building zones (depending on the level it is generated for). Since this report is, on average, roughly 10 pages long, but only contains the reports already shown in Appendix [A.2](#), a reproduction as part of this dissertation is omitted.

Excel Export *Available on both organisation and object level.*

While the Excel export is not an online, interactive report in the sense used above, the exporting functionality on both organisation/community and object/building level does qualify as reporting tool in a common sense. In a set of Microsoft Excel Worksheets, combined into one file, this report lists most of the information already provided in the reports above, but with more detail and absolute numbers. The Excel export function was introduced as a response to observations of multiple

users trying to manually validate report results by hand; having the requisite data available in spreadsheet form and being able to use Microsoft Excel formulas to estimate expected report results was a much requested feature for quite some time.

As noted before, this descriptive list of available reports is not meant to exhaustively describe the reports and their algorithmic implementation, as such a description would transcend the scope of this case study. The following analysis of the system's sources of complexity, as well as transparency and accountability requirements and deficiencies, will expand on this overview where necessary.

4.4 Transparency

The EnerCoach tool is a sophisticated algorithmic system of considerable complexity, and suffers from a distinct lack of transparency towards its various stakeholders. This initial realization as described through my own experience in the exploratory vignette in Section 4.1 was subsequently confirmed through the analysis of interview transcripts, email communication, and usage data and support requests. One energy consultant, upon learning the exact rules governing the inclusion/exclusion rules for buildings, meters and zones when calculating certain reports, provided this succinct summary:

“[...] and that is not written down anywhere? Not even energy consultants know that! That is the blackbox!”²³

Energy consultant “D”

To address these shortcomings from a CAS perspective, three primary questions needed to be answered first:

1. What are transparency requirements in the EnerCoach system?
2. What are distinct deficiencies of the system in terms of transparency?
3. What are the underlying reasons for these deficiencies?

The following sections address these questions in order.

4.4.1 Requirements

To answer the first question, multiple perspectives on transparency requirements can be considered, including *function-specific* and *stakeholder-specific* transparency requirements.

²³Orig. “[...] und das ist nirgends aufgeschrieben? Auch ein Berater weiss das nicht! Das ist die Blackbox!”

4.4.1.1 Function-specific Requirements

Starting with a *functional* perspective, the EnerCoach tool provides three main functionalities: *data entry*, *data aggregation and visualization*, and *auditing*, each of which come with some specific requirements towards transparency.

Data entry involves the collaborative and shared tasks of manually entering consumption data for electricity and water meters as well as **TPSSs**, keeping building zone and object properties up to date, and adjusting the energy mixes utilized by the organisation in question on a regular basis. Each of these tasks is a contribution, by one user, to the shared data pool of all users of a single organisation, and can significantly impact the results of the reports. Changes to the energy mix definitions, for instance, will impact all meters utilizing that mix and, subsequently, almost all reports across the board. Even a single change to an object zone may shift a whole organisation’s performance enough to affect, for instance, the energy ratings shown in the *energy certificate report*, or entirely hide its the target / threshold values, as one EC hotline staffer reports:

“Yesterday, we had an example where it didn’t show the target and threshold values for a building. And so I first validated: was the building zone entered correctly, meaning, are there inactive zones—that’s one of these things that triggers an error, I know that—and then [I found that] in multiple zones, those [usage factors]—100% heating, electricity and water—were not set to 100%. Although we always recommend that, right...[laughs] and that was set to 0% for one zone, for heating!”²⁴

EC Hotline staffer “E”

While most organisations are managed by less than three users, some larger communities have a collaborating team of 15 or more people working together, as Figure 4.6 shows.

For these larger communities in particular, keeping track of who made which changes and validating the data to avoid errors is no small challenge, and can be considered an instance of the *many hands problem* [136] as well. As an additional complication, most communities do not continuously work with the system, but will do so in annual or biannual cycles, leaving months in between interactions and further complicating the process of tracking changes to the community’s shared data pool. By and large, the requirements for transparency of the *data entry* functionalities thus pertain to *ex-post explainability* in the sense of being able to track where specific system inputs resulted in certain system outputs, but including a certain level of *system transparency* about

²⁴Orig. “Also gestern war ein Beispiel wo’s bei einem Gebäude die Ziel- und Grenzwerte nicht angezeigt hat. Und da hab’ ich zuerst überprüft: ist die Zone korrekt eingegeben, also, gibt’s inaktive Zonen — das ist meist auch so irgendwas was Fehler auslöst, das weiss ich — und dann eben war bei verschiedenen Zonen da auf der letzten Seite diese [Nutzungsgrade] — 100%-Wärme, Elektrizität und Wasser — waren nicht bei 100%. Wie wir das jeweils empfehlen, oder... [lacht] und das war bei nur einer Zone auf 0%, bei Wärme.”

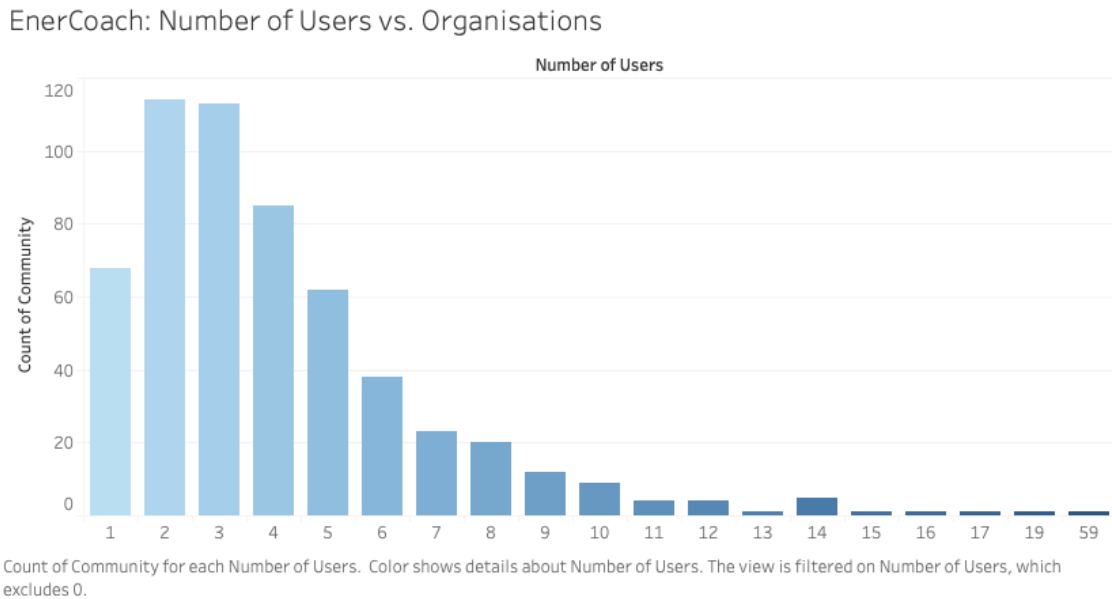


Figure 4.6: EnerCoach user counts vs. communities.

the interconnections between model entities such as mixes and meters, or building zones/categories and community performance.

Secondly, the *data aggregation and visualization* functionalities—incorporated in the reporting system—have requirements towards both *ex-post transparency* and *system* or *model transparency* as well. Regarding the former, interpreting the results of the reports is a complex sense-making process [315, 10] that involves tracing both user inputs (such as consumption data, mix definitions, or building data) and system constants (such as climate data or the energy carrier definitions) through the reporting algorithms to arrive at a certain output. Only a limited number of the various data sources involved in this calculation process are visible to the users; energy carrier data is only visible as part of the mix period definitions, and climate correction factors are only shown in specific reports as secondary axes (e.g., the *key figure report* for *thermal key figures* shown in Figure A.3). In both cases, users only see the end result of what are already quite complex calculations, and have no access to the underlying raw data itself. Similarly, the aggregate nature of most reports—particularly those generated for whole organisations, as opposed to single buildings—complicates tracing specific data points, such as the consumption data entered for one meter in one building out of many. In terms of the latter type of transparency requirements, a minimum of understanding for the modelling and interplay between different entities in the system is a prerequisite to achieve *system transparency* when it comes to the report. The knowledge that, for instance, building objects may be excluded from a report aggregation—on the condition that their data is incomplete—is an important piece of information when tracking down implausible or unusual report results. Similarly, a basic level of domain knowledge about energy

accounting methodologies, such as the definition of key figures as a relative measure of building performance relating energy consumption to floor space, is required to *make sense of* the outputs of these reports.

Finally, the *auditing* functionality requirements are largely focused on *ex-post transparency*, as the audit process itself is mainly one of validation and verification of report results. As energy auditors shepherd communities through the certification process for the EnergyCities / EEA programs, they must confirm the veracity of data entered, and the plausibility of report results, against various data sources located outside the EnerCoach system itself; this may include, for instance, an energy supplier’s billing statements or the actual building’s meters and layouts. Their job requires *system-level transparency* only insofar as they need to be able to confirm that the EnerCoach reporting system adheres to their own expert domain knowledge for energy account methodologies as prescribed by the certification process. In the interviews with energy consultants/auditors, it became clear that requirements towards *system transparency* may often be replaced by the intrinsic *trust* that the system would be working as expected. For energy consultant and auditor “D”, for instance, trusting the calculations was often enough:

“The energy carriers, the CO₂, I don’t need to retrace, I trust the factors.”²⁵

Energy consultant “D”

This may be due to the fact that the tool itself is affiliated with and recommended by the EnergyCity program, whose standards and calculation methods the EnerCoach tool complies to; while energy consultants like “D” mentioned instances where their trust in the correctness of the calculations was disproven due to earlier implementation errors, they still overall stated that they assume the calculations to be sound and regard these issues as “*growing pains*”²⁶ of the system.

This assessment notwithstanding, the complexity of the system combined with the fundamental changes the calculation methods underwent between the Excel version and the current online version still led to a significant number of software errors or bugs. By and large, two types of software errors frequently occurred: outright faulty server responses when trying to calculate a given report (e.g., the server responding with a HTTP 503: Internal Server Error), or more subtle errors leading to false report results (i.e., results that did not adhere to the formal specifications). The former type of errors play little role in questions of transparency, as both their occurrence and the responsibility for remediating actions are always clear: users see an error page instead of the expected report result, and are asked to contact either the EnerCoach hotline or WIENFLUSS to make a bug report; errors of this kind also get logged in the server backend and sent as emails directly to the developers to address quickly. Errors of the latter type, however, are much harder to pinpoint, as the reports may well show a result that would, at first glance,

²⁵ “Und die Energieträger, das CO₂, das muss ich nicht nachvollziehen, ich vertraue den Faktoren.”

²⁶ Orig. “Kinderkrankheiten”

appear correct or at worst implausible. Further inspection (e.g., as part of an audit) may yet reveal that an implausible result is, indeed, an incorrect one. Particularly during the initial years of deployment, e.g. between 2016 and 2018, such errors happened rather frequently and most often were the result of edge cases that were not explicitly described in the original specifications. Most of these errors affected only very specific combinations of object, meter, mix or zone definitions for specific reports and consumptions. For example, in late 2018, a user reported differences in results between two reports that should yield the same results. A closer investigation revealed a small error in one of the reports in a database query related to collecting relevant **TPS** for aggregation: The query did not account for **TPS** utilizing electricity mixes (e.g., a heat pump using electricity from a mix) under certain circumstances. Since this specific constellation of circumstances only occurred in fewer than 10 communities out of more than 600, the problem had gone unnoticed until then.

In terms of *ex-post transparency*, the existence of software errors such as the one described above meant many support requests directed at the EnerCoach hotline ultimately came with the underlying question of whether or not the implausible result may be caused by a subtle software error rather than by mistakes made by the user. Answering this question thus also directly pertains to accountability processes insofar as it reveals who the *actor* required to render the account would be: another user, an EnerCoach hotline staffer or, in the case of an actual software error, WIENFLUSS. Determining the difference between *implausible* and *incorrect*, however, usually had the highest requirements towards *ex-post transparency*, which often could only be met by direct access to the system's code and database on the one hand, and in-depth expert domain and technical knowledge on the other. Consequently, this determination often had to be made by WIENFLUSS rather than by the EnerCoach Hotline.

4.4.1.2 Stakeholder-specific Requirements

From the *stakeholder* perspective, requirements towards the transparency of the EnerCoach system vary significantly depending on the stakeholder groups as visualized in Figure 4.3. By and large, the level of transparency needed correlates with the stakeholder group's involvements with the day-to-day, practical use of the system for *ex-post transparency*, and the involvement with larger scope, policy-based considerations in terms of *system transparency*. User groups such as community users and building managers mostly require high levels of *ex-post transparency* to be able to gauge the impact their actions have on the system's outputs in the form of reports. For building managers, the scope of transparency includes their objects, meters, object zones and consumption data, but also includes energy mix definitions when managing the assignment of *energy mixes* to certain meters. A basic understanding of the various meter types and their relation to energy carriers or mixes as an example of *system transparency* is also required. Community managers need clarity on the same aspects, but on a larger scale corresponding to the larger number of entities in their purview.

Energy auditors and consultants also primarily require *ex-post explanations* to conduct

their analysis and verification tasks; their requirements towards *system transparency* are largely the same as the functional requirements for *auditing* tasks.

On the other side of the spectrum, the stakeholders less concerned with concrete, day-to-day tasks as users of the system have significant requirements towards *system transparency*. Both the EnerCoach Working Group members and WIENFLUSS are concerned with detailed, policy-driven implementation and adaptation tasks that require both an in-depth understanding of the system's architecture, calculation routines and model entities on the technical side, and fundamental domain knowledge about the energy accounting process and methodologies on the socio-technical side. Of these two stakeholder groups, *ex-post transparency* requirements apply to WIENFLUSS only, most pressing in order to provide competent tech support to the EnerCoach Hotline staffers and being able to differentiate between software errors and errors caused by the users (e.g., during data entry).

Finally, the EnerCoach Hotline staffers have the most balanced transparency requirements. As primary point of contact for the various end-users, they must be able to provide answers for specific, questionable system outputs (*ex-post explainability*), but also require significant insights into the system's architecture and calculation paradigms (*system-level transparency*) to be able to answer knowledge-based questions and run the training sessions for new users.

4.4.2 Deficiencies

Through the interview analysis, the deficiencies of the EnerCoach tool in terms of *ex-post explainability* and, to a lesser extent, *model transparency* came to light. These deficiencies affected either certain stakeholder groups or certain functionalities, or both.

By far the greatest number of complaints concerned issues of *ex-post explainability* of report results. As the system's implementation tried to adhere to both the fundamental paradigms laid out by the previous, Excel-based implementation, but also underwent significant transformations in its new, collaborative and online implementation, many users switching from the old to the new system were confronted with reports that looked familiar, but produced different results. The screen capture shown in Figure 4.7 illustrates just one of the many similar emails received by the EnerCoach hotline staffers, a smaller percentage of which also reached WIENFLUSS as official support requests.

One of the reasons for this behaviour was the result of a design choice made early in the implementation process: moving from fixed annual data entry to arbitrary time periods with the higher granularity of daily values. As became clear during the initial design process for the new tool, users of the old tool would have to go through sometimes complex, manual calculations to normalize data such as a consumptions or mix periods to a yearly value. Many communities did not receive energy bills from their suppliers that would reliably start on the same days and run for a year. Communities using multiple energy sources and suppliers, i.e., different suppliers for gas, district heating and electricity, might get one bill running from January 1st to December 31st, and

4. CASE STUDY: ENERCOACH

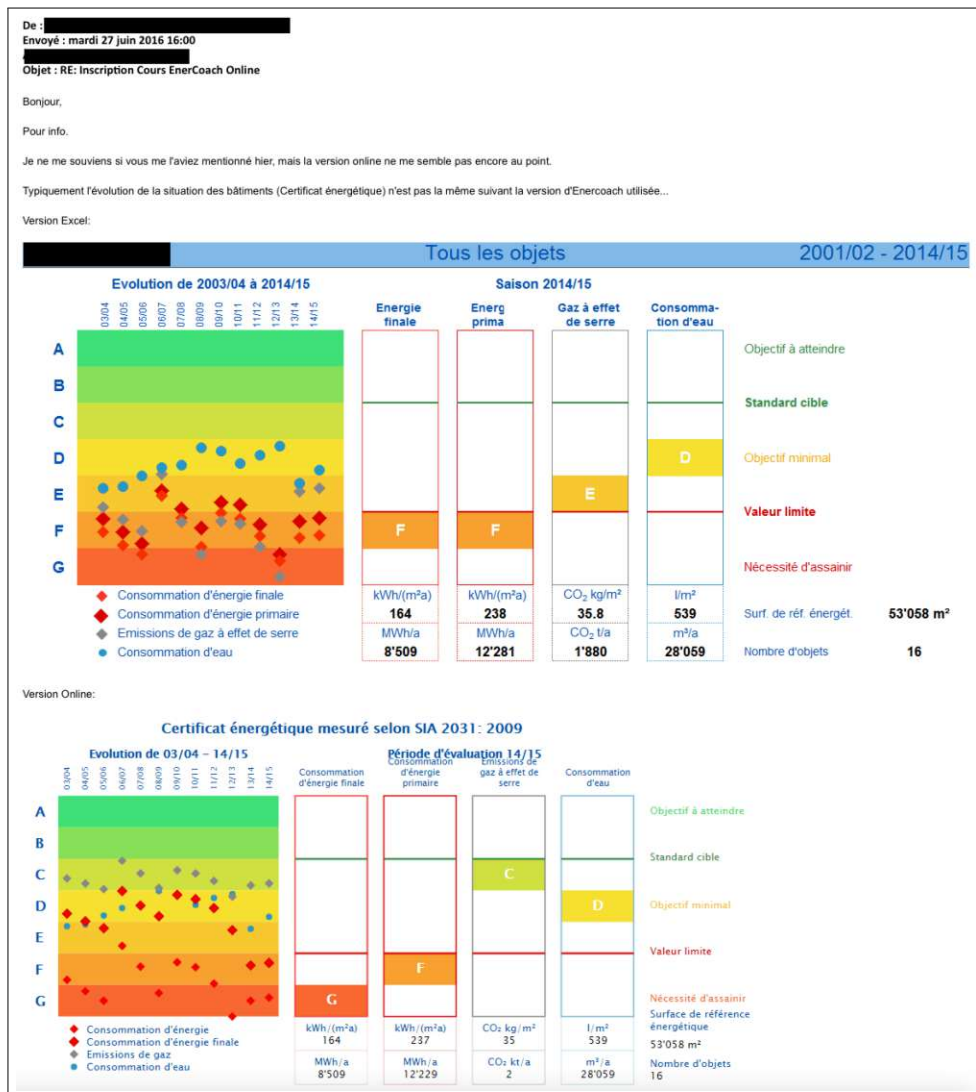


Figure 4.7: Screen capture of an early EnerCoach support email request juxtaposing the report graphics of new (online) and old (Excel) versions of the energy certificate report. The request in French roughly translates to: “Hello. For your information. I don’t remember if you mentioned it to me yesterday, but the online version does not seem to me to be quite ready yet. Typically, the evolution of the building situation (Energy Certificate) is not the same, depending on the version of Enercoach used...”

another running from the 3rd of May until the 2nd of May of the following year. To enter these data into the old system, they would have to split and divide the consumptions according to one pivotal date, which was a notoriously error-prone process. In a bid to address this shortcoming, the new EnerCoach tool was designed to be capable of handling arbitrary time periods by automating this normalization process across all

time-based model entities, including mix periods, consumption periods and the service life of building zones. While reducing the potential for human error, this automation also increased the opacity of the normalization process for end users, and resulted in slightly differing report results, thus giving rise to complaints about a lack of *ex-post explainability* of these reports. Stakeholders with higher requirements towards *ex-post explainability*—specifically, energy consultants and auditors—explained how they would purposefully not make use of this feature in order to retain the ability to verify basic calculations more easily:

“I’ll enter as inputs, as consumption, the yearly consumptions. And I’m making the reports for the whole year as well, so, January to January [...] and I mostly don’t use values from those quarterly billing statements, because I can’t trace those anymore. When I’m looking at the invoices for August to January and then January to the following year, then I don’t have a feeling for what, in the calendar year 2017, how many kilowatt-hours that are, because I can’t see it when I do the input. And I think that’s the chemistry, the secret behind these enormous calculations, that the reporting period has to be calculated as well.” ²⁷

Energy consultant “D”

What “D” describes is precisely the practice that the arbitrary time period feature introduced with the online version of EnerCoach was supposed to make obsolete. For “D”, however, the ability to quickly judge whether or not a given report result was immediately plausible clearly was more important than being able to avoid the tedious process of normalizing input data manually. In other words, this technical feature was made impractical for an important stakeholder group due to its lack of *ex-post transparency*, and they found themselves forced to circumvent the feature.

The example for a support request email shown in Figure 4.7 also perfectly illustrates the challenge of differentiating *implausible* from *incorrect* results, and the underlying issue of identifying the occurrence of subtle software errors. As described in the previous Section 4.4.1, the EnerCoach hotline staffers often struggle to definitively determine the cause for what a user may deem an implausible result, but must negotiate the field of tension between the user’s call for support on the one hand, and the potential cost of escalating the request to WIENFLUSS on the other. As numerous support emails forwarded to WIENFLUSS show, users routinely conflate *implausible* and *incorrect*: if their problem is difficult enough for them to contact the EnerCoach support hotline

²⁷Orig. “Ich gebe als Input, als Verbrauch gebe ich Jahresverbräuche ein. Und ich mache die Auswertungen auch nach dem ganzen Jahr, also Januar bis Januar [...] und ich mache keine Werte meistens von so Quartalsabrechnungen, weil dann kann ich das nicht mehr nachvollziehen. Wenn ich eine Abrechnung habe August bis Januar und dann Januar bis wieder zum nächsten Jahr, dann habe ich kein Gefühl, was jetzt im Kalenderjahr 2017, wieviele KWh das sind, weil bei der Eingabe ist es nicht ersichtlich. Und ich glaube das ist auch die Chemie, das Geheimnis hinter der Riesenrechnung, dass auch die Auswertungsperiode berechnet wird.”

already, they often have already tried different strategies to influence the result to no avail, and—as a consequence of diminished belief in *self-efficacy* [217]—feel a lack of agency to help themselves. The assumption of a software error as the underlying cause of their problem thus is often the only remaining explanation in their eyes. Thus, the system’s lack of tools to help both users and EnerCoach hotline staffers better assess the reason for such behaviour is another major transparency deficiency.

Another source of opacity lies in the mathematical necessities involved in the report calculation process. Certain fundamental calculations, such as the calculation of key figures, cannot be performed if the requisite base data is missing or incomplete, making it necessary to exclude the affected building from the report aggregations. To this end, the system features a complex data plausibility framework utilized before and during the report generation process to determine whether or not buildings, meters and zones fulfilled the necessary preconditions to be included in any given report. Complicating the issue were the different and somewhat inconsistent requirements for the various reports: for instance, the *evolution of consumptions*, *greenhouse gas emissions* and *costs reports* (see Figures A.5, A.6, and A.7 respectively) only require that at least one meter with valid consumption data within a 12-year time frame exists in a given community. By contrast, the *energy certificate report* requires complete consumption data, building zone definitions, climate data and so forth, for the requested time period. This discrepancy meant that it became very difficult for users to determine why, based on the data they entered, one report might show the aggregate results for all 15 buildings, while another only included 12 buildings for the same time period. Although the results of this plausibilisation and exclusion process were provided to the users in the form of the ‘missing data’ table illustrated in Figure 4.8, most interviewees reported not realizing the impact this missing data had on the report in question, with one energy consultant (“G”), after suggesting to implement such a feature, stating she “*had never seen this [list] before...*”²⁸.




| Evolution of energy key figures: Electricity - Missing Data | |
|---|--|
| Missing value | Location |
| Consumption Data for all electricity meters | Demostadt > 11 Sozialamt der Region |
| Consumption Data for electricity meter | Demostadt > 07 Schulhaus Vogelsang > 698 484 Appartement  |
| Consumption Data for electricity meter | Demostadt > 10 Hauswirtschaftsschule > 390 820 Lokal + Heizung  |
| Consumption Data for electricity meter | Demostadt > 13 Friedhofgebäude > 439 858 Atelier-Garagen  |

Figure 4.8: Example of a missing data table displayed underneath each report listing missing data and reference location.

In terms of *system transparency*, the transparency of model entities and their core relationships was one of the most commonly mentioned sources of confusion. As one energy consultant explicated during their interview, the sheer number of factors that influence the calculation of the target and threshold values for energy consumptions of

²⁸Orig. “Also, die [Liste] habe ich noch nie gesehen...”

buildings (based on the building zones and their categories) was particularly confusing, essentially requiring even domain experts like energy consultants and auditors to trust the system, instead of being able to verify its results.

*“If I have two reports, one for a school building with a water heater and one without, and different situations — the target values are different, one is 17 and the other 18, [...] then I can not give an immediate answer, I have to dig deep into the SIA norms. [...] Maybe it has to do with the zones, maybe with the consumptions, or with how compact the building is, or the building envelope number, through the SIA calculations, which is pretty difficult.”*²⁹

Energy auditor “D”

As these target values are dependent on both user-entered data (which, at least, could be looked up manually) and system constants (which are not visible to normal users), verifying the plausibility of these values was particularly challenging. The fact that the calculation methods of these target and threshold values could also vary for different report types completely eluded all but one interviewee.

A similar issue nested deeply in the technical aspects of the implementation turned out to be the conceptual gap between the mathematical specifications on the one side, and their translation into viable algorithmic processes on the other. For instance, the order in which certain aggregations are performed makes no difference in the mathematical representation of a given report, but can change the final result of the report after its manifestation as code. An illustrative example of this is the key figure report: in the system’s implementation, *consumptions* of all meters of all included objects and *reference areas* of all zones are summed up first, and then divided by each other for technical reasons of system performance. In a mathematical sense, this procedure would be equivalent to calculating each object’s key figures first and then calculating a weighted average of all objects; in the concrete implementation, however, the former approach meant that objects with gaps in the time series of reference area could still be included in some reports, since all reference areas were aggregated separately at first, resulting in different results for some specific edge cases. These implementation details may well be highly relevant information for EnerCoach Hotline staffers, consultants and auditors to consider when trying to troubleshoot unexpected report results. At the same time, hotline staffers and energy consultants agreed in their assessment that distinguishing between these two approaches most likely would transcend the levels of algorithmic and technical literacy of most end users. Consequently, any potential measures addressing this kind of opacity in

²⁹Orig. *“Wenn ich dann 2 Reports habe, einmal ein Schulhaus mit Boiler, einmal ein Schulhaus ohne Boiler, und verschiedene Situationen - die Zielwerte sind unterschiedlich, einmal 17, einmal 18, [...] dann kann ich nicht sofort eine Antwort geben darauf, da muss ich ziemlich tief in die SIA-Norm gehen. [...] Vielleicht hat es mit den Zonen zu tun, vielleicht hat es auch mit den Verbrauchern zu tun, oder mit der Kompaktheit der Gebäude, mit der Gebäudehülle, über die SIA-Berechnung, und das ist dann schwierig.”*

the system would most likely need to be designed for the target group of expert users (such as the hotline staffers themselves) to elucidate these details.

An interesting observation in regards to these deficiencies concerns the existing strategies to counteract them, and their effectiveness in addressing these issues. Similar to the example given above of energy consultants manually aligning reporting periods and data entry to improve their intuitive ability to assess the plausibility of report results, other challenges also relied on finding solutions outside the system rather than requesting changes to the system itself. For instance, as outlined in the transparency requirements Section 4.4.1 above, certain tasks performed by auditors or energy consultants require a certain level of *traceability* of user actions, or, in other words, figuring out who was responsible for a given change in a community's data pool. Given the larger concern of algorithmic accountability, this requirement pertains directly to the accountability process between *actor* (the users performing data entry) and *forum* (the auditors or consultants). The system itself offers little to no functionalities to support such a process, beyond the list of user accounts having access to a given community or building, and yet this specific type of transparency—most obviously related to *accountability*—did not feature prominently as problematic in the interviews. Pressed upon their strategies for tracing these changes, the reasons given pointed to established processes in the *social*, rather than the *socio-technical* or *technical* realm of the system's assemblage: auditors, consultants or hotline staffers would use their tacit knowledge of contact persons and their responsibilities to track down the culprits via phone or email, rather than trying to find a function in the system that would allow them to do so on their own. This observation quickly became apparent as a larger pattern; as a general strategy, many of the system's shortcomings in terms of transparency were addressed by social measures (to various extents of success), rather than through requests for additional technical features. For example, most aspects of *system transparency* were supposed to be addressed by (1) the training sessions offered by the hotline, (2) the availability of the EnerCoach Hotline itself, (3) the (very) rudimentary user documentation in the form of a series of PDFs in lieu of a more complete user handbook, and (4) through the auditing and energy consulting processes embedded in the context of application, as energy consultants would work to explain the system's concepts to their point persons in the communities, and auditors would find inconsistencies as a type of human corrective. While the importance of these measures was largely agreed upon by the interviewees, they also noted that they were not satisfactory in resolving the opacities of the EnerCoach system, and rather regarded as a clutch in lieu of better solutions.

4.4.3 Underlying Reasons

Building on the in-depth analysis of transparency requirements and deficiencies given in the previous sections, the reasons underlying these issues can be distilled. Identifying not just *what* transparency requirements the system should fulfil and to which extent it already does, but also *why* it does or does not is an important prerequisite to addressing the issues at hand.

Considering the tentative taxonomy of *transparency challenges* illustrated in Section 2.3.2, Figure 2.2, the EnerCoach system would be categorized as an *interpretable system*. Neither are inherently opaque machine learning or AI methodologies used in the system, nor are the data models themselves opaque in and of themselves. The *curse of dimensionality* does not apply in the strict sense of the term either: while the aggregation and calculation procedures can be overwhelming due to the large number of objects, meters, and mixes, the interacting variables do not increase at the same rate as the problem space for machine learning applications.

Considering EnerCoach as an interpretable system, one might consider *intentional opacity* as an underlying reason for certain transparency deficiencies experienced by some stakeholders, first and foremost community users and those working with the system directly. In fact, this suspicion featured prominently in the interviews with members of the EnerCoach Working Group, as withholding certain internal processes as a matter of policy might well be an intentional measure to discourage attempts to game the system. However, this suspicion was strongly rebuked by all interview partners: Neither did the EnerCoach Working Group ever consider actively or intentionally hiding or obfuscating the internal processes of the system, nor did they see attempts to cheat the system as a realistic or common enough problem warranting such secrecy. Asked specifically, what areas or aspects of the system should not be available or explained in detail to the users, EnerCoach working group member “B” describes their stance as follows:

*“Well, I don’t see anything right now. I would, at most, consider the topics where there might be attempts to influence the input data in relation to those results that are relevant for the point assignment in the EnergyCity process, but by and large those can’t be faked as easily either.”*³⁰

EC Working Group member “B”

While they conceded that cases of specific users or certain communities trying to manipulate their data (in order to reach better ratings for their certification processes) did occur, they also described these attempts as naive and trivially identifiable as such by auditors, consultants or the hotline:

*“In my experience: almost all of those that are trying to artificially improve their energy key figures, are doing it the other way around. They don’t understand that it is a simple division, and input smaller utilization reference areas or a lower utilization factor, and this yields the opposite effect.”*³¹

³⁰Orig. “Also ich sehe im Moment nichts. Ich kann mir höchstens einen Themenbereich vorstellen wenn es um die Beeinflussung der Daten geht in Bezug auf diese Resultate die dann punkterelevant sind für den Energiestadt Prozess, aber in der Regel kann man die auch nicht so einfach faken.”

³¹Orig. “Also meiner Erfahrung nach: fast alle die die Energiekennzahl versuchen künstlich zu verbessern, machen das umgekehrt. Sie verstehen nicht, dass es eine einfache Division ist und schreiben dann weniger Energiebezugsfläche rein oder einen niedrigeren Nutzungsgrad, und das hat dann die umgekehrte Wirkung.”

As they described it, these attempts also show a fundamental misunderstanding of the relation between *energy consumptions* and *utilized floor space*, with users artificially *underestimating* the floor space of buildings in their attempt to improve the key figures, and only achieving the reciprocal effect of *increasing* the calculated energy consumption per m² instead.

Consequentially, all interview partners agreed that the primary source of opacity in the system lies in its complexity, and the relation between the underlying energy accounting policy and methodologies on the one hand, and the concrete implementation as algorithmic code on the other. While the general idea behind energy accounting is deceptively simple—keeping track of energy consumption, greenhouse gas emissions, and estimating the relative performance of energy-consuming entities in a community—the aggregations and calculations involved in the EnerCoach system manifest an entangled mix of *descriptive* and *normative* aspects of the underlying policies that, in the words of Bruno Latour [80], represent technology as *society made durable*³². To disentangle these two aspects, and to understand the internal processes of the EnerCoach system as a specific implementation of the guiding principles and policies defined by the EnerCoach Working Group, thus became a challenging and sometimes impossible task for many of the system’s stakeholders—including, at times, the EnerCoach working group members themselves.

On a technical level, the aforementioned granularity of possible data entry on a daily basis for energy consumptions, mix periods and object zone service life, as well as the interdependencies between the various data entities grossly exacerbate the system’s complexity. In software engineering terms, the slow, piece by piece availability of report specifications during the system’s development process also resulted in *high coupling* and *low cohesion* [316], making these issues equally challenging for both WIENFLUSS, the EnerCoach Working Group members and hotline staffers, and particularly difficult to grasp for the end users.

From a socio-technical perspective, the heterogeneous nature of the various stakeholder groups in terms of both their algorithmic and domain literacy is a major complicating factor in the transparency of the system. As the analysis of requirements and deficiencies above reveals quite clearly, the needs and problems faced by energy consultants, auditors, community managers, and hotline staffers differ greatly, and require nuanced and targeted measures to address. Any attempts to improve the system for all stakeholders at the same time, and with the same measures, are prone to exemplify the saying “*one size fails all*”. The current EnerCoach system and its deficiencies in terms of transparency thus are also the result of the difficult trade-offs between attempting to accommodate a large

³²This *durability*, in particular, becomes particularly apparent when considering how slowly the transition between the old Excel and the new, online, version of EnerCoach was progressing, and how difficult it proved to introduce policy changes to the system without prompting adverse reactions by the various stakeholder groups.

variety of community and building configurations as accurately as possible on the one hand, and creating an energy accounting system that could still be used effectively by these different stakeholder groups on the other.

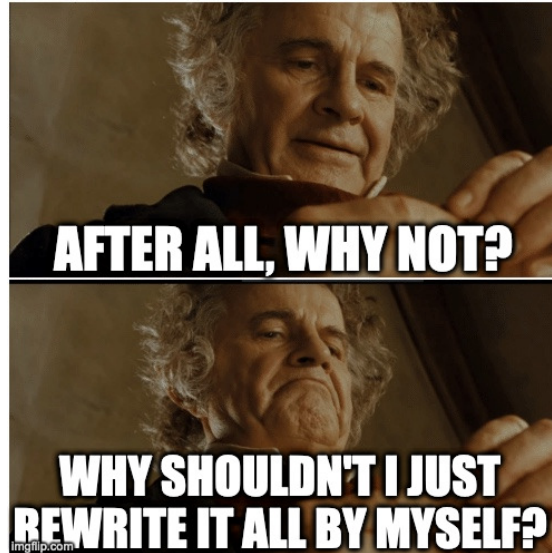


Figure 4.9: Illustration of a typical software engineer tempted by notions of grandeur and overconfidence. *Image courtesy of imgflip.com.*

Faced with these results, software engineers might be tempted to conclude the only plausible remedy would be to scrap the system, and attempt a complete redesign from scratch: after all, now that the technical and usability issues related to transparency are *known*, why *shouldn't* we be able to avoid the mistakes that lead to this situation the next time around (see Figure 4.9 for an illustration of this thought process). This notion, while surprisingly common and likely familiar to any programmer of complex, large-scale systems, can quickly be dispelled as infeasible and, indeed, short-sighted in light of the nature of these systems. For the EnerCoach example in particular, the monetary budget needed to undertake such a complete redesign would have significantly transcended the available financial resources of the involved actors.

Beyond the obvious issues of financing, however, the idea that a new implementation would be able to avoid the issues of transparency *a priori* is misleading in and of itself, as many of the challenges related to transparency (and, by extension, accountability as well) are the result of the evolving, *ontogenetic* [87] nature of algorithmic systems: to assume that a new system would not be prone to the same issues as the old one after a multi-year reimplementation process is folly. Finally, addressing issues of transparency from a purely techno-deterministic standpoint, i.e., assuming that technical solutions alone can address complex, socio-technical problems, has already led to various cautionary tales and case studies in CAS, as explicated in Chapter 2.

Taking a different approach, the *interventionist* part of this case study was explicitly

designed to avoid the lure of such simplistic solutions, and attempted to address these issues through the participatory approach described in the following sections.

4.5 Participatory Design Workshop

Following the analytic part of the case study, a one-day *participatory design workshop* was held in Aarau, Switzerland with five participants, led by myself. The workshop consisted of two phases: (1) a collaborative exercise in algorithmic visualization, and (2) a brainstorming session utilizing design thinking tools such as mock-ups and paper prototypes to derive concrete measures to address issues of transparency and explainability. The methodological design and considerations are described in more detail in Section 3.2.3. The following sections describe the workshop process, results and findings in details.

4.5.1 Part 1: Collaborative Exercise

As the system analysis identified the reports in general, and the *energy certificate report* (see Figure A.1) in particular as a source of much confusion and opacity for many users, the collaborative exercise in Phase 1 of the workshop was aimed at (1) taking stock of the level of shared understanding of the five participants in regards to this report and (2) creating a visual aid to illustrate the algorithmic processes involved in the calculation of this report. In other words, this portion of the workshop was focused on questions of *system-level* and *model-level transparency*. To this end, the participants were chosen to cover both domain experts in energy accounting and the EnerCoach system itself, as well as in-depth knowledge of end-user needs and complaints when using the system. Additionally, four out of five participants had (at least some) prior experience as energy consultants/auditors, and they all had participated as interviewees prior to the workshop as well.

After an introduction of the general goals of the workshop, participants were given instructions to discuss, explain and visualize an abstract representation of the algorithmic process of calculating the *energy certificate report* to the best of their knowledge and understanding, until a consensus about the accuracy of the result was reached. The energy certificate report was chosen specifically for its complexity and universality. First and foremost, generating this report involves the most complex combinations of various data sources and comes with the highest requirements for data integrity and completeness out of all available reports. Secondly, the specific calculations involved are universally relevant throughout the system insofar as they, in their modular forms, feature in various other reports as well.

To guide their process and support them in this task, all relevant model entities known to the EnerCoach system (see Section 4.3.2.1) were provided in the form of *entity cards*. The creation of these cards was prompted by a suggestion made by one of the prior interview participants that, in order to explain the system entities' interrelations, a

common visual language would be helpful to recognize commonalities between reports, terminologies and concepts. Figure 4.10 shows a subset of these cards, which were given to the participants without further explanation of their specific meaning, leaving them to interpret their significance among themselves. For each of these entities, they received multiple cards, allowing them to create alternative visualizations side-by-side or to visualize their contributions to the report in multiple, parallel ways. For this exercise, I remained a silent observer, with the exception of answering questions about the exercise process and instructions itself.

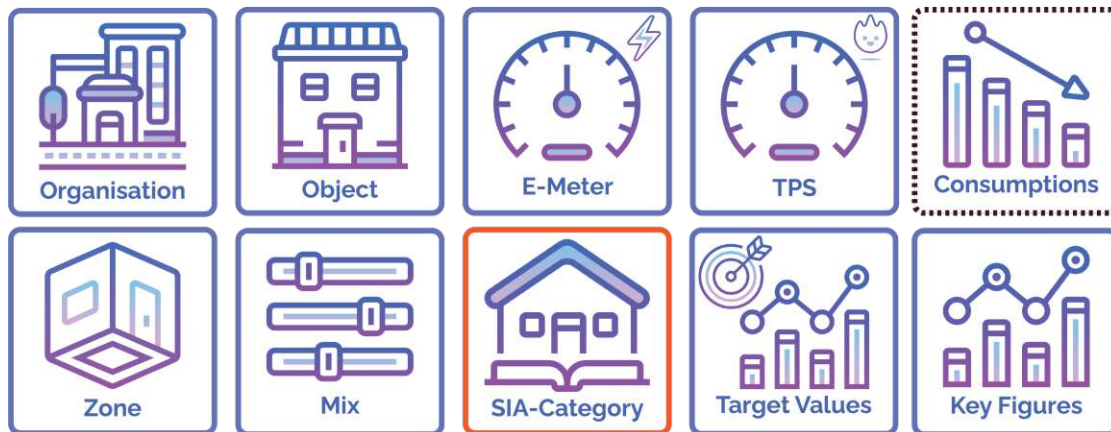


Figure 4.10: Selection of EnerCoach entity cards used in the participatory design workshop.

With the help of these entity cards, pens and flip-chart paper, the group was asked to visualize the calculation process in any way they deemed accurate, and narrate the calculation process with the help of the resulting visualization at the end.

First, the group engaged in a lively discussion centred around the meaning of the various entity cards in relation to where and how the underlying data was entered in the EnerCoach system. While the group reached a consensual understanding of these cards very quickly, the relation of the represented entities to each other, and the dependencies between them, were not immediately clear and led to some confusion. For instance, one exchange related to the entity card *target value* revealed that energy consultant “G” had assumed the calculation of target values were dependent on *energy consumptions*, and thus, *meters*; EC Working Group member “A” clarified that this was not the case, and that target values depended solely on the *object zone* definitions and their related constants as defined by the [SIA](#) norms.

Similarly, the participants hotly debated how and when climate correction factors were included in the calculation processes—a point of confusion not just for them, but for the end users as well, as “G” explicated:

*“What is interesting and what we have to explain constantly, is that [...] people see the total consumption values for heating, for instance, and they enter what they have on their energy bills or on their meters, and they get a number. But when you look at the evolution of consumptions, that number is already factoring in the heating days, so it’s [climate] corrected, and then there are people that use those numbers [...] and then I get the phone calls: ‘Why is this different?’ [...] I mean, I can explain everything, but we have to explain it, too, and even we took a while to figure out that these values are corrected for the heating days already, [...] which is correct, of course, but it is not identical to the number that you entered before.”*³³

Energy consultant “G”

As the group struggled to combine these realizations into a coherent visualization, energy consultant “F” offered a different approach starting with the energy consumptions, and suggested building the report from there based on the dependent entities required to calculate the next step. Incidentally, this strategy resembles that of a ‘greedy’ heuristic [317], i.e., always selecting the locally optimal choice at each stage of the process. The group quickly adapted his approach, and collaboratively assembled a type of process flowchart similar to those reported on by Hundhausen et al. [318] in their study of the effectiveness of various algorithm visualizations for computer science students. Figure 4.11 shows some impressions of this cooperative process in action, as well as the resulting flowchart.



Figure 4.11: Impressions of the PD workshop visualization exercise.

³³Orig. “Also, was ja spannend ist und was wir immer erklären müssen, ist dass [...] die Leute die Verbrauchswerte im Total für die Wärme sehen, zum Beispiel, und hier füllst Du ab was Du in den Rechnungen oder auf den Zählerwerten hast, dann hast Du eine Zahl. Und wenn Du dann nachher schaust bei der Entwicklung, ist diese Zahl bereits hochgerechnet mit den Heizgradtagen, also korrigiert, und dann gibt es Leute, die diese Zahl verwenden [...] und dann hab’ ich sie am Telefon: ‘Wieso ist das anders?’ [...] Also, kann man alles erklären, aber das muss auch erklären, und wir haben auch eine kurze Zeit gebraucht bis wir es hatten, dass das da Heizgradtag-korrigiert ist, [...] was auch richtig ist, aber es ist nicht identisch mit der Zahl, die Du vorher eingegeben hast!”

The remaining discussions during this collaborative process provided some additional, valuable insights into the discrepancies between the participant's abstracted understanding of the calculation process on the one hand, and the concrete implementation on the other, highlighting particularly opaque aspects of these calculations and the system in general.

One of the most controversially discussed aspects pertained to the *origins* of the data represented by the entity cards, putting the spotlight on the difference between whether certain data would be provided by the users or the system itself, and at which point in time (e.g., the difference between system constants such as energy carrier definitions or climate data on the one hand, and user data such as consumption data or mix definitions on the other). In the end, a consensus was reached on a taxonomy of data entities, categorizing them as either (1) *system-wide data points* unaffected by user input, (2) *static data points* rarely changed by users (such as community managers or building managers), and (3) *dynamic data points* changed or updated at least annually (such as mix periods or consumption periods). Being able to differentiate clearly between these three types of data was shown to be very helpful to the participants, particularly in terms of validating implausible report results. Since the *dynamic data points* were the most likely culprit for such questionable outputs, they would be the first target for revalidation, often explaining the result and making further investigation of the more difficult to correlate *static* and *system-wide data points* unnecessary. Following this insight and upon request by the participants, the entity cards were adapted after the workshop to more clearly depict this taxonomy, with a blue border frame depicting static data points, dashed blue frame denoting dynamic data points, and orange frames showing system-wide data points respectively. The selection of entity cards shown in Figure [4.10](#) already incorporates this adaptation.

Another insight gleaned from the discussion was the issue of sequential vs. parallel processing, as mentioned briefly in Section [4.4.2](#). The group visualization showed a parallel processing approach to calculating key figures and target consumptions, i.e., aggregating data per building and averaging them for the resulting report output. The actual implementation follows a sequential approach for each of the constituent entities of this calculation, aggregating consumptions and floor space separately at first, and calculating the final key figure the the community in question overall at the end. As described before, mathematically, this should lead to equivalent results; however, as this process involves multiple steps of plausibilisation for missing data on the various *community*, *object*, *meter* and *object zone* levels, the results in the technical implementation can be quite different, and thus a source of errors difficult to track down or trace.

Finally, as the group was trying to decipher the specific ways in which target values for energy and water consumptions are calculated depending on, among others, *object zone* attributes describing the building, the *utilization reference area* of these zones, and *climate data*, energy consultant “F” made this observation:

*“It is kind of funny, if I may say so, that we, in this group, do not seem to be 100% clear on how the target values are being calculated, and so it is obviously equally difficult for a user to say: am I performing well [with my buildings] or not?”*³⁴

Energy consultant “F”

This provocatively formulated statement spawned a longer discussion on the reasons for why the calculation of target values was so complicated. This exchange revealed the various trade-offs made to accommodate communities with a large number of old buildings in need of redevelopment or renovation: On the one hand, target values for energy consumption and CO₂-expenditure are based on accepted SIA standards for new buildings; on the other, judging old or historic buildings by the same standard would lead to significantly worse ratings for those communities still relying on older or historic buildings. As the EC Working Group members with knowledge of the system’s history explained, the normative aspects of the EnerCoach system—energy certificate ratings or target values, for instance—grew into an exceedingly byzantine set of formulas, factors and exceptions that exacerbated the system’s transparency deficiencies further simply as a consequence of trying to accommodate everyone. Not all decisions necessarily made sense scientifically or reflected the state of the art in sustainable building technologies or sustainable energy practices, but instead followed an overarching political agenda. Providing communities a gradual path forward to improve their situation regardless of where they started from was seen as more important than a factual comparison of community buildings to the state of the art.

Similarly, the most fundamental constants defined in the system, the definitions of energy carriers and their primary energy factors, which influence almost all reports and calculations in the system, were initially considered as descriptive, objective constants by some of the participants. This notion, however, was quickly dispelled by EC Working Group member “C”, who was performed the initial collection of these energy carriers

³⁴Orig. *“Ist ja eigentlich ganz lustig, wenn ich nochmal kurz dazu sagen darf, dass wir innerhalb der Gruppe anscheinend was den Zielwert angeht auch nicht 100% immer im Klaren sind, welches ist wie berechnet, und damit ist es natürlich auch schwierig für den Benutzer, danach zu sagen, bin ich jetzt eigentlich ‘gut’ unterwegs [mit meinen Gebäuden] oder nicht?”*

and their assigned factors for the original Excel version of EnerCoach:

“I had to consider all this, back when [EnerCoach] was all new. I also searched for all the energy carriers, but there are people working on this, expert groups, that publish about this. I had to do a lot of work for this, and looked up all those energy carriers [...] I consulted the experts multiple times, too, [and asked] ‘Why is it like this?’ The answer is simply: those are political values, purely political!” ³⁵

EC Working Group member “C”

Just as for the calculation of target values, her explanation showed that the exact factors describing the energy carriers are the result of a complex interplay between scientific expertise and energy policy decisions. Both cases illustrated how explanations aimed at improving *system-level transparency* may need to satisfy not just the question of *how*, but also of *why* a given behaviour was implemented, as purely descriptive explanations may not cover the intentions behind the decisions that fundamentally shaped the final outcome.

4.5.1.1 Baseline Comparison

Once the group was satisfied with their result, and had reached an agreement that their visualization was as accurate as possible given the constraints imposed by the medium (i.e., the abstract representation of internally complex entities), the participants were asked to compare their result with a visualization prepared by me and based on the actual code-level implementation of the *energy certificate report*. For this visualization, I had followed the same constraints in regards to medium and abstraction, only using the same model entity cards and visualizing their connections and interdependencies. Figure [4.12](#) shows this first version created by me.

This discussion focused on the different perspectives and interpretations of the same process, and the learnings that could be derived for the usefulness of such visualizations for the various stakeholders. A core insight emerging from this juxtaposition was to highlight the importance of finding a delicate balance between technical accuracy on the one hand, and comprehensibility through abstraction on the other: *system-* and *model-level transparency* for the EnerCoach system does not necessarily need to be absolutely technically accurate in order to facilitate understanding and sense-making processes. In fact, as one participant (“F”) noted, the technical specifics may complicate the big picture of the algorithmic process to the point where it masks the important

³⁵Orig. *“Diese Überlegungen hab’ ich mir stellen müssen, damals als das ganze [EnerCoach] neu war. Ich hab auch alle Energieträger gesucht, aber es sind eben Leute dahinter, Kompetenzgruppen, die daran publizieren. Ich hatte da eine Riesenarbeit und habe die Energieträger herausgesucht [...] Ich hab mehrmals die Experten konsultiert, [und gefragt] warum ist das so? Die Antwort ist eben: das sind rein politische Werte, rein politisch!”*

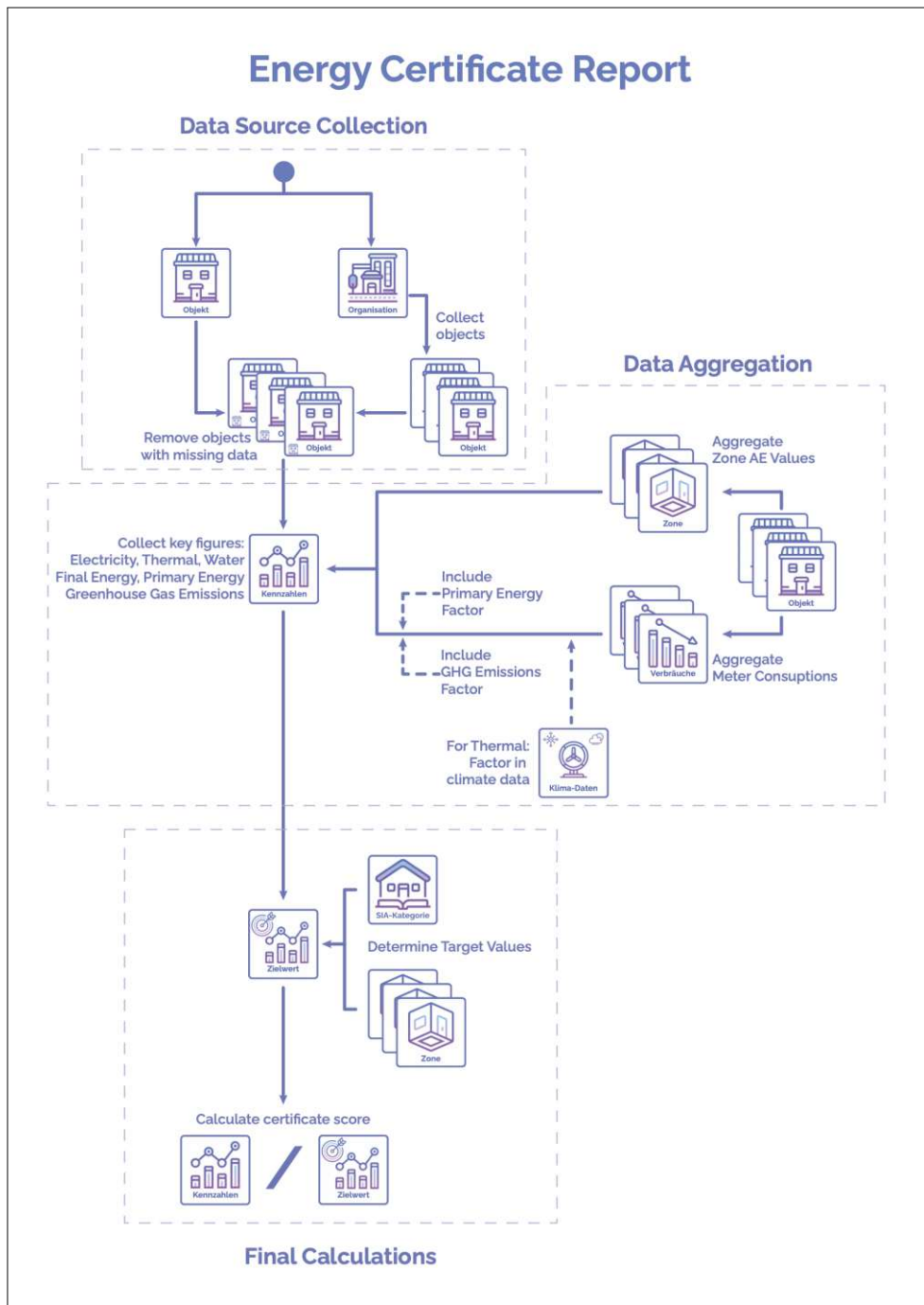


Figure 4.12: Initial comparison chart visualization representing the technical implementation provided by me.

learnings (expert) users could and should take from such a visualization, instead of giving them a quick overview and ideally pointing them in the right direction towards further, more detailed documentation. Opting for less accurate—or sometimes even outright inaccurate—representations, while staying true to the larger picture, proved to be the more promising approach.

Another important observation emerging from this process suggested the applicability of these flowchart representations to visualize high-level algorithmic processes for an expert audience. Asking the group to assess for whom this kind of visualization would be most useful prompted a broader discussion on the various levels of algorithmic literacy and domain knowledge required to interact with the EnerCoach tool in general. The result of this discussion was a taxonomy of user groups according to their knowledge, split by tasks performed with the help of the EnerCoach tool. The group identified user groups with high (“red”) , medium (“blue”) and low (“green”) requirements towards their literacy regarding the tool in particular and knowledge about energy accounting processes in general, and assigned them to various data entry and reporting tasks. The resulting categorization is shown in Figure 4.13.

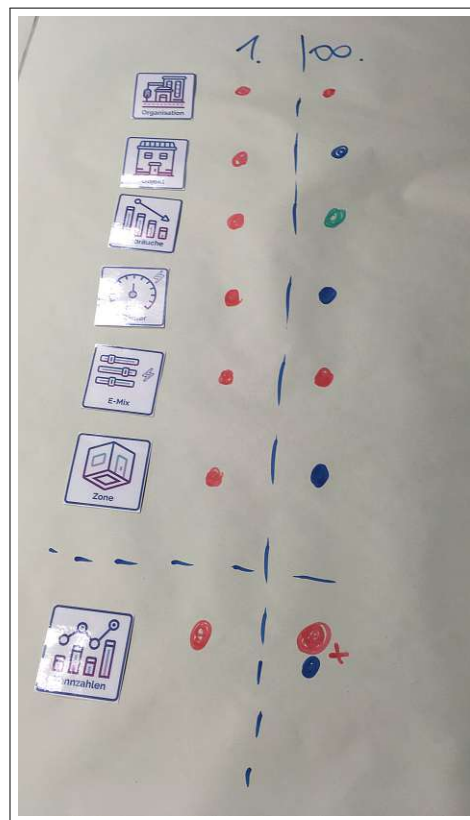


Figure 4.13: Workshop participants’ assessment of knowledge requirements for various data entry and reporting tasks. Red marks high, blue refers to medium, and green denotes low requirements for a given task.

Here, the group also differentiated between the first time a task was performed for a community and repeated task executions, generally agreeing that—ideally—the first time required a higher level of understanding than subsequent iterations. For instance, discussing the requirements for entering the yearly consumption data led to this exchange:

“Consumption data [entry] is ‘green’, but meter definitions are ‘red’. But I think, if we have ‘red’ at the object level, then the definition and collection of how many meters there are must be ‘red’ as well.”³⁶

EC Working Group member “A”

“The same goes for the consumptions though... I mean, it’s only ‘green’ for the repeated times, and to collect consumptions the very first time should be ‘red’ as well, because you have to first find out, which product is that, how does the invoice look, do I have an invoice... You have to follow up with this, and a layperson is overwhelmed by this! The facility manager doesn’t have the invoice, and it’s quite a process to get it...”³⁷

Energy consultant “G”

“That should be the project manager, actually. They should be part of all these tasks, actually, or even do it alone for a community, and then do it all themselves. But here they could delegate to a ‘blue’ person, and the ‘blue’ person could enter data here, here and here [...]”³⁸

EC Working Group member “C”

Considering that the group agreed that reporting activities would most likely be done by expert users with few exceptions for repeated report generation that might be delegated, the group also concluded that process visualizations for those reports would be most helpful for the stakeholder groups with advanced domain knowledge, as opposed to the larger group of end users. In particular, the participants noted that the hotline staffers might profit from such a process overview when answering support requests, and could use the flowcharts as a kind of checklist to validate the various data inputs one by one.

³⁶Orig. *“Also, Verbrauchsdaten [Eingaben] sind grün, aber Zählerdefinitionen sind rot. Aber ich denke, wenn wir auf Objektebene Rot haben, dann ist die Definition und das Zusammentragen, wie viele Zähler gibt’s da, für mich auch rot. [...]”*

³⁷Orig. *“Und das ist dasselbe bei den Verbräuchen auch... also, es ist erst beim wiederkehrenden grün, und die Verbräuche das erste mal zu erfassen gehört eigentlich auch ins Rote, weil Du musst auch beim Verbrauch schauen, welches Produkt ist das, wie sieht die Rechnung aus, hab’ ich eine Rechnung... Ich muss dem nachgehen, da ist ein Laie überfordert! Der Hauswart, der hat die Rechnung nicht, und es ist ein Prozess, daran zu kommen. [...]”*

³⁸Orig. *“Das sollte eigentlich der Projektleiter sein. Der kann eigentlich überall dabei sein, oder auch alleine bei einer Gemeinde, und in Folge macht der einfach alles. Aber hier kann er den Blauen delegieren, und der Blaue kann hier und hier und hier eingeben [...]”*

Finally, the discussion as part of the collaborative exercise also yielded a few alternative ways of representing the algorithmic process with the help of the entity cards (e.g., a circular visualization depicting the relative contribution or importance of the various data entities, or a timeline reminiscent of a Gantt chart showing the static and dynamic data point inputs in a given year). These alternatives notwithstanding, the participants settled on a visualization remarkably close to the one provided by me, and had no problem comprehending these general flow-chart type visualizations. It should be noted as a limitation to this insight that subliminal cues nudging the participants towards one type of visualization over another (such as the use of the term ‘process’) can never be avoided with the instructions for such a collaborative exercise, and their impact is very difficult to quantify. Furthermore, the choice of using entity cards to provide a common visual language may likely have impacted the groups direction towards these visualizations as well.

4.5.2 Part 2: Concrete Measures

The second part of the workshop provided the participants with an open forum to brainstorm and design concrete technical measures for the existing user interface of the EnerCoach tool to address issues of *ex-post explainability*. The participants were provided with a set of laminated screenshots for the various interfaces of the EnerCoach system, including the different reports, for the same sample community shown in the report graphics in Appendix [A.2](#).

With these laminated screenshots as reference, and through the use of removable markers, as well as additional pens and paper, the participants were asked to create mock-ups of potential interface features and paper prototypes for potential interactive widgets that would help them trace and validate report results, troubleshoot implausible outputs and generally support the sense-making processes users need to accomplish in their daily use of the system. Figure [4.14](#) shows examples of some of these annotated mock-ups for the *renewable energy report* and the *energy certificate report*. The resulting suggestions were ranked by the workshop participants in order of both feasibility and usefulness, and were subsequently implemented by WIENFLUSS and rolled out to the production instance of the system. For this part of the workshop, I contributed as a *technical expert*, answering questions about the technical feasibility of their ideas and designs, and offering advice on user interface design paradigms and best practices where applicable.

The most promising concrete suggestion the group designed was aimed at helping bridge the gap between user-entered, dynamic data points—such as consumption data—and the aggregate, joint results of the *key figure reports*. As part of the discussion, some workshop participants noted a frequent issue they found themselves confronted with when giving support to community users: After the community had finished entering data for a given year, and generated the first key figure reports (e.g., Figure [A.3](#)), they would often find significant outliers in key figures for a specific year. Since larger communities may have entered data for dozens of buildings and hundreds of meters per year, tracing which building or which meter may be responsible for these outliers was a challenging task.

4. CASE STUDY: ENERCOACH

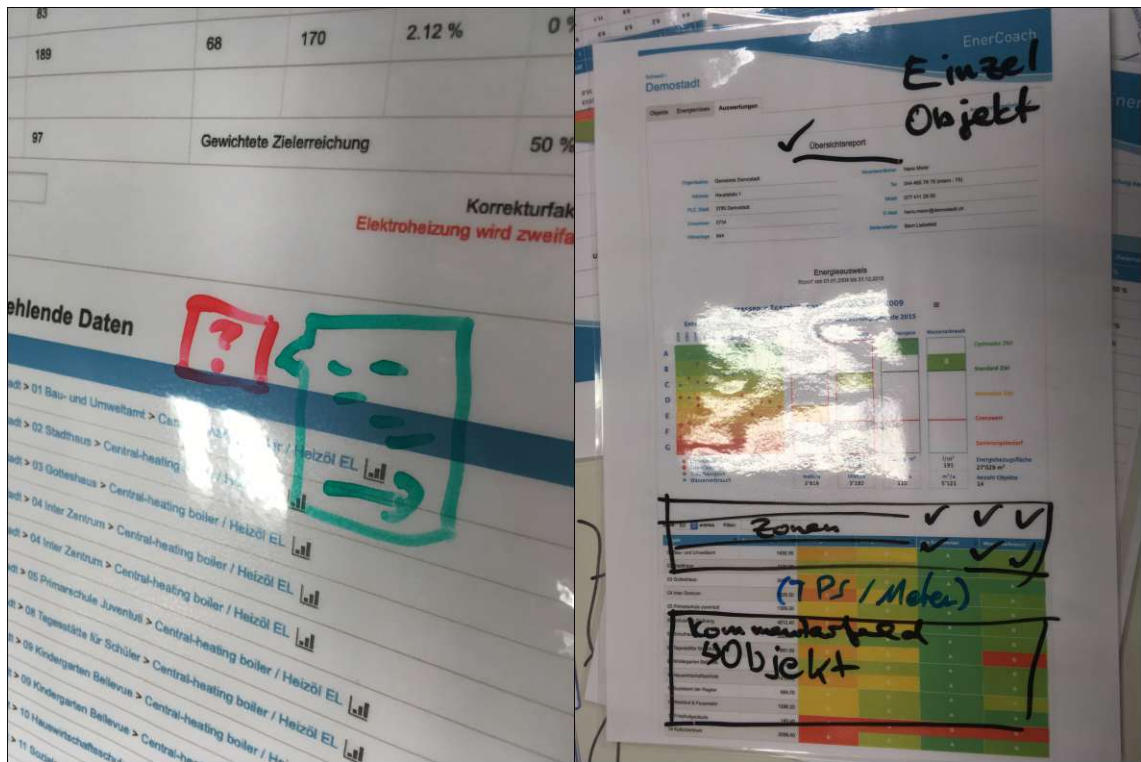


Figure 4.14: Two examples of mock-ups through annotating laminated screenshots, showing a mock-up design for a contextual help button, and an addition of a contextual data table for the overview report.

Even after identifying which buildings may be contributing to these outliers the most, distinguishing between anomalous results (due to, e.g., data entry errors), actual software errors, and legitimately lower key figures often proved very difficult. To support this sense-making task, the group suggested the addition of user-defined contextual comments or notes that could be entered at the same time as the energy consumption data was entered in the form of *annotations*. When generating a report, these annotations should be displayed adjacent to the report graphic, to help explain potentially anomalous, but legitimate results. This measure directly relates to the example given towards the end of Section 4.3: buildings that have significantly reduced energy consumption due to extraordinary circumstances in a given year (such as parts of the building being shut down for renovations, and the subsequently reduced energy consumption for heating or electricity) would receive a note during the data entry process, and would thus be easily recognizable in the reports. As Wood et al. [311] note in their study of sense-making processes in eco-feedback tools, this kind of contextual annotation has been shown to positively impact user understanding, and thus seemed a promising approach to reduce the potential for drawn-out and frustrating validation processes.

In addition to this specific measure, the group also provided a list of smaller, but potentially impactful measures and improvements to the user interface. Stemming from the success of the common visual language of the entity cards, the various inconsistencies in language and terminology previously identified across the system would be homogenized to avoid confusion. Other suggestions included a site-wide implementation of a contextual help system, allowing users to access minimal explanations or help texts for various interface and report elements, as well as a new layout and design for the list of missing data shown underneath each report.

4.5.3 Post-Workshop Implementation and Evaluation

At the end of the workshop, the group prioritized the potential measures to be implemented and evaluated by WIENFLUSS.

Following the success of the visualization for the *energy certificate report*, the group also suggested the *evolution of energy consumptions* report as a potential subject for visualization. Based on the feedback given in the workshop, I adapted and designed the improved versions³⁹ of the visualizations for both reports shown in Figure 4.15.

The resulting flowcharts were provided to EnerCoach hotline staffers, and feedback was gathered via email for a final round of evaluation. Generally, this feedback was positive, as the charts were seen as helpful tools to gain a system- and model-level overview of the constituent parts and their relative interactions for the reports depicted, with one hotline staffer describing the charts as “*super practical*”⁴⁰. They also noted where the visualizations closed certain gaps in documentation relevant to their support work. For instance, the *evolution of energy consumption* report visualization detailed the contribution of the various energy carriers to the colour-coded groups shown on the right of the stacked area chart (see Figure A.5). Some of these groups are comprised of energy generated by multiple carriers, e.g., “*Wood/Biomass Heating*” aggregates consumptions of *Wood logs*, *Wood chips*, *Wood pellets*, *Charcoal* and *Biogas*. Also noted as a positive, the *common visual language* approach was seen as helpful in the iterative sense-making processes, connecting the terminology used in various other parts of the system to the report system. While not mentioned directly, the use of pictorial representations may have had positive effects on memory and recognitions, as a wealth of prior research in HCI and Cognitive Sciences research suggests [319, p.225-227].

As proven by the workshop, the visualizations could be helpful for *system-* and *model-level transparency*; however, their limitations most prominent lie with supporting ex-post explanations of report results. Since they only offer an abstracted, high-level depiction of the process, they do little to explain the specific algorithmic processes that lead to a given system output in the form of a report graphic; at best, they can illustrate “*where*

³⁹For the improved versions, both German and English versions were produced; the German versions were the ones feedback was given on, but the English version is reproduced here to allow better comprehensibility for a broader audience.

⁴⁰Orig. “*super praktisch*”

4. CASE STUDY: ENERCOACH

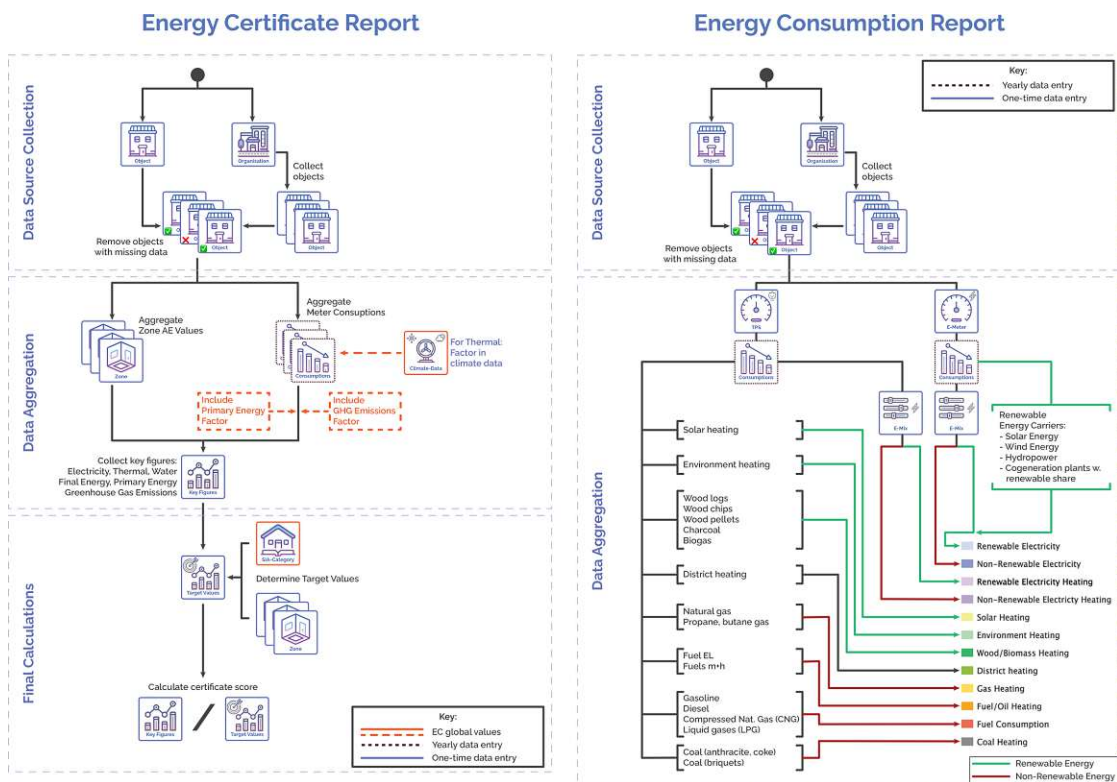


Figure 4.15: Final report visualizations for *energy certificate* and *energy consumption* reports.

to look”, but offer no help beyond that initial indication. The second larger feature that was implemented after the workshop—the contextual annotations for user-entered data—addresses this aspect specifically. After the feature was green-lit and rolled out to the production instance of the system, the key figure report for single buildings would show annotations given by the users next to the reports themselves. Figure 4.16 shows such a report with a comment explaining usage patterns for the building’s heat pump.

A detailed qualitative evaluation of this feature was not possible due to limitations in access and availability⁴¹ of potential interview partners. In lieu of such an approach, a quantitative analysis of usage patterns and the utilization of this feature was performed. This analysis showed that, out of a total of 15853 thermal production systems, roughly 23% have user-defined comments as of the publication of this dissertation. For electricity meters (19928 in total, roughly 9% with annotations) and water meters (12146 in total, roughly 6% with annotations), the percentages that have these annotations attached are lower, but still show a significant overall use of the feature. While it may not be valid

⁴¹At the time, the COVID-19 pandemic had reached Austria and Switzerland, and most of the participants in the workshop, as well as my other points of contact, were unavailable at the time for personal or medical reasons.

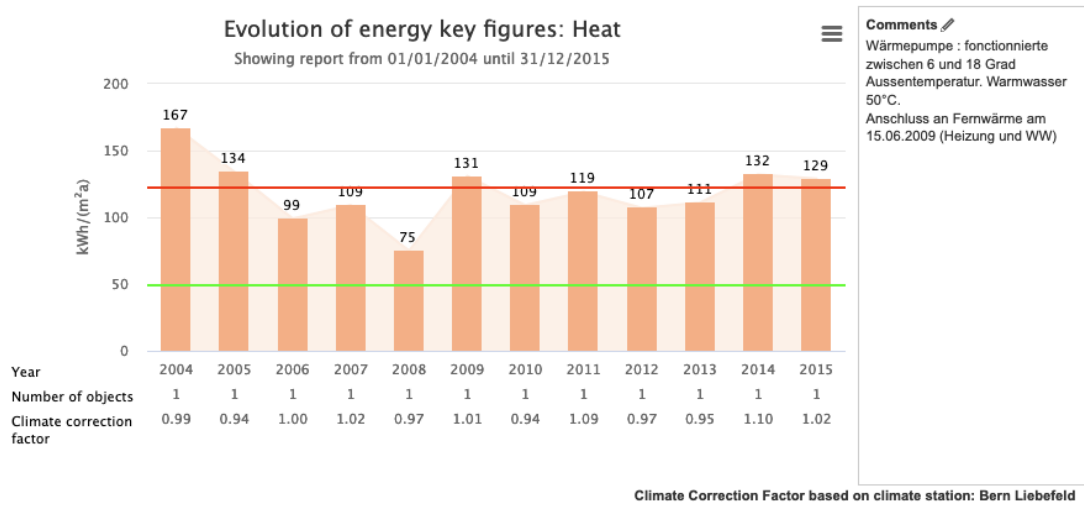


Figure 4.16: Adapted key figure report showing a user-entered annotation on the right side.

to conclude that the *use* of a feature equals its *usefulness*, an outright rejection of this functionality would surely have indicated its failure.

4.6 Chapter Summary

Summing up the learnings derived from this case study of the EnerCoach energy accounting system, two perspectives are worth considering: first, the results of the system analysis and description, transparency measures and improvements resulting from the PD workshop, and second, the evaluation of the methodologies employed in this study.

For the former, an in-depth situated ethnography revealed the complex interplay between stakeholders, the system's history, aims and goals, its technical implementation and its contextual operationalization. Tracing the choices and value judgments made as part of the specification and implementation process helped elucidate the more opaque aspects of the system, and provided answers to [SRQ2.2](#) and [SRQ2.3](#). In particular, this put a spotlight on some of the primary factors contributing to its shortcomings in terms of *model transparency* and *ex-post explainability*: the interwoven nature of descriptive and normative aspects of energy accounting coded into a technological artefact, the complexity of the underlying reporting system, and the heterogeneous group of stakeholders with differing needs and levels of literacy. Equally important as identifying these factors was to determinate what was *not* the cause for a lack of transparency in this system: the opacity of the EnerCoach system is clearly an unwelcome side-effect and not an intentional design choice or policy. Based on these observations, the next steps to mitigate the issues could be guided towards very different types of measures. For system-level transparency, the co-created process visualizations were shown to support advanced and expert users in their day-to-day tasks as they interacted with the system and its end-users. For ex-post

explainability, the small—but significant—improvements to the user interface were shown to support the cognitive sense-making processes of its users.

For the latter, methodological aspect, the case study has shown the positive impact of a holistic and inclusive perspective on an algorithmic system as a socio-technical assemblage. The insights gathered as part of the situated ethnography would not have been possible by simply reading the system’s code, analysing its user logs, or reading the original specifications, without the relevant contextual information provided by the stakeholders, or the analysis of their interactions, tasks and use of the system. Taking the opposite stance, an analysis relying solely on the observations of the various users would also have resulted in a biased and limited perspective on the system’s underlying technology, the complexities inherent to its implementation, and the limitations imposed on *system-* and *model-level transparency* and *ex-post explainability*. Few scholars would refute the merits of more empiric data and more diverse perspectives on an algorithmic system; however, the fact that such an analysis is only possible in very rare cases should not be overlooked as well. In the case of EnerCoach, this access came in the form of the *auto-ethnographic nature* of the study due to my own involvement in the system, and has proven to be a valuable approach to uncovering the most relevant avenues of inquiry.

Finally, the use of participatory paradigms and trans-disciplinary methodology, borrowing from the scientific communities of [HCI](#), [CSCW](#), and Design Studies [\[320\]](#), has proven to be a particularly fruitful approach to study issues of algorithmic transparency and provide answers to [SRQ2.1](#). While the efforts involved in organizing and conducting participatory design workshops may be prohibitive in some cases, where possible, they promise a deeper level of reflection for both scholars and participants, and subsequently, more applicable and targeted results. Bridging the gap between the *social* and the *technical* in *socio-technical systems*, participatory design with heterogeneous groups of stakeholders can help improve the agency of all involved parties alike through a shared process of learning and emancipation. Seen through the lens of assemblage thinking, participatory design intrinsically promotes a shift in agency towards the participants, as they gain the ability to more directly influence the non-human, technical components of the assemblage. How they utilize their agency may vary from case to case: while the participants in this case study chose to design measures aimed at improving their own, human agency to understand and explain the technical components, other stakeholders in different assemblages may have chosen to delegate more power and agency towards the technical parts through automation. In either way, however, participatory design used well empowers the participants to make these choices themselves, and thus—as suggested by Kemper and Kolkman [\[28\]](#)—elevates stakeholders of an algorithmic system to a *critical audience* as a prerequisite to achieving true algorithmic transparency.

4.7 Chapter Conclusions

Beyond the concrete results of the EnerCoach case study, some larger conclusions in the overall context of this dissertation can be made.

Relevance to the field

The first point to address retrospectively is the relevance of this case study to the field of [CAS](#) as a study of algorithms and their power, as opposed to a study in [HCI](#) on system design or visualizations. The reason this relevance may not be as obvious as, for instance, with the [AMAS](#) case study, lies in the immediacy of what is *at stake*. With AMAS, the consequences of biased classifications, e.g., systematic discrimination of individual jobseekers, are immediate and obviously dangerous at a potentially large scale. For EnerCoach, by contrast, these consequences are not as immediate, and much less clearly defined. Quantifying a loss of trust in the system over time, or measuring the impact of all the time lost by stakeholders trying to pinpoint the reason behind an implausible or false report result is much harder than identifying the harm done to a jobseeker who was just denied access to potentially vital resources based on questionable predictions. To dismiss EnerCoach as a case study in [CAS](#), however, would be a mistake: after all, the EnerCoach’s algorithms embed values and normative assumptions just as much as [AMAS](#), and EnerCoach exercises algorithmic power and authority on its stakeholders as well. For EnerCoach’s stakeholders, being able to trace and assess the plausibility of report results to their own inputs as well as correlate them to their understanding of the system-internal assumptions, calculations, and constants, is more than just a matter of convenience. For some of the stakeholders, knowing that they can rely on the system is the foundation upon which they build their professional reputation, i.e., their career as energy consultants that present reliable suggestions for improvements to the communities they work with. For others—like the EnerCoach Working Group—the overall correctness and reliability of results may determine how successful this tool is and continues to be—a tool that they clearly believe to be a vital component in the fight against climate change and that they invested years to develop and promote.

EnerCoach also represents the same challenges of perceived objectivity and trust in automation as many other systems do, and the case study helped shed light on these issues. As the thick description of the system showed, the calculations and reports are decidedly normative in nature, including both explicit assessment of performance (in the form of target/threshold values and ratings) and implicit implementation choices such as the silent doubling of electric heating consumptions. At the same time, the study also showed a limited awareness and lack of clarity for many stakeholders regarding exactly which parts of the system were based on value-laden choices. Transparency, as I have argued, extends not only to answering *how* the system behaves, but also disclosing *why* it does so. But the mere availability of these answers alone is not enough if stakeholders do not ask the question or seek out that information. Trust in the system may, perhaps surprisingly, thus lead to less overall system transparency, as users assume they do not *need* to know all the details, because they trust that the system presents an objective truth or accurate representation of reality—both of which are, of course, inaccurate assumptions. Only in the moments of doubt, when a lack of *ex-post explainability* spotlights a gap in their understanding, those questions become more pressing, and the need to determine the cause of implausible results forces a reevaluation of that trust

and create the need for *micro-accountability*. To productively utilize that moment as an opportunity means also to share the information about the underlying assumptions and normative aspects of the system, and to raise the awareness of the stakeholders about the possibility of faults in the system. In other words, micro-accountability processes in EnerCoach are part of a cyclical relationship with the issues of transparency and explainability: limited *ex-post explainability* may both hinder and trigger *accountability* processes, while a successful *accountability* process carries the potential to inform the *forum*, thus increasing the overall *system-level transparency*—which, in turn, provides a better foundation for *ex-post explainability* again.

Thus, being able to both *count on* and *hold to account* the system makes a significant difference to those involved, even if the stakes are not as immediately obvious as they may be for other systems. EnerCoach, as a case study, thus should be seen as an example of the importance of looking beyond the high-profile and more placative cases of algorithmic systems. The fact that a system like EnerCoach is, at first glance, neither particularly controversial, nor seems capable to cause immediate harm, does not mean that such a system should not be held to account according to similarly high standards as other, more controversial systems. At the same time, the non-controversial nature of the system also means that system stakeholders may be more open to participation in research and design processes, presenting an opportunity to gain access and generate insights that are more difficult to attain in controversial cases. Finally, I would argue that the strong focus of the field of **CAS** on controversial and problematic cases represents a lost opportunity to showcase the potential of non-controversial cases of algorithmic systems and technologies: Using frameworks like the **A³ framework** presented in Chapter **6** can help design algorithmic systems that serve as models of good practice of *transparent, explainable* and, ultimately, *accountable algorithmic systems*.

The Benefits of Theoretical Contradictions

Many of the theoretical foundations of this dissertation presented in Chapter **2** offer useful abstractions and high-level perspectives: they allow identifying overarching themes and attributes of algorithmic systems, and the issues related to them. Some of these are easily and directly applicable, such as Bandura’s concept of human agency: considering where and how an individual stakeholder’s agency is affected by their knowledge about the system, the technical capabilities, but also by their own belief in *self-efficacy* offers clear conclusions about what aspects of the system may warrant improvement, and how to approach them. Furthermore, the introduction of *proxy* and *collective human agency* spotlights the strategies stakeholders may follow to achieve their goals if their individual agency is insufficient to do so, and suggest strategies to support their collective agency (such as the introduction of custom annotations).

Other theoretical foundations, however, present certain difficulties and create tensions arising from their application to concrete and specific examples. Assemblage thinking, in particular, exemplifies these tensions well: As an analytic tool, considering EnerCoach as a socio-technical assemblage helps foreground important facets beyond the human

and technical aspects, and the role they played in its co-production. The underlying assumptions and intentions, the contextual limitations placed on the system by legal, technical or social boundaries, are all part of its assemblage, and worthy of consideration in terms of their impact on the agency of the entire assemblage and its component parts, human and non-human alike. At the same time, this perspective does not come particularly naturally, as the human experience does not typically include considering such facets as having agentic properties—indeed, considering the agency of technical aspects of the system already requires some serious mental legwork. Thus, the perspectives offered by such theoretical considerations, as helpful as they may be, require a disciplined reflection in order to avoid falling back to more conventional perspectives on algorithmic systems. This challenge is particularly difficult when engaging with the system and its stakeholders more closely, as was the case with this study due to its auto-ethnographic and participatory nature.

From these challenges, an observation on the useful ambiguity of different theoretical conceptualizations of algorithmic systems emerges. As different approaches—from narrow technical definitions to socio-technical systems, assemblage thinking and ANT or functional perspectives—offer different insights, so varies their applicability and usefulness in different contexts. Since none of those perspectives represents an absolute truth or exclusive definitions, however, the tensions and contradictions between them that arise from these variations also offer important insights. As we struggle to reconcile, for instance, individual and distributed agency of actors, or a lack of *ex-post explainability* of a demonstrably *interpretable* system, a core benefit of interdisciplinary research becomes evident: a forced change of perspective that, in all its contradictions, offers more potential for insight than following a single, well-fitting but ultimately limited model.

Auto-Ethnographic Reflection

As I conducted this case study in an auto-ethnographic manner, some personal reflections can help elucidate the process and the benefits and challenges that come with this approach.

As a long-time contributor and technical lead for the system for many years prior to the start of the research project, I experienced the process of empiric data collection as both rewarding and challenging. Throughout the project, there were many small moments of clarity, insight and realizations that I could directly relate to my prior engagement with the system and the knowledge about its technical aspects I already possessed. These moments always carried both the excitement of sating my epistemological curiosity in a way another researcher with less prior knowledge might not have been able to, and a cautionary voice telling me to critically reflect on my assumptions, to verify these insights through other, empiric data, and to question my own biases. Thus, much of the process was a balancing act between advancing quickly and slowing down purposefully, leading to an, at times, exhausting back-and-forth and making it difficult to gauge when and where to end the case study. For the future, a clear decision on the boundaries of research

in terms of when to stop collecting data thus seem an important strategy to avoid an endlessly drawn out continuation of the auto-ethnographic process.

At the same time, I was surprised how liberating the repeated reflection on my dual role as a researcher and developer turned out to be. Having had a personal and professional stake in the system for years as a developer, I had experienced many moments of frustration due to the limitations posed by tight budgets and limited resources—moments in which I felt I knew how to improve the system, but could not do so. As I was engaging in the constructive participatory design process, I kept reminding myself of the role of a researcher trying to approach the stakeholders' ideas with an open mind and unhindered by the specific limitations posed by scarce resources. As a result, exploring these ideas as a researcher also felt incredibly motivating as a developer, regardless of their potential for actual realization, and re-opened avenues of thought I had abandoned earlier in light of their feasibility.

Finally, I feel incredibly lucky for the respect both WIENFLUSS and the EnerCoach stakeholders showed towards the scientific process. Reviewing the empiric material I gathered, including the interview and workshop process, as well as email communication surrounding the project, I realized that any feelings of pressure to succeed, to provide tangible and useful outcomes as a result of the project were my own, and not plausibly triggered by the study participants. Whether or not this was the consequence of the honesty and clarity with which I tried to communicate the boundaries between scientific research and commercial interests, I can not determine with certainty. The fact that I had to consider these questions for myself before starting the research process, in order to be able to communicate them to the (at that point) prospective study participants, however, almost certainly helped regulate my own expectations and allowed me to reflect on my own expectations more critically. Thus, such *a priori* considerations before engaging in auto-ethnographic research have proven invaluable, and I would recommend prospective auto-ethnographers to do the same.

Case Study: AMAS - The AMS Algorithm

This chapter introduces the [AMS](#)'s [Arbeitsmarkt-Assistenz-System](#) ([AMAS](#)) as a case study of a highly problematic [ADS](#) system used to profile and categorize jobseekers according to their predicted chances of re-integration into the labour market. After an exploratory vignette to contextualize my engagement and history with researching the system in [Section 5.1](#), relevant prior research on [ADM](#) and [ADM](#) systems from a critical perspective is summarized in [Section 5.2](#). Following this, I provide a detailed description of the system, its components, stakeholders, as well as its operationalization in the context of the AMS in [Section 5.3](#). The core issues of bias, discrimination, transparency and explainability addressed in this case study are the focus of [Section 5.4](#), and a summary of the results rounds out this chapter in [Section 5.5](#). As to the larger focus of this dissertation, the comparative case study and [A³ framework](#) presented in [Chapter 6](#) will address issues of algorithmic accountability related to this case study.

I have (co-)authored various previous publications on the [AMAS](#) system, which contribute to and partially overlap with the content presented in this chapter. First and foremost, the research cooperation between the [CIS](#) and the [ITA](#) at the [ÖAW](#) resulted in the first English-language journal contribution on the system [\[3\]](#)¹. While it represented an important first step towards our understanding of the system, this publication was limited by the lack of access to internal documents, and subsequently only provided a specific perspective on the system shaped by those documents that were publicly available at the time. Following this, the research group behind this journal contribution—including² [Doris Allhutter](#) ([ITA](#)), [Fabian Fischer](#) ([CIS](#)), [Gabriel Grill](#) (University of Michigan at Ann Arbor), [Astrid Mager](#) ([ITA](#)) and myself—embarked on an in-depth analysis of

¹Publication title: “*Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective*”

²Collaborators listed in alphabetical order by last name.

the system, supported and co-financed by the [AKOÖ](#), which resulted in a significantly more comprehensive research report published by the [ITA](#) [\[4\]](#)³. As a consequence of this cooperation and the contributions to this case study, any content written in the first-person grammatical form that is pertinent to this case study will be done in *plural* to honour and reflect the collective effort contributing to the research; any personal deictic references and possessive pronouns such as *we*, *us*, or *our* thus refer to our research group.

Other relevant publications include the comparative study presented in detail in Chapter [6](#) [\[6\]](#)⁴, and a publication in the Austrian journal “Juridikum” [\[5\]](#)⁵.

Given the fact that the most in-depth and up to date analysis of the system [\[4\]](#) was published in German, this chapter will provide access to this analysis for English-speaking scholars as well. Furthermore, the most relevant research results from these previous publications I contributed to will be synthesized into a coherent case study. As such, this chapter is a necessary foundation for the arguments made in the following Chapters [6](#) and [7](#) as well.

5.1 Exploratory Vignette

While my research on the [AMAS](#) system was not auto-ethnographic in nature, the following exploratory composite vignette [\[60\]](#) as a narrative introduction to my engagement with this system still provides a valuable contextualization of, and introduction to, the AMS Algorithm.

In the early fall of 2018, in a moment of weakness, I succumbed to the urge to procrastinate. I opened the website of Austrian newspaper [derStandard](#)⁶ to find a lead article titled “*AMS to assess jobseekers by algorithm in future*”⁷ [\[265\]](#). Since I was looking for both practical and Austria-specific case studies of algorithms to discuss with my students at the time, my interest was piqued immediately, and I set out to investigate. After skimming the article, and the related interview with [AMS](#) CEO Johannes Kopf posted alongside [\[264\]](#), two things became immediately apparent: (1) this system was both highly relevant to my research, and potentially highly problematic for a variety of reasons, and (2) I had successfully failed at procrastinating and was back on track to start a new research project.

My first concern about this system was the issue of transparency: while the article explained vaguely that the system would “*assess all Austrian job-seekers’ prospects*”⁸

³Publication title: “*Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*”

⁴Publication title: “*The agency of the forum: Mechanisms for algorithmic accountability through the lens of agency*”

⁵Publication title: “*Der AMS-Algorithmus. Transparenz, Verantwortung und Diskriminierung im Kontext von digitalem staatlichem Handeln*”

⁶<https://derstandard.at>

⁷Orig. “*AMS bewertet Arbeitslose künftig per Algorithmus*”

⁸Orig. “[...] *die Perspektiven aller Arbeitslosen in Österreich bewertet*”

and would “categorize them into those with low, medium and high chances to enter the labour market”⁹ [265], how exactly that system would make such a determination was not explained. However, following the trail of breadcrumbs of what information was available online about this system at the time, I was surprised to find the publication “The AMS Labor Market Chances Model: Documentation of Methods”¹⁰ [DOK_1] claiming to transparently describe the system on the AMS Research Network¹¹ website. Given the amount of intentional opacity I was used to encounter when looking into algorithmic systems, this was, at first glance, a refreshing departure from the otherwise dissatisfactory status quo of transparency in algorithmic systems. Reading the document, my positive surprise soon turned into worry, as I found the illustration shown in Figure 5.1, described in the document as estimated equation coefficients of short-term integration chances for jobseekers with fully available data. The equation lists the estimated average positive or negative impact certain personal attributes of jobseekers had on their chances of finding a job, and featured such telling lines as “-0.14 x GENDER_FEMALE”, “-0.70 x AGEGROUP_50_PLUS” and “-0.15 x RESPONSIBILITIES_OF_CARE”¹².

```

BE_INT
= f ( 0,10
    - 0,14 x GESCHLECHT_WEIBLICH
    - 0,13 x ALTERSGRUPPE_30_49
    - 0,70 x ALTERSGRUPPE_50_PLUS
    + 0,16 x STAATENGRUPPE_EU
    - 0,05 x STAATENGRUPPE_DRITT
    + 0,28 x AUSBILDUNG_LEHRE
    + 0,01 x AUSBILDUNG_MATURA_PLUS
    - 0,15 x BETREUUNGSPFLICHTIG
    - 0,34 x RGS_TYP_2
    - 0,18 x RGS_TYP_3
    - 0,83 x RGS_TYP_4
    - 0,82 x RGS_TYP_5
    - 0,67 x BEEINTRÄCHTIGT
    + 0,17 x BERUFSGRUPPE_PRODUKTION
    - 0,74 x BESCHÄFTIGUNGSTAGE_WENIG
    + 0,65 x FREQUENZ_GESCHÄFTSFALL_1
    + 1,19 x FREQUENZ_GESCHÄFTSFALL_2
    + 1,98 x FREQUENZ_GESCHÄFTSFALL_3_PLUS
    - 0,80 x GESCHÄFTSFALL_LANG
    - 0,57 x MN_TEILNAHME_1
    - 0,21 x MN_TEILNAHME_2
    - 0,43 x MN_TEILNAHME_3)

```

Figure 5.1: Estimated coefficients for short-term integration chances of jobseekers with fully validated data. Graphic excerpted from [DOK_1]

⁹Orig. “[...] in drei Kategorien einteilen: in jene mit hohen, mittleren und niedrigen Chancen, am Arbeitsmarkt unterzukommen.”

¹⁰Orig. “Das AMS-Arbeitsmarktchancen-Modell: Dokumentation zur Methode”

¹¹<https://www.ams-forschungsnetzwerk.at>

¹²Variable names translated from German by the author

This specific graphic would later rise to some infamy as it was published in various newspapers and articles, as either an example for the discriminating nature of the system or the labour market as a whole, depending on who was writing the article. At the time, I realized that a critical perspective on the use of such systems in general, and the goals and implementation of this system in particular, was not just of academic interest, but relevant to the public discussion as well. Reaching out to a technology journalist I had given interviews to previously, I offered my help as an expert from the perspective of [CAS](#), and gave an interview later published on the newspaper's website [\[321\]](#). The original article would subsequently lead to a whole series of articles, investigating various perspectives on the system, and describing what was known at the time. At this point, my decision to conduct an academic research product focusing on the system as a case study was already made.

Fast-forward a few months into early 2019, the public debate in the system had already risen to include controversial articles, opinion pieces and other elements of discourse, and the public's interest in the system was already beginning to wane again. At the time, I was mostly concerned with trying to collect as much information about the system's technical implementation and internal workings in order to make sense of what contradictory documentation and statements the various stakeholders of the system had published. This interest would broaden, however, as I had the chance to accompany a non-native German speaking jobseeker to the AMS for their initial, mandatory consultation, as a friendly supporter and translator. While I knew that the system was only being evaluated in select pilot offices of the AMS at the time, I still could not pass the opportunity to ask the AMS worker after the consultation, as an interested citizen, what her opinion was on the new system, and whether or not they were already using it to assess jobseeker's chances of reintegration into the labour market. Her answer was unequivocally dismissive: Neither did she think the system was going to be helpful in reducing her workload, but more importantly, she could not understand how the system would be able to provide the same depth of assessment that her many years of experience working with jobseekers would give. In the end, she shrugged her shoulders and said: *"I don't think I'll look at the scores much, even when they become more widely available."*

While the interaction described above was purely anecdotal and should not be taken as more than it was—an offhand impression given by a single AMS caseworker—it made me realize the tensions between the public narrative pushed by the AMS and its representatives, and the practical, real-life impacts the system may or may not have on those affected by it, be they AMS caseworker or jobseekers. On the one hand, the AMS promised a modern, efficient, and fair tool to support caseworkers in their job, and on the other hand, caseworkers might see the system in quite a different light, perhaps even as useless or a nuisance.

These impressions I gathered in my early engagement with the [AMAS](#) system would continue to shape my interest in the system, and eventually lead to the research project resulting in this case study. The various topics and inter-disciplinary aspects relevant to this analysis led me to collaborate with a diverse group of colleagues, who introduced me

to a new set of perspectives and methodologies, and broadened my horizons in [CAS](#) as well. Finally, coming full circle in this vignette, my experiences in studying this system confirmed that sometimes, there may be value and scientific merit in the occasional act of procrastination.

5.2 Prior Research

Among other types of algorithmic applications, [Automated Decision-Making \(ADM\)](#) and [Automated Decision Support \(ADS\)](#) systems have increasingly permeated many areas of society, including the spheres of governance and administration, throughout the last decade. The distinction between [ADM](#) and [ADS](#) systems is fluid and not clearly defined in all cases, but can be roughly made along the scale of human agency in terms of the decision itself: systems that are autonomously making a decision or suggesting a specific decision to a human overseeing the process would be closer or entirely [ADM](#), whereas systems that only contribute to a human’s decision-making process would be classified as [ADS](#)¹³ [\[322\]](#).

In the European Union alone, examples for such systems are abound: For instance, “Algorithmic work activation” in Belgium entails a system predicting long-term unemployment risk of jobseekers based on their online behaviour on the Flemish unemployment service’s job platform; Denmark automatically determines student stipends for higher education; Germany’s credit scoring system “SCHUFA” is just one of many such systems employed throughout the EU; and numerous countries (including Austria, Denmark, France, Germany, Italy, The Netherlands, and Spain) now employ various predictive policing systems in the hopes of improving the efficiency of their police forces [\[323\]](#). While the context of use and the goals of these systems may differ, they share the underlying promises of *objectivity* and *neutrality*—promises that often boil down to little more than a “*mathwashing*” of complex value-laden decisions and unpopular policies: After all, “[*m*]odels are opinions embedded in mathematics”[p.24][\[40\]](#), as Cathy O’Neill put it so succinctly. The public’s trust in this promise seems to be less-than absolute: in terms of the response by the general public, various examples exist of systems whose proposal was met with significant public outrage and resistance, leading to their eventual scrapping (e.g., Denmark’s tracing of vulnerable children as part of their “*ghetto plan*” (sic!) [\[323, p.50\]](#)).

This increase in algorithmic applications of controversial nature across the board has been met with calls from the academic community for increased oversight in the form of [AIAs](#) [\[181\]](#), algorithmic audits [\[324\]](#) and accountability measures supporting *reviewability* [\[325\]](#). While the current state of regulations calling for these measures offer little in terms of *administrative accountability* (see Section [2.4.2.2](#) for a detailed discussion of the state

¹³In the spirit of brevity and comprehensibility, the following section refers to the spectrum of [ADM](#) and [ADS](#) systems as ‘ADM’ systems as a compound term, following the logic that even support systems contribute, in an automated way, to decision-making processes. For the later discussion of the AMS Algorithm itself, the distinction between the two is explicitly made where relevant.

of the art in this regard), some existing regulation is applicable to **ADM** systems in the European Union. The EU's GDPR [170] is a concrete example of such regulation aimed at protecting citizen's privacy and guaranteeing a *right to explanation* for those affected by algorithmic systems (including **ADM**). It has, however, rightfully been criticised as “*uncertain, convoluted, [and] rife with technical difficulties*” [326, p.52]. For the public sector in particular, Kuziemski and Misuraca [327] diagnose a difficult tension between the obligation of governments to protect their citizens from potentially harmful **ADM** systems on the one hand, and the use of such systems for their own purposes on the other hand: “*to govern algorithms, while governing by algorithms*” [327, p.1]. These tensions create a contradictory situation not unlike the conundrum of governmental surveillance through the use of software exploits or spy software relying on unpatched security issues in smartphones. As the controversy [328] around the use of NSO Group's “Pegasus” spyware by more than 45 countries around the world shows, these technologies present a difficult to reconcile double-bind for state actors, as they are both obligated to keep citizens safe from the illegal uses of software exploits, and also support the private businesses dealing in these exploits and their weaponization by paying them for their spyware under the guise of anti-terrorism surveillance. For **ADM** in the public sphere, this double-bind may not hinge on the question of supporting or prohibiting the use of illegal software exploits, but on the question of undue bias and subsequent discrimination, and thus presents a similar conundrum.

These developments in the use of **ADM** systems must be evaluated against the backdrop of a decades-long transformation process of unemployment services in Austria, Germany and Switzerland. Starting the early 1990s and following a neo-liberal doctrine of ‘the market’ as a central organizing principle of the state, the traditional bureaucratic welfare state services were slowly transformed into what Penz et al. describe as “*post-bureaucratic service providers*” [329, p.2]. The Austrian **AMS** exemplifies this process as a semi-autonomous legal entity, governed by a board of politically appointed representatives of state and state-adjacent entities (such as the Ministry of Labour or the Austrian Chamber of Labour). It is financed by the state of Austria, but, at the same time, run by a **Chief Executive Officer (CEO)** in the spirit of ‘New Public Management’ [330]. As a self-governing service provider, the AMS thus is facing expectations towards efficient and effective servicing of the citizens in its care as *customers* or *clients*, in what Penz et al. call “*customer-oriented interaction work instead of bureaucratic administration*”¹⁴ [331, p.1].

¹⁴Orig. “*kundenorientierte Interaktionsarbeit statt bürokratischer Verwaltungsarbeit*”

With this transformation of unemployment services comes a new approach to how jobseekers should be supported in the form of the *activation paradigm*. Penz et al. (summarizing Pascual [332]) characterize this paradigm by its three distinct features:

“[F]irstly, its ‘individualised approach’ aims at changing the behaviour, motivation, and competencies process of individuals in contrast to structural measures against unemployment. Secondly, it is assumed that wage labour is a necessary precondition for social participation and autonomy. Thirdly, ‘contractualisation’ is a ‘core principle’ of the relation between the state and its citizens. Citizens have to sign a contract with public institutions and thereby agree to obligations that need to be fulfilled in order to obtain benefits and to be recognised as a full citizen.”

[329, p.5]

The third characteristic in particular further introduces the notion that receiving unemployment benefits puts a burden on society, and that recipients of these benefits thus have a moral obligation to do everything they can to reach gainful employment (again) and stop receiving such benefits [329].

With this activation paradigm in mind, it comes as no surprise that algorithmic support of such post-bureaucratic services promises a seductive *future imaginary* [333] of a lean, efficient, customer-oriented and personalized services that optimizes government spending on measures supporting the unemployed, thus reducing the overall “burden” placed on society by those requiring support. As we describe in more detail in chapter 6 of our in-depth analysis of the AMS Algorithm [4, pp.89-96], various European countries including Denmark, Germany, The Netherlands, Poland, Sweden, and Switzerland have adopted **ADM** or **ADS** systems in the context of unemployment services. Desiere et al. [334] present an international comparison of profiling approaches across **OECD** member countries, and outline their differences and commonalities in terms of algorithmic support, data sources, profiling methodologies and operationalization. They distinguish between *caseworker-based* profiling, *rule-based* profiling and *statistical* profiling systems (see their Figure 5.2 for an overview of countries employing these methodologies and their overlaps).

Rule-based profiling of the unemployed is used to categorize them into various client groups based on a predefined set of administrative eligibility criteria [334, p.9]. These rules and criteria may be internally (within the service entity) or externally (by policy or laws) determined and are aimed at prioritizing certain subpopulations over others. It is worth noting that these prioritizations do not necessarily follow the doctrine of supporting the most marginalized or disenfranchised subpopulations more than others. Following the “*activation paradigm*” outlined above, an overall strategy of reducing the number of total recipients of unemployment benefits may dictate prioritizing younger, potentially shorter-term jobseekers over the long-term unemployed in order to satisfy quantitative **Key Performance Indicators (KPIs)** imposed on the unemployment service [334].

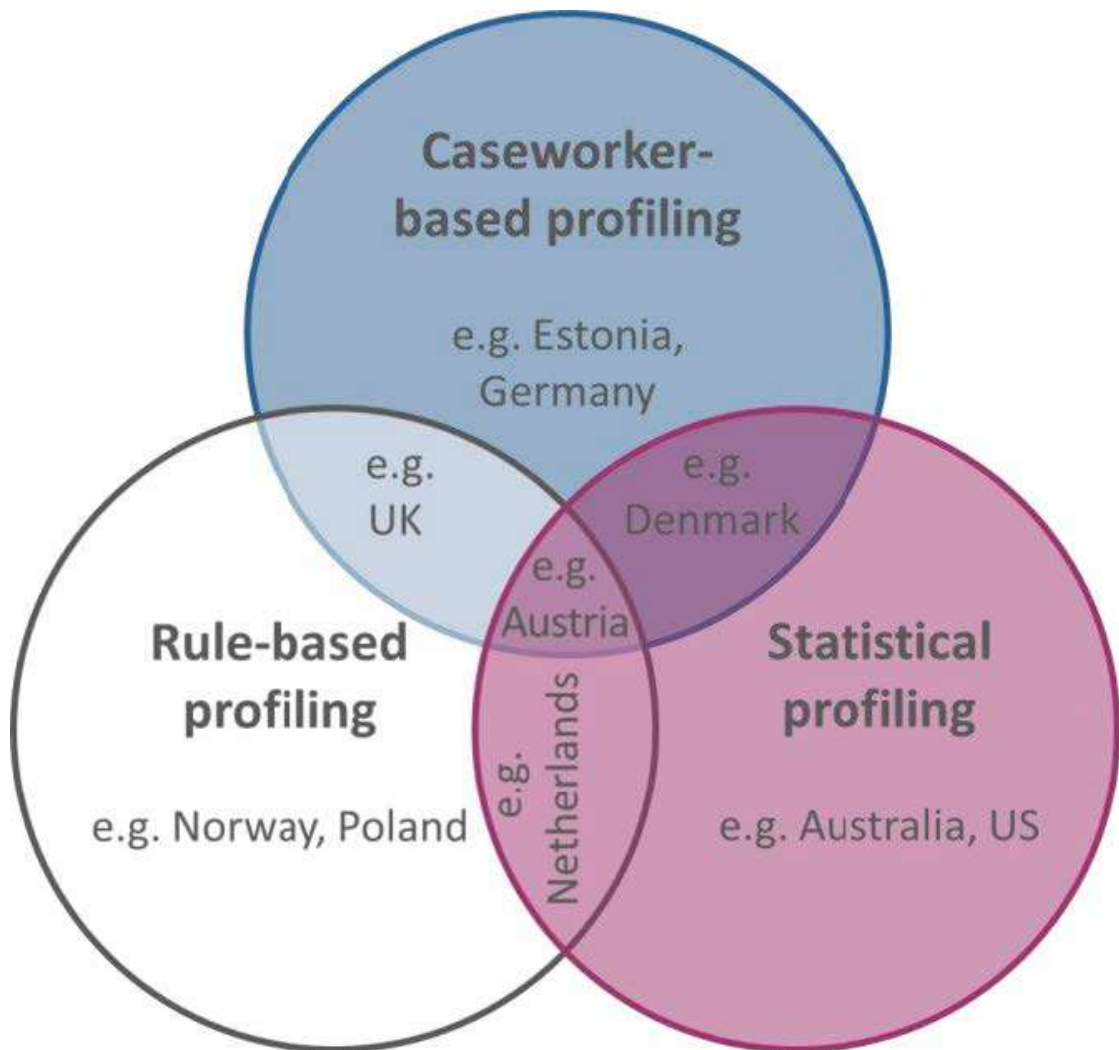


Figure 5.2: Venn diagram of the three main methodologies for profiling systems employed by [OECD](#) member countries as presented by Desiere et al. [334, p.10]

Caseworker-based profiling relies on caseworkers to make an assessment of the jobseekers, often supported by other analytic tools. While this approach relies heavily on the caseworkers experiential wisdom and personal judgment and does not necessarily involve automated or algorithmic processes, standardized assessments can be used to streamline the process (e.g., requiring an assignment of the jobseeker to predefined categories such as “easy” or “hard-to-integrate” as is the case in Germany [334, p.8]).

Finally, *statistical* profiling relies most heavily on automatic processing of jobseeker data to predict various probabilities about individual jobseekers, based on prior observations. Predicted values can include risk of long-term unemployment (used in Australia, Belgium, Denmark, Italy, Latvia, The Netherlands), lifetime income support costs (New Zealand),

or chance of exhausting limited entitlement for benefits (USA). Desiere et al. mention the reduced time and resource costs required by the use of such systems as beneficial over caseworker-based profiling. They also make the highly disputable claim that *statistical* profiling is advantageous over rule-based profiling because it “*has the advantage of considering all jobseekers as individuals and not simply as members of a particular group.*” [334, p.9]. As our analysis of this case study shows, this claim is dubious at best, given the fact that statistical profiling as exemplified by the AMS Algorithm works by categorizing jobseekers into various, supposedly homogenous, groups. It is precisely this problematic assumption of homogeneity in a group of jobseekers sharing a limited set of available data points that leads to bias against certain subgroups, both in terms of overall predictive quality (error rates or precision) and issues of cumulative disadvantage [91].

These various approaches are not mutually exclusive; as Desiere et al.’s choice of visualization in Figure 5.2 attests, overlaps in approaches are not only possible, but rather common. The AMS’s profiling system, however, represents a special case where all three approaches are combined: (1) rule-based profiling for certain specific subpopulations¹⁵, (2) statistical profiling in the form of the AMS Algorithm as the foundation for (3) caseworker-based profiling as the final assessment of a jobseeker’s presumed chances of reintegration into the labour market. This overlap makes this case study a particularly compelling example, as it showcases the complexity and interrelatedness of these approaches in regards to their benefits and dangers.

5.3 Socio-Technical Description of the AMAS System

The AMAS system is an algorithmic system for the statistical profiling of the unemployed. Based on prior observations of a four-year period, the system predicts an individual’s *chance for re-integration into the labour market* or *IC score*. For each jobseeker, two IC scores depending on two *integration criteria* are calculated; based on the values of both scores, the jobseeker is then categorized into one of three possible groups: those with *high*, *medium* or *low* chances of re-integration. Both the percentage values of the IC scores and the resulting categorization are presented to the AMS caseworkers alongside other information about the jobseeker’s case as a supportive measure. Based on the categorization given by the system, the AMS caseworker has limited agency to grant various levels of supportive measures, including educational training, subsidized job placements or referring the jobseeker to external institutions. As the AMS insists, the AMAS system itself makes no definitive decision, and the caseworkers may change the jobseeker’s classification (but not the underlying IC score) if they deem it incorrect. Following this argumentation, AMAS can be characterized as an ADS system, as it does not autonomously make decisions by itself.

¹⁵An example of this rule-based approach is Austria’s implementation of the European Union’s *Youth Guarantee* directive in the form of an *education guarantee*: regardless of real or predicted chances on the job market, furthering their education must be supported by the AMS for those jobseekers under 25 years of age that choose so.

The following sections cover a more exhaustive socio-technical description of the system. The four main aspects discussed include (1) an outline of the [AMS](#)'s reasoning for implementing the system in terms of its history, goals and overall aims, (2) a short overview of system stakeholders, (3) an in-depth technical description of the system's functionalities, variables, data sources and calculations, and (4) a description of how the system is operationalized in the context of the AMS caseworker-client consultation.

5.3.1 History, Goals and Aim

Prior to announcing its intentions to deploy the system in late 2018, the [AMS](#) had already attempted using various other profiling approaches as part of their unemployment services. Almost a decade before the planned rollout for [AMAS](#), pilot studies for profiling jobseekers into three categories were conducted in three AMS branches (Grießkirchen, Zell am See and Wien/Schlosshoferstrasse respectively) [\[BER_11\]](#). This initiative aligned with an international trend towards *profiling and targeting* as a reaction to the 2008 financial crisis. The pilot study was already conducted in cooperation with the *Synthesis Research GesmbH*, the same consulting firm responsible for the design and statistical implementation of the current [AMAS](#) system. Similar to the current system, the pilot study differentiated between short- and long-term prognoses, and used personal variables reminiscent of the current system's data points, including *age, gender, days of employment in the previous year* and *regional branch office* of the case. Contrary to the current system's definitive classifications, this initial profiling approach determined four rough "*orientational guidelines*" for further support: "*No measures necessary*", "*qualification measures can be supportive*", "*activation measures can be supportive*", and "*more intensive measures may be useful*" [\[PRÄS_5, p.3\]](#). Of those four guidelines, the first, second and fourth seem congruent with the current system's three categories; the group comprised of jobseekers requiring "*activation*" was not part of [AMAS](#) in its current iteration. Results of this pilot project included the importance of the system's "*accuracy*"¹⁶ (albeit not yet in the quantified sense of precision or error rates), and the assessment that profiling should, indeed, be considered as a probate tool to standardize internal processes through performance indicators [\[BER_15\]](#). This overall goal of standardization of internal processes also may be understood as the [AMS](#)'s response to latent criticism regarding systemic bias and discrimination among its caseworkers.

In the years following the 2008 global financial crisis, the [AMS](#) found itself challenged by stagnating yearly budgets, while unemployment rates continued to rise [\[HAND_1\]](#). In the introduction section of the internal handbook for [AMAS](#), the authors describe the necessity for a "*strategic reorientation*"¹⁷ in light of these circumstances: given the diminishing effectiveness of expensive "*active*" support measures, particularly for jobseekers with "*migration background*"¹⁸, limited additional funds for additional personnel barely able to

¹⁶Orig. "*Treffsicherheit*"

¹⁷Orig. "*strategische Neuausrichtung*"

¹⁸The term "*migration background*", as used by the [AMS](#), is in and of itself highly problematic (for a discussion of this terminology in the German geolinguistic context, see Will et al. [\[335\]](#)). Since a

retain the previous caseworker-to-case ratios, and the projections for continued increase in client numbers due, to improved access to the labour market for refugees led to the decision to implement a jobseeker profiling system (“Kundensegmentierung”) [HAND_1, p.4]. An internal strategic process was initiated involving various working groups tasked with model design and a new concept for low-threshold, external support measures. After the completion of this process, a public tender was initiated in December 2015, with the winning bid going to the same research company that had already collaborated on the pilot study 7 years prior. Although the system was slated to be deployed towards the end of 2016, the project was halted after “*political intervention*” according to the response to one of our inquiries [BRIEF_3, p.1]. Following the constitution of Austria’s new coalition government in 2017, the project was revived, implemented and rolled out to a subset of regional offices as a pilot test in late 2018, coinciding with the official announcement of the system in various news media and press releases.

After the initial plans for a roughly 6-month long test phase in selected AMS branch offices, the actual rollout of the system was delayed multiple times—due to undisclosed reasons—until the beginning of the global COVID-19 pandemic. As the onset of the pandemic necessitated a number of changes to the system and its operationalization [4, p.22], a new projected release date was set for January of 2021. Before this milestone was reached, a surprise ruling by the [DSB] prohibited the use of the system for the time being. As they stated in their reasoning [DSB], the legality of the use of personal data for the explicit goal of profiling was not sufficiently guaranteed by the Austrian implementation of the GDPR [336, 170] or by the laws governing the [AMS] itself [337]. Although the AMS appealed the ruling at the Austrian Federal Administrative Court and was granted an initial reprieve, the appeals process is ongoing as of 2022, and the system is not currently in use.

At the time of the announcement and subsequent deployment in the fall of 2018, a shift in expectations for the system’s goals and aims occurred. The original arguments for a statistical profiling system were founded on the stagnation of already scarce resources, the hopes for increased efficiency of the consultations, and improved effectiveness of the support measures granted to jobseekers.

detailed discussion would transcend the scope of this dissertation, the term will only be used in quotation marks going forward to denote its problematic nature and explicate it as a reference to [AMAS] internal nomenclature.

In 2019, the (at the time) *Ministry for Labour, Social Security, Health and Consumer Protection*¹⁹ stated—in line with the prevalent climate of austerity politics in Austria—that the use of the system should, as a concrete goal, make further *budget cuts* possible:

*“To support a rapid placement, the expansion of automation will be accelerated. In this context, personalized labour market measures will provide the support that will enable the AMS to offer quality counselling services despite a reduction in resources expenditure.”*²⁰

[BER_13, p.5]

Internally, the **AMS**'s reasoning for the system has remained the same: (1) increasing the *efficiency* of the consultation process, (2) maximising the *efficacy* of support measures granted and money spent on the unemployed, and (3) *reducing* the effects of *caseworker's personal bias* through the standardization of the consultation process. According to the AMS, the goal of a more efficient consultation process would be achieved through the support of the system, streamlining the assessment the caseworkers have to make by adding an additional, supposedly reliable source of information about the jobseekers' projected prospects. The classification provided by the system, taken together with the rule-based profiling of jobseekers into three categories would also increase the efficacy of measures granted, as most resources would only be spent on the group of jobseekers with *medium* chances, presuming that, statistically, this group would profit the most from these measures. Finally, as the **AMS**'s **CEO**, Johannes Kopf, argued repeatedly in public statements and interviews, the **AMAS** system would make more objective assessments than the caseworkers, who tended to “*assess the jobseeker's chances more pessimistically than the model*”²¹ [338].

It is worth noting that these stated goals, while consistently appearing in one form or another in the documents and public statements, are not clearly connected to the promised “evaluation” process the AMS claimed to have initiated in late 2018 and early 2019 by rolling out the system to select AMS locations throughout Austria. Since the AMS never disclosed a specific methodology for this evaluation, it is unclear if the system's success was ever measured against these stated goals or other criteria.

5.3.2 Stakeholder Analysis

Based on the description of the development process and the history of the system, the stakeholders of the system deserve a closer look, as their needs and interests significantly shaped the system's design and implementation.

¹⁹Orig. “*Ministerium für Arbeit, Soziales, Gesundheit und Konsumentenschutz*”

²⁰“*Zur Unterstützung einer raschen Vermittlung wird auch der Ausbau der Automatisierung forciert werden. Dabei wird die personalisierte Arbeitsmarktbetreuung jene Unterstützung liefern, die es dem AMS auch bei reduziertem Ressourceneinsatz ermöglichen wird, eine qualitative Beratungsleistung anzubieten.*”

²¹Orig. “[...] *unsere Beraterinnen und Berater die Chancen einer zukünftigen Jobaufnahme pessimistischer einschätzen, als es das Modell errechnete.*”

Two stakeholder groups directly interacting with the system stand out: the *AMS caseworkers* and the *jobseekers*. Analysing the system description, its goals and aims as provided to the AMS workers through the internal AMS handbook [HAND 1], and juxtaposing the public statements made by AMS CEO Johannes Kopf [338, 264], reveals a curious tension in *who the system is designed for*: On the one hand, the documents frame the system strictly as a support system for the AMS caseworkers; it should support them in making the—often quite difficult—assessment of a jobseeker’s chances to find a job, in addition to determining which measures would be most promising to help the jobseeker do so. On the other hand, this ultimate goal of helping jobseekers to find employment features prominently in public statements made by the AMS, at times even suggesting the system’s fundamental aim is to support jobseekers (through supporting the caseworkers assessment, and by providing the jobseekers themselves an indication of their chances on the labour market). With these competing and overlapping narratives, the purported and real needs of these two stakeholder groups become difficult to disentangle.

For the AMS workers, the official narrative presented by the AMS highlights the challenge of resource scarcity and the complexity of their assessment. On the one side, the caseworkers need to follow a complex set of procedures in processing their cases and making their assessment determining what, if any, type of assistance the jobseeker should be provided in order to maximise their chances to find a job. These procedures formalize the options available, and restrict what caseworkers can and cannot offer their “clients”. On the other side, the caseworkers have to contend with extremely limited resources in terms of time spent per jobseeker. Currently, the ratio of AMS workers to jobseekers can be as high as 1:250 according to the head of Austria’s Chamber of Labor [339]. In 2017, that ratio was slightly better, sitting between 1:110 and 1:150²², as an experimental study [340] conducted by the *Austrian Institute of Economic Research (WIFO)* shows. The study also delivered concrete statistical evidence that by improving this ratio—through hiring more AMS workers—the chances of jobseekers to find a job improved significantly; this effect was also described as cost effective due to the reduction of unemployment benefits paid to those jobseekers entering gainful employment faster. The decision to not pursue this strategy, i.e., to increase the time each caseworker has to spend with each of their assigned jobseekers, was thus a political, rather than an evidence-based, decision. Based on the approximate 1:100 ratio, the average time an AMS caseworker had for each consultation with a jobseeker was roughly 15 minutes. Concrete numbers for 2022 are difficult to track down, but the increase in jobseekers per caseworker to 250 suggests either a significant reduction in that time (on average), or a significant reduction in the number of times jobseekers and AMS caseworkers can meet. Relating these challenges—complexity of their tasks, and very high caseloads—to the system’s design calls into question whether caseworkers were substantially involved in the requirements analysis that shaped the system. Although the official documents state that multiple workshops with caseworkers were held to gather their input for the system, the participants’ agency to suggest alternatives to the profiling approach pursued by the

²²These fluctuation is related to seasonal variations in the number of jobseekers.

AMS may have been limited at best. It is thus highly questionable whether the system's purported goal to support caseworkers by making the consultation process more efficient was founded in the actual needs of this stakeholder group, as opposed to an overarching and politically informed strategy by AMS management.

For the jobseekers, there simply was no participation at all in the design or conceptualization process of the **AMAS** system. Whether or not any jobseeker would have expressed needs that could be fulfilled by the **AMAS** system (or any other, similarly designed profiling system) must thus be relegated to realm of speculation. It is worth noting, however, that jobseekers find themselves burdened by multiple impact factors when interacting with the AMS in general, and their assigned caseworker in particular. Beyond the obvious and well-documented psychological impact of unemployment in general [341, 342, 343], they also find themselves in a precarious hegemonial power relation to the AMS. Although the AMS is framed as a service institution in a supportive role to find employment, the caseworkers can mandate certain measures (including mandatory courses) or even sanction the jobseeker by halting unemployment benefit payouts should the caseworker deem the jobseeker's efforts a violation of the 'contractual' agreements between them. For specific and particularly vulnerable subpopulations, this precarious relation may be further complicated by language barriers and the challenges of navigating sometimes byzantine requirements posed by the bureaucratic processes of the AMS. Consequently, while it is certainly possible that some jobseekers could benefit from the assessment of their chances by the AMS algorithm as a "*second opinion*", as framed by the AMS, it is highly questionable if even a simple majority of jobseekers would regard the system as supportive in any way. For those categorized into the group with "*low chances*", the psychological impact of this assessment (regardless of its accuracy) may even lead to further stigmatization and de-motivation [343, 344], and thus may even negatively impact their chances of finding a job.

Compared to these primary stakeholders interacting with the system directly, the larger organisational stakeholders, namely the AMS management and board of directors, as well as the research company Synthesis Research GmbH subcontracted to develop the model, clearly had a much more significant role in shaping the system. The overall narrative of the system's goals and aim is clearly situated in the realm of organizational strategy rather than individual needs of either jobseeker or AMS caseworker. Competing strategic responsibilities of these bodies complicate the analysis of these roles. For instance, the overall design of the system, the modelling aspects, and even the pilot version rolled out in late 2018 was determined by AMS management and the project working group. In contrast, the *operationalization* of the resulting categorization (i.e., what measures and options would be available to jobseekers classified in the *low*, *medium* or *high* groups) was still unclear in 2019 and pending decisions by the board of directors [HAND_1, p.9]. Another example would be the requirements formulated towards the system as characterized by the subcontractor, which include the fact that the "[...] *variables used should be highly recognizable by the AMS caseworkers and clients, [...] particularly length of unemployment, age, health impairments, gender, responsibilities of care and highest*

*level of education attained*²³ [BEGL_1, p.10]. This requirement is only referenced in this specific document and seems to suggest that the recognizability of variables used in the system trumps other considerations (including other potential data sources, or alternative profiling approaches altogether); it was not possible to determine which of the other other stakeholders might have put forward this specific requirement. The only possibly related references are to be found in the official requirements document [PFLI_1, p.8], stating that *traceability*²⁴ of the results should be guaranteed; later in the document, this seems to refer more to the *ex-post explanation* texts the system should provide to account for the major factors impacting the classification. Assumptions on whether or not the reasoning provided by Synthesis Research GmbH was simply their own interpretation of these requirements remain speculative.

As these examples illustrate, the larger stakeholder groups held the primary definitorial power over the systems modelling, design and implementation. At the same time, tracing exactly where these decisions originate, and what motivated them, remains a difficult and sometimes seemingly impossible task due to the limited interpretability of conflicting reference documents. What the documents do show, however, is the complex, overlapping spheres of influence shaping this system, and the strong impacts the subcontractor's views had on the final result.

5.3.3 Technical Description

Given the complex and conflicting nature of information published by both the AMS and the Synthesis Research GmbH, and the public statements made by individuals representing these organisations, a coherent description of the technical inner workings of the [AMAS] system is as difficult to achieve as it is sorely needed. While our initial journal publication [3] presented this technical description as accurately as was possible at the time—given the limited availability of reliable sources—the in-depth analysis of the trove of documents we received only after this initial publication clarified certain aspects and led to the more accurate publication of our research report [4] in German. To discuss the system as a case study, the following section provides an up-to-date and accurate overview of these technical aspects, based on the latter publication (albeit in English). Discrepancies with the former publication are a direct result of the new information available after the receipt of the trove of internal documents listed in Appendix [A.3].

²³Orig. “Die Merkmale sollen einen hohen Wiedererkennungswert für die AMS-Beraterinnen/-Berater und ihre Kundinnen/ Kunden haben; [... d]azu gehören: Dauer der Arbeitslosigkeit im laufenden Geschäftsfall, Alter, gesundheitliche Belastungen, Geschlecht, besondere Betreuungsaufgaben im Haushaltsverband, Bildungsabschlüsse.”

²⁴Orig. “Nachvollziehbarkeit”

5.3.3.1 Predictive Variables

The **AMAS** system bases its profiling of jobseekers on a set of 13 core variables²⁵ or attributes as detailed in **[SPEZ_1]**. The variables can be roughly grouped into three distinct categories, namely

- personal and biographical information on the jobseeker (e.g., Age, Gender or Highest level of education),
- contextual variable meant to reflect the Regional labour market, and
- a set of 4 compound variables detailing the personal labour history of a jobseeker, including the number of Days of gainful employment within the four years prior, number of AMS cases within four one-year intervals, number of Cases longer than 180 days, as well as the number and quality of any Measures claimed by the jobseeker.

This data model uses only *discrete* values with hard thresholds for each variable. For some variables like Gender, Health impairments or Obligations of care, these discrete values only offer binary choices (e.g., “Male” or “Female”), while others can take up to four distinct, conditional values in escalating order. For instance, the variable measures claimed can take the values “0”, “Min. 1 supportive measure”, “Min. 1 educational measure”, or “Min. 1 subsidized employment measure”: a jobseeker having taken advantage of both a *supportive measure* and an *educational measure* would still only be assigned the superseding value of “Min. 1 educational measure”. For this example, the order of values is determined by escalating expenditures for each type of measure; for instance, subsidized employment is significantly more costly than educational measures. Table **5.1** lists all 13 core variables and their potential values.

The last variable in the list given in table **5.1** takes on a special role, as it clusters the jobseekers based on the duration of their current case of unemployment. The assignment happens at the beginning of each case and at the given tri-monthly intervals with each of the subsequent, regular consultations mandated by the AMS procedures.

The origins of this dataset on jobseekers are compound; basic biographical data is supplied by the **Austrian Federation of Social Insurances (AFSI)**²⁶, while contextual data and employment history stems directly from the AMS’s own databases. Base data from these datasets are normalized and transformed in the discrete values as outlined above before they are used as part of the **AMAS** system.

Upon closer inspection, a number of observations emerge from this set of predictive variables. First and foremost, a certain discrepancy in clarity and granularity exists: while some variables and their available values (e.g., Age or Citizenship) are fairly straightforward, others are vague and unclear, and feature only seemingly coarse potential

²⁵For clarity, variables referenced in these paragraphs will be formatted as sans-serif font (e.g., age group), and variable values will additionally be enclosed in quotation marks (e.g., “30-49”)

²⁶Orig. “Dachverband der Sozialversicherungsträger”

| Variable | Nominal Values |
|---|--|
| Gender | "Male" "Female" |
| Age group | "0-29" "30-49" "50+" |
| Citizenship | "Austria" "EU except Austria" "Non-EU" |
| Highest level of education | "Grade school" "Apprenticeship, vocational school" "High- or secondary school, university" |
| Health impairment | "Yes" "No" |
| Obligations of care (only women) | "Yes" "No" |
| Occupational group | "Production sector" "Service sector" |
| Regional labour market | 5 types of employment prospects specific to assigned AMS job centre: "Type 1" "Type 2" "Type 3" "Type 4" "Type 5" |
| Days of gainful employment within 4 years | "<75%" ">= 75%" |
| Cases within 4 one-year intervals | "0 cases" "1 case" "Min. 1 case in 2 intervals" "Min. 1 case in 3 or 4 intervals" |
| Cases with duration longer than 180 days | "0 cases" "Min. 1 case" |
| Measures claimed | "0" "Min. 1 supportive" "Min. 1 educational" "Min. 1 subsidized employment" |
| Duration of current unemployment | "Beginning of case" "3" "6" "9" "12" "15" "18" "21" "24" "30" "36" "48+ months" |

Table 5.1: List of all 13 variables and their potential values that are part of the statistical model.

values. **Health impairment**, for instance, is modelled only as a binary value, and was originally derived from [AFSI](#) base data, but later changed to refer to AMS internal data [\[PROT_ORG_11\]](#). A closer investigation of the applicable federal directive [\[RICHT_1\]](#) reveals a set of preconditions based on various Austrian federal laws regulating disability and equality definitions—as well as doctor's notes and certificates by accepted institutions—for declaring health impairments as hindering factors for certain jobs. The assignment of the **Health impairment** variable, however, seems to still depend on whether or not the AMS caseworkers deem the provided documentation acceptable as proof for limitations; the fact that certain health impairments (e.g., people using a wheelchair) may have significant impacts on jobseekers in one job sector, but little impact in another, is not taken into account, since the variable only allows a binary choice independent of the other variables. Similarly, the classification for **Occupational group** is extremely coarse; the myriad of different jobs falling under either "Production" or "Service sector" are condensed into these two categories only, ignoring any presumably significant variations in job opportunities within one or the other of these two sectors.

A second observation concerns the questions which types of data *were not* included in the

data set, as opposed to which were. Notably absent are any data points characterizing (potential) employers or a differentiated view of the regional labour market. The variable named *Regional labour market* itself is slightly deceiving in this regard, as it only represents a classification of the *performance* of the local AMS branch the jobseeker is assigned to. Specifically, each regional branch of the AMS gets assigned to one of five types based solely on the ratio of newly unemployed jobseekers vs. those having found a job at that location. The AMS's claim that this variable depicts an accurate characterization of the local job market is highly questionable, as (1) not all jobseekers in a given region choose to rely on unemployment benefits and thus are known to the AMS, resulting in a potential sampling bias, and (2) this variable completely ignores the fact that jobseekers might search for employment in a larger radius than the regional office would cover. This is particularly evident in the capital of Austria, Vienna, where various districts show different classifications for their local branches, but jobseekers assigned to one branch (based on the district they live in) are most likely looking for employment in neighbouring districts or even the surrounding countryside of Lower Austria as well. This obvious limitation makes the accuracy of this predictive variable highly questionable for specific, individual cases, particularly in the urban settings of Austria.

Besides these absent or misleading variables, other factors potentially influencing a jobseeker's chances to find employment are absent as well. Difficult to quantize variables, such as personal motivation, competencies and skills, ambition, or tolerance towards frustration are simply not modelled in the set of variables, although the predictive value of such factors is clearly known to the AMS. The final report [\[BER_5\]](#) for the pilot program for a newly developed supportive measure, the [Evaluation of Prospects Measure \(BBEP\)](#)²⁷, shows all of these personal factors (and more) as part of the evaluation. Since the [BBEP](#) remains one of the optional measures granted only to a subset of jobseekers—those with “*low chances*”—these factors are not immediately taken into account by the AMS for those jobseekers classified otherwise.

Finally, the nature of the majority of variables stands in a stark contrast to the *activation paradigm* informing the use of this profiling system. As explicated in Section [5.2](#), a core tenet of this paradigm is the *individualising* approach and its underlying assumption that jobseekers are somewhat responsible for their own (mis-)fortune on the labour market. The variables used to predict said (mis-)fortune, however, are by and large outside the sphere of influence for any individual, precluding any agency for change. Neither can a jobseeker (reasonably) influence their Age, Health impairments, the performance of the regional branch they are assigned to (*Regional labour market*), nor their Citizenship. Even for those variables that a jobseeker may, arguably, have an influence on (e.g., *highest level of education*), a jobseeker's lack of *belief in self-efficacy* [\[217\]](#) may preclude them from taking any action: since there is no way for a jobseeker to gauge the specific impact that furthering their education or attending job training would have on their chances as predicted by the system, they have little incentive to do so based on their score alone. In summary, to base the AMS's labour market policy on an individualist and

²⁷Orig. “*Perspektivencheck*”

personalised approach, while predicting jobseeker’s chances on that same labour market through largely predetermined, unchangeable attributes of jobseekers without giving them a reasonable path forward to actively change their situation seems paradoxical at best, and particularly cynical at worst.

5.3.3.2 Populations

To arrive at the final classification into one of the three groups with *high, medium or low* chances on the labour market, the **AMAS** system performs a series of clustering and calculation steps on the total population available in the dataset. The following sections outline this process step-by-step; an illustration of this process and its constituent parts can be found in the form of a flowchart in Figure 5.3.

Before any calculations, classifications or predictions occur, the dataset is sequestered into 4 distinct subpopulations, depending on the completeness of the available data. This process is linear and always follows the same order: First, all jobseekers for whom a *complete employment history during the 4 years prior* is available at the start of their current unemployment case are taken as the first group. The remaining population is considered as having *fragmented* employment histories if they (1) show more than 150 days of gaps in their employment history (as determined by their automatically deducted social security payment periods) and no data on education, being or having a dependent during this period to explain their lack of employment, or (2) have less than 1310 days of employment history overall available. The jobseekers of this subpopulation with incomplete data are further split up in the next step **[SPEZ_1][DAT_1]**. First, those jobseekers classified as having a “*migration background*” are grouped into a second subpopulation; “*migration background*” is determined either by a foreign citizenship or a recently acquired Austrian citizenship for the jobseeker themselves, or if both or one of their parents are foreign nationals. Within the system, this group is called “*partially valid assessable population with migration background*”²⁸. Of the remaining population, the third subpopulation includes those under 25 years of age as “*partially valid assessable youth population*”²⁹. The fourth and final subpopulation named “*partially valid assessable population with previously fragmented employment history*”³⁰ includes all remaining jobseekers; i.e., those who neither have a complete employment history, nor are classified as having a “*migration background*”, and are also older than 25 years. For all those groups with incomplete data, certain variables are not considered by the model for further processing, including Cases with duration longer than 180 days and Days of gainful employment within 4 years for those with a “*migration background*”, or Age and Citizenship for young people.

According to **[DOK_2]**, roughly 31% of jobseekers fall under one of the three subpopulations with only *partially valid assessable* data at the beginning of their unemployment case. While the special treatment of young people under 25 can be plausibly explained

²⁸Orig. “*partiell valide schätzbare Population mit Migrationshintergrund*”

²⁹Orig. “*partiell valide schätzbare jugendliche Population*”

³⁰Orig. “*partiell valide schätzbare Population mit zuvor fragmentierter Erwerbskarriere*”

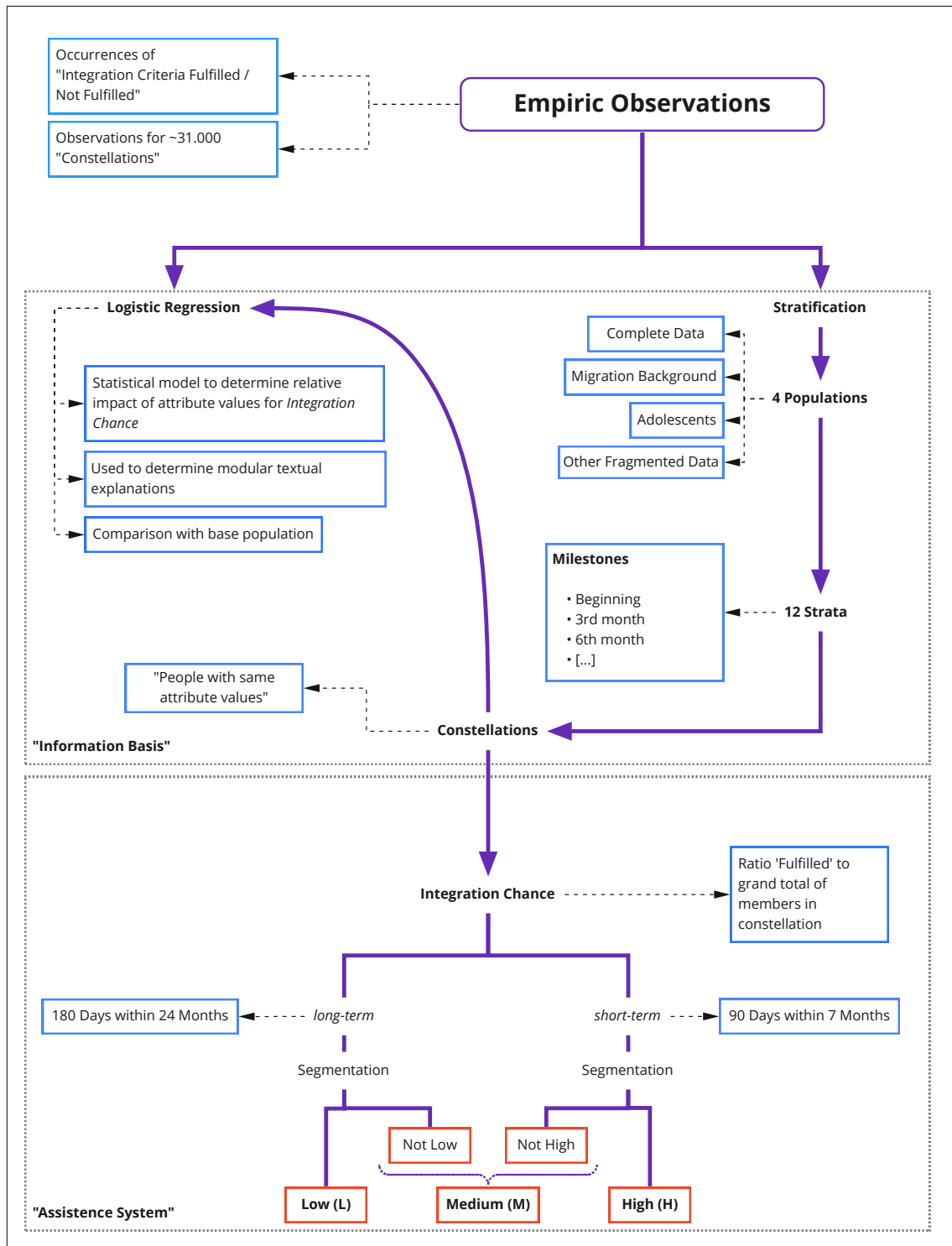


Figure 5.3: Flowchart illustrating **AMAS** constituent parts and profiling process.

due to the rule-based aspects of treatment including the *education guarantee* (see Section 5.2), the separate treatment of jobseekers with a “*migration background*” remains unexplained, as they should, by law, receive the same treatment and opportunities as Austrian jobseekers with fragmented employment histories. The specific order of execution for these sequestering steps is also curious, as jobseekers with incomplete histories who are both seen as having a “*migration background*” and are under 25 years of age are classified into the *migration* group first, instead of being treated as young people with the rest of Austria’s youth. As the *youth guarantee* applies to them as well, the only possible explanation remaining is an optimization of the system towards precision, although this line of reasoning remains speculative as of the publication of this dissertation.

5.3.3.3 Integration Chance, Criteria and Classification

The profiling process for each of these 4 subpopulations is based on the concept of “*constellations*”. Given a specific jobseeker’s set of variable assignments or attributes, all jobseekers of the previous 4 years that share their exact same attributes are considered to be part of the same *constellation*, and taken by the system as the basis for predicting said jobseekers’ Integration Chance (IC) value. As a simplified, yet illustrative example with a subset of variables, consider the following case: a female jobseeker with a German citizenship, 35 years of age, having completed high school as highest level of education, with no health impairments or obligations of care would be compared to all previous jobseekers with the same variable values, i.e., “Female”, “30-49 years” of age, “EU except Austria”, “High- or secondary school, university”, and neither Health impairments nor Obligations of care for dependents.

To determine such a jobseekers’ IC value, and subsequently classify her into one of the three categories based on that IC value, two *Integration Criteria* are evaluated for each member of her *constellation*. The first, short-term criterium is considered as *fulfilled* by a jobseeker in this constellation if they achieved at least *90 days of employment* within the first *7 months* after the start of their case (i.e., after they first became unemployed and registered their case with the AMS). The second, *long-term* criteria considers a period of *180 days of employment* within the first *24 months* of unemployment as a successful integration into the labour market.

In mathematical terms, the calculation of IC scores then is defined as follows: Let S_C be the sum of members in a constellation C defined as $S_C = N_s + N_{-s}$, with N_s as the number of members that fulfilled the short-term integration criterium, versus N_{-s} as those that did not. The Integration Chance IC_{short} of a new member of C is thus calculated as

$$IC_{short} = \frac{N_s}{S_C} \quad (5.1)$$

For the long-term integration chance IC_{long} and N_l as the number of members of C that fulfilled the long-term integration criterium, the calculation is equivalent:

$$IC_{long} = \frac{N_l}{S_C} \quad (5.2)$$

In other words, if a given jobseeker's *constellation* consists of a total of 50 observations of jobseekers with the same set of variables in the previous 4 years, of which only 12 have fulfilled the long-term integration criterium, the current jobseeker's long-term integration chance IC_{long} is calculated as the following ratio:

$$IC_{long} = \frac{12}{50} = 24\% \quad (5.3)$$

For any given jobseeker at the beginning of their case, both IC_{short} and IC_{long} are calculated based on their comparative constellation \mathbb{C} . Depending on these resulting IC values, the jobseeker is classified into Q_{high} as having *high chances* of (re-)integration into the labour market if their short-term IC value is greater than or equal to 66%. If their long-term IC value is less than or equal to 25%, they are classified as Q_{low} having *low chances*; otherwise, they are assigned to the group Q_{medium} with *medium chances* by default.

$$Q = \begin{cases} Q_{high} & \text{if } IC_{short} \geq 66\% \\ Q_{low} & \text{if } IC_{long} \leq 25\% \\ Q_{medium} & \text{if } IC_{short} < 66 \wedge IC_{long} > 25\% \end{cases}$$

Considering the number of variables and their potential values, a hypothetical number of roughly 81.000 constellations could exist, if all possible combinations of variable values were to occur. In reality, the much smaller number of roughly 31.000 actual constellations exist, as [DOK_2] explicates. However, given the limited number of observations and the fact that the different variations are not equally distributed over all possible observations, it comes as no surprise that only about 7.800 constellations contain more than 10 members. For a threshold of 50 observations, that number shrinks down even further to roughly 1.900 constellations. From a statistical standpoint, these observation numbers are miniscule compared to the total population of roughly 504.000 jobseekers interacting with the AMS on yearly basis. The developers of the system, the Synthesis Research GmbH, are evidently fully aware of this fact, but—statistical evidence notwithstanding and without further explanation—declare 50 or more observations “(statistically) extraordinarily satisfying”³¹, and 10 or more observations as “sufficient”³² [DOK_2, p.13]. In doing so, however, they implicitly concede a less-than-satisfactory statistical validity of predictions for for the 39% or 195.000 jobseekers per year whose constellations only contain 50 or fewer members as comparison. Figure 5.4 illustrates these relative numbers. To address the even more pressing issue of the roughly 12% of

³¹Orig. “(statistisch) außerordentlich befriedigend”

³²Orig. “ausreichend”

jobseekers for whom only 10 or fewer comparative observations are available in their constellation, the system merges adjacent constellations following an undisclosed logic. As requests for further information remained unanswered, it is still unclear as to whether this merging occurs automatically—e.g., based on certain measures of neighbourhood—or as an *a priori* and manual process.

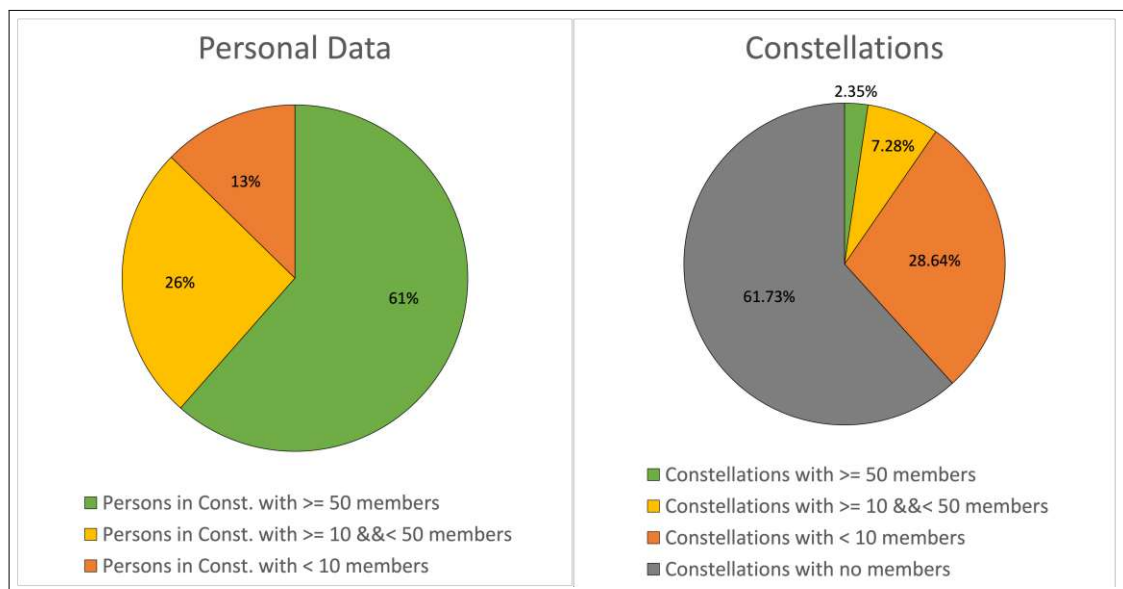


Figure 5.4: Personal data vs. constellations and members. [DOK_2, p.13]

Looking back at Figure 5.3, the pathway originating after the creation of *constellations* towards the statistical analysis through *logistic regression* remains to be discussed. The documentation originally published by the Synthesis Research GmbH on behalf of the AMS [DOK_1] does not mention the specific calculation of the IC value as described above, but seemingly suggests the IC value is solely the result of the coefficients exemplified in Figure 5.1, determining the relative impact of various variable values on a given jobseeker's chances, relative to a predefined "*base group*": young Austrian men with only compulsory education (grade school), no obligations of care or health impairments, working in the service sector, with no previous gaps in employment and at least 1028 days of employment in the four years prior [DOK_1, p.11]. This normative definition of the base group itself is quite telling and should be considered value-laden as well, but was presented in the documents as a purely statistical consequence. While such a method could yield similar, if not equivalent results for the classification process (e.g., by applying the coefficients directly to the jobseekers in question to determine IC_{short} and IC_{long}), we later learned through the interviews with representatives of the AMS and Synthesis Research GmbH, as well as through the additional documents we received, that this approach based on logistic regression was not used in the concrete calculation of IC values at all. Consequently, our initial analysis of this specific, technical aspect of the system as described in [3] was factually incorrect, and corrected in the subsequent

publication of our research report in [4]. This does not mean, however, that such a regression approach is not utilized in the system at all. On the contrary, the *textual explanations* functionality of the tool implicitly suggests that the relative impacts of these variable values played a role in the creation and assignment of these explanations, and was used to determine which jobseekers receive which of the modular text fragments explaining their IC value calculated as outlined above. A more detailed description of this explanation functionality can be found in the following Section 5.3.4.

5.3.4 Operationalization in the Context of the AMS

As previously established in this dissertation, an algorithmic system’s practical operationalization influences both its success and critical issues arising from its context of use. In the case of AMAS, this becomes particularly evident due to the specific requirements the AMS caseworkers face when interacting with the system and the jobseekers.

The interaction between AMS caseworker and jobseekers as determined by the federal directive KP-1 [RICHT_1], which governs the entire process, from first contact to the closing of the case. This process is roughly categorized into four phases: (1) first contact, (2) clarification, agreement and matching, (3) continued support and securing subsistence of livelihood, and (4) conclusion. These phases can stretch over multiple in-person (or since the advent of the global COVID-19 pandemic also online) meetings. Of those four phases, AMAS impacts phases two and three the most. After the initial collection or validation of personal data in phase one, the IC values and subsequent categorization into one of the three groups with *low*, *medium* or *high* chances are an integral part of phase two, and directly contribute to the assessment of whether the jobseeker can immediately start looking for a job, or if they will be offered additional support measures. Depending on this determination, an agreement is signed by both jobseeker and AMS caseworker detailing the future efforts required of the jobseeker (e.g., “*x number of job applications sent within y weeks*” or “*attends a training course*”). Throughout subsequent interactions, the jobseeker’s prospects are re-evaluated via the AMAS system periodically to reflect potential changes over time.

The outputs generated by the system are integrated into the AMS’s own software suite used by the caseworkers to collect jobseekers’ personal data and document their case. A sample form generated by the system showing the jobseekers IC values and categorizations is depicted in Figure 5.5. As a caseworker evaluates the jobseeker’s case, they open this form via a menu entry titled “*Labour market chances*”³³, and are presented with both the the short-term and long-term IC values (top right), the resulting automated classification (top left), and a section for the caseworker’s manual classification. Within the system, two classifications are shown: the “*Computer-Supported Labour Market Chance (CAM)*”³⁴ determined by AMAS, and the “*Labour Market Chance as assessed by Caseworker (BAM)*”³⁵.

³³Orig. “*Arbeitsmarktchance*”

³⁴Orig. “*Computergestützte Arbeitsmarktchancen*”

³⁵Orig. “[...] von BeraterInnen erfolgte Einschätzung der Arbeitsmarktchance” [HAND_1, p.10]

The screenshot displays a web-based form titled "Arbeitsmarktchance". It is divided into several sections:

- Computergestützte Arbeitsmarktchance:** Contains a dropdown menu for "CAM" (set to "CAMN"), a date field "erstellt" (09.10.2019), and a code field "S001". To the right, under "IC Values", are two input fields: "Arbeitsmarktchance in %" with "kurzfristige" (13) and "langfristige" (10).
- Beraterinnen-Arbeitsmarktchance:** Contains a dropdown menu for "BAM" (set to "BAMM"), a date field "erstellt" (30.04.2019), a code field "G894", and fields for "bis" and "geändert". Below this is a text area for "Begründung" containing the text: "Frau [REDACTED] ist jung, gesund und hat keine Betreuungspflichten. Sie sucht in der Gastronomie und hat eine große Auswahl an freien Stellen."
- Protokoll:** A table with headers "Code", "ab", "bis", "Ben.", and "Begründung". The table body is currently empty.

At the bottom of the form are buttons for "Ändern", "CAM übernehmen", "Segmentzusatzinformation anzeigen", "Schließen", and "Hilfe".

Figure 5.5: Annotated sample form showing jobseeker's IC scores and classifications [SCHU_7, p.38]

The **CAM** is represented through four possible values in the top right text field as either (1) "CAMH" (*high chances*), (2) "CAMM" (*medium chances*), (3) "CAML" (*low chances*), or (4) "CAMU" (incomplete data, no automated classification possible). Based on this **CAM** classification, caseworkers must, mandatorily, make their final assessment of the jobseeker's chances and categorize them by assigned the **BAM** with one of three possible values—"BAMH", "BAMM" or "BAML"—equivalent to the first three of the four **CAM** values described above. Depending on their assessment, caseworkers can immediately accept the system's classification *as is*, adopting the **CAM** suggested by the system as their **BAM** by clicking the button labelled "*CAM übernehmen*". Should they disagree with the system's assessment, they can manually adjust the classification and overwrite the system's **CAM** with a new **BAM**. If—and only if—they choose to assign the jobseeker a **BAM** differing from the algorithmically generated **CAM**, they must provide a mandatory written justification for this change in the text area shown underneath the **BAM** text field. The example in Figure 5.5 shows such a change: here, the caseworker has adjusted the jobseeker's classification from "CAMN" (*low chances*) to "BAMM" (*medium chances*), and

provided the justification that “*Ms. [redacted] is young, healthy and has no obligations of care. She is looking for a job in gastronomy and can choose from many open positions.*”.

The result of this process is the final classification, either the automatically generated classification generated by **AMAS** and accepted by the caseworker, or the manual classification and justification provided by the caseworker overriding the automated assessment. As defined by the federal KP-1 directive **[RICHT_1]**, the automated classification should be taken as the starting point for a reflective discussion between caseworker and jobseeker to cooperatively evaluate their chances and arrive at a final classification. Differences between the jobseeker’s, caseworker’s and automated assessments should be either resolved through this discussion, or must be documented in the jobseeker’s case file, if no agreement can be reached. The AMS, both in its public communication and in the KP-1 directive, insists that only the caseworker makes the final assessment, and that the automated **CAM** is only supportive in nature.

A jobseeker’s classification has many procedural implications for their further interaction with the AMS and the potential resources available to them. This is reflected throughout the KP-1 directive **[RICHT_1]** by the different instructions and rules that must be applied to jobseekers with (presumably) *high*, *medium* or *low* prospects. Based on this classification, the AMS considers those with high chances as “*service clients*”³⁶, those with medium chances as “*consultation clients*”³⁷, and those with low chances as “*support clients*”³⁸. Table 5.2 provides an overview of these classifications, groups, and rules and limitations applied to each group.

The AMS considers “*service clients*” as requiring only minimal support and controlling, and assumes they will find new employment mostly on their own accord. Those jobseekers who already have confirmation of a new job must be automatically assigned to this group independent of their **CAM** classification, and young people under the age of 18 cannot be assigned to this group under any circumstances **[RICHT_1, p.13]**. “*Service clients*” are only required sporadic contact with their caseworker (at least once every 2 months), and their profile is only matched with potential job offers through the AMS’s own job platform once a month **[RICHT_1, p.87]**. Finally, only a very limited subset of the educational measures offered by the AMS are available to them in special cases.

Jobseekers in the second group with projected *medium* prospects are considered “*consultation clients*” by the AMS. For them, a much more intensive frequency of meetings—at least on a monthly basis, but bi-weekly as a suggestion—is prescribed by the KP-1 directive **[RICHT_1, p.88]**. This group also automatically includes all young people under 18 years of age, as well as those young people between 18 and 25 years of age that would

³⁶Orig. “*Servicekundinnen/-kunden*”

³⁷Orig. “*Beratungskundinnen/-kunden*”

³⁸Orig. “*Betreuungskundinnen/-kunden*”

³⁹If a caseworker thinks a different measure is necessary, even though the jobseeker is classified as having *low* chances, the caseworkers *must* re-classify the jobseeker via their **BAM** into the group with *medium* chances, regardless of whether or not they assess their chances as *medium*, before they can grant them this measure.

| Classification | AMS Label | Preconditions and Limitations |
|-----------------------|------------------------|---|
| <i>High Chances</i> | “Service Clients” | <ul style="list-style-type: none"> • All jobseekers with confirmed job offers • No jobseekers under the age of 18 • Minimum interval of one contact every two months • Limited educational measures available |
| <i>Medium Chances</i> | “Consultation Clients” | <ul style="list-style-type: none"> • All jobseekers under the age of 18 • All jobseekers between the ages of 18 and 25 if assigned to “CAMN” (<i>low</i>) • No jobseekers with confirmed job offers • Minimum interval of one contact per month, suggested twice per week • All measures potentially available |
| <i>Low Chances</i> | “Support Clients” | <ul style="list-style-type: none"> • No jobseekers under the age of 25 • No jobseekers with confirmed job offers • Minimum interval of one contact per year • Offered BBEP as voluntary measure • Offered referral to External Counselling and Support Institutions (BBE) as voluntary measure • No other measures available³⁹ |

Table 5.2: List of AMAS classifications, corresponding client group labels, and preconditions/limitations for group assignment and measures granted.

have been assigned to the *low* segment by the automated classification. Jobseekers with a confirmed job offer are excluded from this group, as they are automatically assigned to the . In accordance with the AMS’s general strategies towards resource allocation, this is the group it intends to spend to the most on, with all levels of supportive, educational or subsidized employment measures potentially available to them.

The group with the *lowest* projected chances are labelled “*support clients*”. This group also excludes unemployed people with a confirmed job offer and those under 25 years of age, and the frequency of (mandatory) consultations are significantly reduced up to

a minimum of once per year under certain circumstances. Jobseekers categorized into this group are also offered a special type of evaluation, the **BBEP**⁴⁰. This evaluation involves a much more detailed look at the jobseeker's chances, including health and efficiency, competencies, as well as personal values and motivation [BER 5, p.5]. Finally, these "support clients" are being referred to special types of institutions called **External Counselling and Support Institutionss (BBEs)**⁴¹. These institutions follow the *activation paradigm* insofar as they offer low-threshold support to jobseekers in order to provide "personal stabilization and support to cope with everyday life"⁴², "strengthening and activation of their potential for self-help"⁴³ and "support the transition towards adequate social security systems"⁴⁴.

The internal power of these classification assignments is particularly evident in light of the specific rules for when caseworkers are supposed to override and adapt the automated classification. Although the final decision for any offers made to the jobseeker in terms of supportive, educational, or even subsidization measures lies with the caseworker, they nonetheless need to override the automated assessment to 'fit' the jobseeker in the appropriate category before assigning the measure. For instance, a jobseeker classified as having *low* chances could still be offered certain measures only available to the *medium* group, but only if the caseworker reclassifies them as a member of the "BAMM" group. Even after doing so, however, the original automated assessment remains in the system as a "CAMN" marker and is shown alongside the overwritten "BAMM" marker until a re-evaluation by the **AMAS** system is triggered [**HAND_1**].

5.3.4.1 Explanations

In addition to numeric and categorical outputs of AMAS, caseworkers can open an additional form to show textual explanations for these scores via the button "*Show additional segment information*"⁴⁵. The resulting form is depicted in Figure 5.6. The example shown here provides both current and past explanations (in case the classification changed over time), and is automatically generated from a set of explanatory text fragments. Caseworkers are encouraged to refer to these explanations when jobseekers ask for a justification of their score or classification.

Explanation texts provided by **AMAS** in this interface follow a set of rules and conditions dependent on the presence and values of variables for the jobseeker in question. Whether or not explanations are provided depends on two factors: First, only jobseekers in the population with a *complete employment history* as outlined in Section 5.3.3.2 will have explanations provided, leaving the other three populations with incomplete data (either those with "migration background", *young people* or *others with incomplete data*) guessing

⁴⁰ Orig. "Perspektivencheck"

⁴¹ Orig. "Beratungs- und Betreuungseinrichtungen"

⁴² Orig. "persönlichen Stabilisierung und Unterstützung bei der Alltagsbewältigung"

⁴³ Orig. "Stärkung und Aktivierung des Selbsthilfepotenzials"

⁴⁴ Orig. "Unterstützung beim Übergang in das adäquate Sozial- und Versorgungssystem."

⁴⁵ Orig. "Segmentzusatzinformation anzeigen"

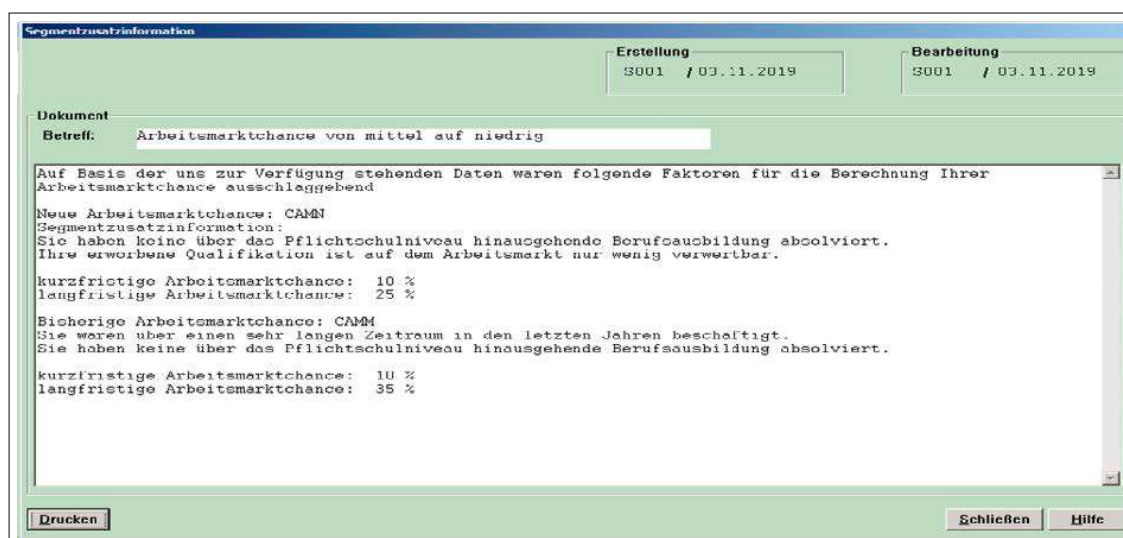


Figure 5.6: Explanations for the IC score and classification provided by AMAS [SCHU_7, p.45]

as to the reasons for their score and classification. Secondly, the explanations generated by the system are only available for those in the *high* or *low* segments, but not for the majority of jobseekers classified with *medium* prospects.

Even for the remaining jobseekers eligible to receive these explanations, the available explanations are limited in detail and somewhat generic. The table in Appendix A.4 lists these texts and conditions in detail, and shows a total of four possible text fragments for those with high chances, and seven fragments for those with low chances. Among the former, *high continuity in employment* or *prior experience with job applications* are noted as beneficial for a jobseekers chances, while *deficits in competency*, *health impairments* or *advanced age* are given as “*particularly challenging*” for the latter. The assignment text fragments to jobseekers is solely depending on the conditions show in Appendix A.4, which implies that the actual IC value for a given jobseeker has no impact on the explanations themselves. Not all of these conditions are immediately plausible, either: for instance, the *advanced age* text fragment is only shown to those jobseekers classified as *low chances* if they are both over the age of 50 and also have at least attained the educational levels of an apprenticeship, vocational school or higher (up to and including university graduates). Similarly, the text fragments suggesting a lack of education—“*deficits in competency*” and “*no vocational training*”—share their first condition (only compulsory/grade school education), but differ in their second condition, which hinges on the jobseeker’s citizenship. While jobseekers with a non-EU citizenship are told that they “[...] *only completed [their] mandatory education and/or have limited German language skills, which makes [their] job search harder.*”, Austrian and EU citizens are told that they “[...] *have no additional vocational training beyond mandatory school.*” The fact that neither the conditions for these two explanatory text fragments in particular,

nor the [AMAS](#) system in general includes any information about language proficiencies at all makes this discrepancy misleading at best, and carrying presumptuously racist undertones at worst. Finally, while the rules governing the assignments of these texts are accessible to the caseworkers for review in [\[HAND_1, p.18\]](#), it is questionable whether or not they will be able to relay these complex rules and factors⁴⁶ to the jobseekers in the very limited time they can spend with them.

5.4 Critical Issues: Bias, Discrimination, Transparency and Accountability

The [AMAS](#) system is fraught with a number of critical issues, many of which have already been touched upon in the previous sections. As our previously published research [\[3, 4\]](#) focused mostly on issues of bias and discrimination, this chapter will cover these problematic aspects only in a cursory manner. For the core topics of this dissertation, transparency and accountability, this section expands on the previous work published *ibidem*.

5.4.1 Technical, Pre-Existing and Emergent Bias

Applying Friedman and Nissenbaum's [\[102\]](#) framework for bias in computer systems⁴⁷ to the [AMAS](#) system reveals a variety of (socio-)technical, pre-existing and emergent biases embedded within the system. Many of these issues are a direct consequence of design decisions and trade-offs related to the data sources and model variables. Like any algorithmic system, [AMAS](#) abstracts the complex reality of jobseekers and the labour market into a simplified data model that can never capture all aspects of that reality [\[345\]](#). As explicated in Section [2.2.1](#), these decisions are never solely *technical* in nature, but rather a consequence of socio-technical circumstances and value-laden decisions.

In the case of AMAS, a number of factors influenced the choice of variables. According to [\[BEGL_1\]](#), the AMS was limited insofar as data points needed to be available throughout Austria in a standardised form, as well as cover a range of previous years. Localized factors, although possibly highly relevant for certain areas of Austria, were thus not considered, and no efforts were made to collect additional data points not already available in the AMS data warehouse. This choice is remarkable insofar as the original pilot study performed by Synthesis Research GmbH for the AMS in 2009 [\[BER_11\]](#) would have left ample time for an in-depth evaluation—and subsequent adaptation—of data collection practices to include additional data sources; for instance in terms of evaluating the impact of supportive measures granted to jobseekers. The stated requirement that the model variables needed to be recognizable by both caseworkers and jobseekers (see Section [5.3.1](#)) further limited the selection of data points. Finally, an AMS senior consultant and

⁴⁶There are a number of exceptions influencing *which* texts are shown *when* in addition to the ones described in Appendix [A.4](#) which were omitted here as they add little to the substance of this analysis and are exceedingly byzantine in their complexity.

⁴⁷For a more detailed description, see Section [2.2.1](#).

manager working on the project described the considerations influencing the choice of variables as including “*quality, [...] validity, [...] empirically proven relevance for the dynamics of the labour market [and] ethical considerations*”⁴⁸ [BRIEF_3, p.4]. While the definition of quality and validity in this case remain vague, he explicates the “*ethical considerations*” as having purposefully chosen not to include data on *marital status, previous sanctions* (i.e., sanctions levied by the AMS for non-compliance with the agreements made between caseworker and jobseeker) or a *previous foreign citizenship* (i.e., recently naturalized Austrian citizens). This last claim regarding foreign citizenships can be easily disproved through the model description of what constitutes a “*migration background*”, as this specifically includes categories for those jobseekers with either a previous foreign citizenship, or as having at least one parent with a foreign citizenship. In addition, this claim of ethical considerations playing a role in the modelling process implicitly means that the variables that *were* chosen in the end were considered to be ethical, including the variables on *Obligations of care*⁴⁹ and *Health impairments*. These tensions and contradictions illustrate the complex interplay between the various social and technical impact factors leading to the final model, and underscore that technical aspects of a system are never determined by purely objective factors.

5.4.1.1 (Socio-)Technical Bias

A consequence of these socio-technical decisions are (socio-)technical biases embedded within the system. In Section 5.3.3.1, I already discussed the lack of *granularity* and *clarity* of variables such as *Age, Health impairments, Regional labour market* and *Occupational group*, as well as the absence of variables reliably characterizing the (local) labour market, potential employers, or hard-to-model qualitative data such as personal motivation or skills and competencies. Taken together, these model limitations contradict one of the implicit, foundational assumptions underlying the model: that a *constellation* of jobseekers sharing the same attributes as modelled by the system are a *homogenous* group whose observed prospects of fulfilling the short- and long-term integration criteria can be directly and uniformly applied to (future) members of the same constellation. The arguments made by both the AMS and Synthesis Research GmbH to support this claim rest almost entirely on the quantitative evaluation of the *accuracy* of the system and their claims of an average *precision* of roughly 85%. In other words: they argue that, were the constellations not homogenous, the system could not reach these supposedly high levels of accuracy, leading to a higher rate of misclassifications. The validity of these claims itself is debatable in light of the questionable nature of their claims towards the *precision* of the system (for a more detailed look at the system’s purported accuracy, see

⁴⁸Orig. “*Qualität der Daten und deren Validität, [...] für das Arbeitsmarktgeschehen empirisch bestätigt hohe Relevanz [..., und] nach ethischen arbeitsmarktbezogenen Gesichtspunkten vertretbar*”

⁴⁹The *Obligations of care* variable is a particularly controversial example, as contradicting information was published on whether or not this variable would *only* be applied to women. The initial documents released by the Synthesis Research GmbH suggested so and claimed that the regression models showed that this variable had negligible predictive value for men, but AMS representatives refuted these claims in later statements, leaving the exact meaning and use of this variable unclear.

the following Section [5.4.2](#)). On an individual level, however, the model itself provides counterexamples contradicting these claims regardless of the overall precision of the system. Coarse variables like the Occupational group or Health impairments mean that two jobseekers with identical variable assignments, but with differing previous careers and health impairments would be considered part of the same constellation. Considering that, for instance, a trained plumber and a software engineer that both require the use of a wheelchair are considered as equal in terms of the system's modelling, this assumption seems highly questionable. Similarly, the hard thresholds and discretization for variables like the Age groups mean that the system considers a 30-year-old and a 49-year-old jobseeker as more comparable in terms of their job prospects than a 49-year-old and a 51-year-old. Finally, since these constellations are applied equally across the board for all jobseekers, systematic biases against certain subgroups are an unavoidable consequence of these modelling decisions.

Considering these issues, a naive approach to mitigate the coarseness of variables would be to raise the granularity of data, creating more differentiated constellations. As the discussion of data coverage and constellation membership numbers in Section [5.3.3.3](#) shows, this approach is likely infeasible: if, in the present system, roughly 39% of jobseekers are assigned to a constellation with fewer than 50 members, a further differentiation of variables or introduction of additional variables would reduce these member numbers even more. For the roughly 13% of jobseekers currently in constellations with fewer than 10 members, this could even lead to constellations with no observations at all, rendering the predictive value of the model non-existent for these observations. Consequentially, any claims towards statistical validity of predictions with such low numbers of observations per constellation would be highly questionable at best.

5.4.1.2 Pre-existing Bias

Throughout the controversial public discussion of the [AMAS](#) system after its initial announcement, the most frequently stated defensive argument as to why the system was *not* discriminatory against certain jobseekers was that it simply represented existing inequalities of the labour market as a reflection of reality: in other words, the system is not discriminatory, but reality is, and [AMAS](#) just reflects this reality [\[BER_14\]](#), [\[NOTES_10\]](#). While a closer look into discrimination as a result of the use of the system is discussed in Section [5.4.2](#), what this argument implicitly concedes is that pre-existing societal biases are, indeed, present in the system by design. The fact that jobseekers assigned to different constellations, depending on personal attributes that are clearly linked to discrimination in the larger societal context of Austria, are profiled differently, is thus an intended functionality, and not an unwelcome side-effect of the system. This intentional replication of societal bias notwithstanding, whether or not the system *accurately* and *proportionally* reflects the reality of individual jobseekers and their prospects is highly questionable. Structural inequalities cannot be causally linked to either personal attributes of jobseekers or other contextual impact factors, such as whether or not a given labour market with a high percentage of workers sharing these attributes simply has no open positions at

a given time. While both of these may result in lower chances as projected by [AMAS](#), determining the difference is impossible due to the coarseness of the model.

| | Zielfunktion 1 (3 in 7) | | Zielfunktion 3 (6 in 24) | | Anteil Geschäftsfälle | | Trefferquote ⁵ im Bereich | |
|--|----------------------------|----------------|-----------------------------|----------------|--------------------------|------------------|---|------------|
| | Zahl GF ¹ | M ² | Zahl GF ¹ | M ² | »N« ³ | »H« ⁴ | »N« | »H« |
| Alle Populationen zusammengefasst | 617.390 | 45% | 627.467 | 71% | 5% | 26% | 84% | 81% |
| Männer | 368.710 | 50% | 374.991 | 75% | 3% | 31% | 84% | 81% |
| Frauen | 248.680 | 38% | 252.476 | 65% | 6% | 18% | 83% | 80% |
| Alter | | | | | | | | |
| bis 29 Jahre | 232.748 | 45% | 239.166 | 72% | 3% | 21% | 82% | 78% |
| 30 bis 49 Jahre | 273.831 | 48% | 278.252 | 74% | 2% | 30% | 82% | 82% |
| 50 Jahre u. älter | 110.811 | 33% | 110.049 | 53% | 15% | 25% | 84% | 83% |

Figure 5.7: Excerpt of documentation table listing profiling results for all populations differentiated by gender and age [\[BER_7, p.21\]](#)

Conversely, what *can* be determined from the documents is the fact that particularly marginalised groups, as represented in the model by attributes such as Gender, Age, Health impairments or Citizenship face socio-economic hardship, and may be affected by disproportional, structural and cumulative disadvantage [\[91\]](#). People with disabilities are historically disadvantaged in their access to education and public services in Austria, which is also reflected in their prospects on the labour market and modelled by the system's Health impairment and Highest level of education variables. Similarly, intersectional disadvantages exist for women of non-EU citizenship compared to men of the same group [\[346, 347\]](#). Finally, the pre-emptive sequestering of sub-populations by complete vs. incomplete employment histories, as well as the additional differentiation within the group with incomplete data in “*migration background*”, young people and all others, reflects further pre-existing biases manifest in [AMAS](#).

Figures [5.7](#), [5.8](#) and [5.9](#) illustrate these discrepancies for all populations, the population with complete and valid data, and the population with incomplete employment histories due to “*migration background*” respectively. For instance, women are shown to be twice as likely to be classified into the group with *low* chances in Figure [5.7](#) (3% vs. 6%). The overall percentage of people projected to have *high* chances drops from 35% for the population with fully available data (Figure [5.8](#)) to 26% for all populations (Figure [5.7](#)), reflecting the lower average chances for the populations with incomplete data.

Not all pre-existing biases in the system are directly represented through the variables: some variables also have the potential to serve as *proxy variables* correlating disadvantages with seemingly disconnected attributes. The assignment of the Regional labour market variable, for instance, as the relative performance of a given regional AMS branch location depends solely on the jobseekers address. There are stark discrepancies in AMS branch performance even within the city of Vienna: While the AMS branch in the district

| | Zielfunktion 1 (3 in 7) | | Zielfunktion 3 (6 in 24) | | Anteil Geschäftsfälle | | Trefferquote ⁵ im Bereich | |
|--|----------------------------|----------------|-----------------------------|----------------|--------------------------|------------------|---|------------|
| | Zahl GF ¹ | M ² | Zahl GF ¹ | M ² | »N« ³ | »H« ⁴ | »N« | »H« |
| Basispopulation (voll valide beobachtbar) | 438.724 | 53% | 442.856 | 79% | 4% | 35% | 84% | 82% |
| Männer | 268.413 | 60% | 271.638 | 83% | 3% | 43% | 85% | 82% |
| Frauen | 170.311 | 42% | 171.218 | 70% | 5% | 23% | 84% | 82% |
| Alter | | | | | | | | |
| bis 29 Jahre | 116.854 | 59% | 119.780 | 84% | 1% | 40% | 82% | 79% |
| 30 bis 49 Jahre | 218.232 | 53% | 220.167 | 79% | 1% | 36% | 82% | 82% |
| 50 Jahre u. älter | 103.638 | 36% | 102.909 | 55% | 13% | 29% | 85% | 83% |

Figure 5.8: Excerpt of documentation table listing profiling results for populations with complete data, differentiated by gender and age [BER_7, p.22]

| | Zielfunktion 1 (3 in 7) | | Zielfunktion 3 (6 in 24) | | Anteil Geschäftsfälle | | Trefferquote ⁵ im Bereich | |
|------------------------------------|----------------------------|----------------|-----------------------------|----------------|--------------------------|------------------|---|------------|
| | Zahl GF ¹ | M ² | Zahl GF ¹ | M ² | »N« ³ | »H« ⁴ | »N« | »H« |
| Migrationshinter- grund | 102.452 | 29% | 104.596 | 55% | 10% | 10% | 82% | 75% |
| Männer | 60.207 | 34% | 61.269 | 59% | 7% | 12% | 79% | 74% |
| Frauen | 42.245 | 24% | 43.327 | 50% | 16% | 8% | 83% | 77% |
| Alter | | | | | | | | |
| bis 29 Jahre | 57.980 | 26% | 58.674 | 52% | 10% | 6% | 81% | 75% |
| 30 bis 49 Jahre | 39.645 | 35% | 41.282 | 60% | 9% | 16% | 82% | 75% |
| 50 Jahre u. älter | 4.827 | 37% | 4.640 | 55% | 25% | 17% | 86% | 75% |

Figure 5.9: Excerpt of documentation table listing profiling results for populations with incomplete data due to “migration background”, differentiated by gender and age [BER_7, p.23]

“Döbling” (19th district of Vienna) is classified as RGS “Type 2” and thus received the second best rating possible, the neighbouring “Brigittenau” (20th district) is classified as the worst RGS Type 5. As the 19th district is historically home to Vienna’s richest inhabitants (as reflected in expensive real estate and elite schools), the 20th district is home to a more diverse population with overall lower socio-economic backgrounds, lower levels of education, and an overall higher percentage of inhabitants with non-Austrian citizenships. Consequently, while a jobseeker’s socio-economic background is not directly represented in the **AMAS** model as a variable itself, the home address determining the branch office the jobseeker gets assigned to (and cannot influence) impacts the projected **IC** values and categorization, thus reflecting their socio-economic background as a proxy.

5.4.1.3 Emergent Bias

The third type of bias as described by Friedman and Nissenbaum's [102] framework involves bias emerging through the continued use of an algorithmic system over time. For [AMAS], these changes may stem from a variety of developments.

First, societal values and norms are subject to continuous transformation. The AMS's announcement of their plans for [AMAS] in the fall of 2018 came less than four months after the Austrian Constitutional Court confirmed the constitutional right of all people living in Austria to be registered with a third gender option [348]. This development, while reflecting a shift in society's values, is not reflected in the [AMAS] system at all, leaving the system open to bias towards people exercising their newly confirmed right in any other interaction with the Austrian government and its bureaucratic institutions. Even if the system were to be adapted, the lack of historic information on this population would likely result in questionable predictions at best.

Secondly, large scale global events impacting the national Austrian labour market impact the viability of predictions based on the past. The advent of the COVID-19 pandemic, for instance, had profound and lasting impacts on Austria's labour market, massively skewing the job prospects for certain professions. Following the series of lockdowns and waves of business closures in hospitality services from restaurants to hotels it became all but certain that the historically high chances for workers in Austria's strong tourism industry were going to be sincerely diminished, rendering predictions for newly unemployed jobseekers in these sectors unreliable at best, and downright wrong at worst. While the COVID-19 pandemic represents an extreme example, the underlying approach of profiling future chances based on historic data going back years is sensitive to many less extreme events as well. Long-term trends like the digitalisation of many service-industry jobs, outsourcing of jobs as part of the global supply chain transformation can also skew the predictive value for jobseekers in these sectors even in the short term.

Finally, locally impactful developments such as the bankruptcy of a large, local employer in a rural region may significantly impact the predictive quality of the system for people in that area. As the current system can only reflect such changes through the **Regional labour market** variable, and does so only for the average values aggregated over the four-year sliding window of observations, the system is simply not equipped to accurately reflect the changes to an individual's chances if they find themselves affected by such developments. While the overall, nationwide impact of such changes in the system might be minimal in terms of the average accuracy of predictions, the impact for jobseekers in the region will most likely be disproportional and thus bias against them.

These issues of emergent bias are, of course, not specific to [AMAS], but generally applicable to predictive algorithmic systems based on historic data. Flexibility and rapid adaptation, as the case of [AMAS] illustrates particularly well, are often made impossible by limited resources. The burden of these emergent biases thus falls on the shoulders of affected stakeholders, who may already face additional hardships as a consequence of the developments leading to the very same emergent biases.

5.4.2 Potential Discrimination of Jobseekers

Discussing the critical issue of the various biases embedded within the **AMAS** system it is important to distinguish between the existence of bias on the one hand, and their potentially discriminating effects on the other. This does not mean uncritically accepting the AMS’s narrative that no discrimination would be caused by the use of **AMAS**, but rather a more differentiated look at what discriminating effects may occur as a result of (socio-)technical, pre-existing and emergent bias.

The system’s practical implementation and operationalization as part of the consultation process between caseworker and jobseeker determines the real-world impact for the jobseekers. If a caseworker accepts the automatic categorization as is, the system’s impact is most direct and immediate. In this case, one of the most pressing issues of discrimination stems from the system’s (purported) accuracy. Both AMS representatives and Synthesis Research GmbH tirelessly emphasized the system’s *accuracy*⁵⁰ of roughly 85% overall as a key performance indicator. Following their arguments, this level of accuracy serves as a measure of success for the **AMAS** project in light of the original requirement towards a minimum accuracy of 75% during the planning phase **[AUSSCHR_1]**. It is unclear how the stakeholders arrived at this original number or what the justification for accepting a 25% error rate would have been. To put that number into perspective, an average error rate of 15%, as the documents **[BER_7]** claim **AMAS** achieved overall, would still mean an average of roughly 75.000 jobseekers would be misclassified every year, assuming the number of half a million jobseekers would be processed by the system. Given the potentially serious consequences of a misclassification for jobseekers, including the negative psychological effects of being stereotyped as “*low chances*” and restricted access to resources for both the *high* and *low* group, it is debatable whether the use of **AMAS** should be considered ethically justified.

A closer look at those accuracy rates reveals an even darker reality for some groups of jobseekers. Discrepancies in accuracy exist across both populations and constellations, as Figures **[5.7]**, **[5.8]** and **[5.9]** illustrate. The disparate impact of these variations in accuracy and precision is a well-documented issue of big data and machine learning systems **[349, 350]**, and the insights gathered from these case studies are directly applicable to the **AMAS** system. Dubbed “Simpson’s Paradox” **[351]**, these effects may even manifest in equal average error rates for a given level of granularity masquerading significantly worse error rates for other levels of granularity. While strategies to address these issues have been investigated by scholars for a while now **[352, 353]**, there is no indication that any such strategies were used during the implementation of **AMAS**. As a result, the accuracy for categorization to either the group with *high* or *low* chances—to the best of our knowledge, the *precision* (as opposed to *recall* or True Positive rate) of the

⁵⁰The term ‘accuracy’ is a non-literal, but—to the best of our knowledge—correct translation of the term “Trefferquote” used throughout the documentation. For a detailed discussion of the rather laissez-faire attitude of Synthesis Research GmbH towards terminology in statistics and algorithmic classification in general, and the questionable meaning behind these ‘accuracy’ rates, see our previous publication **[4]**, pp. 53-56].

system—ranges from 65% to 100% for different, sometimes very small, sub-populations. Statistics provided in [DOK_2] for the overall accuracy across milestones, i.e., as the jobseekers case of unemployment progresses over a period of months, show a decrease in correct classification for the *high* group from 82% at the beginning of their case to only 69% after 9 months. Certain small subpopulations, e.g. young adults under 25 years of age in Tyrol, are misclassified at even higher rates with an accuracy of only 65% in some cases [BER_7]. As these examples for very small constellations and subpopulations show, the lack of available observations directly manifests in a reduced accuracy of predictions.

The use and method of calculation for these accuracy rates raises further questions, particularly in light of the fact that they were used as key performance indicators shaping the development and implementation of the system. Seeing as the classification into the *medium* group happens as a compound classification (i.e., neither *high* nor *low*), the differentiation between type 1 (False Positive) and type 2 (False Negative) errors is not always discernable based on the provided documentation. As most of this documentation lists only precision values for the *high* and *low* groups, the true number of people misclassified as *medium* remains unknown. This is evidently not accidental, as the threshold values for group assignment were adapted to favour misclassification into the *medium* group over misclassifications into *high* and *low* [NOTES_10, q.13]. Additionally, the Synthesis Research GmbH concedes in [DOK_2, p.62] that, for the calculation of these *precision* values, only “cases which had adequately populated observations ($n \geq 10$ for the constellation including all model variables) were considered”⁵¹.

Our analysis shows that the purported accuracy of the system can often be misleading and is highly questionable as an overall reliable measure of the system’s performance. It is particularly telling that, in the same document [DOK_2, p.61] that lists the most in-depth statistics on the system’s accuracy, the Synthesis Research GmbH argues how little impact misclassifications into the *low* or *high* segment supposedly would have. For those misclassified as having *high* chances, the document claims these jobseekers simply stay unemployed longer than expected and, eventually, get reclassified into the *medium* group, where they can take advantage of all resources available. For the *low* group, they even paint misclassification as a fortuitous happenstance, as they see it as proof that the special measures available to the *low* group worked better than expected and resulted in an earlier return to employment for those affected. Both justifications seem callous and almost cynical, and betray a clear disconnect in perception from the reality of jobseekers presented with their presumably objective and reliable prospects of finding employment. The fact that being stereotyped as having *low* prospects may even create a self-fulfilling prophecy for jobseekers who would, otherwise, have had better chances at finding a job due to other factors (such as their personal motivation): after all, they may interpret the results of their classification as proof of the futility of their efforts, and lose the very same motivation that would have given them an edge in an interview with a potential employer.

⁵¹Orig. “all jene Geschäftsfälle herangezogen, die in ausreichend besetzten Zellen ($n \geq 10$ bei Verkreuzung aller einbezogenen Modellvariablen) zu finden sind.”

In light of these arguments for the supposedly negligible impact of misclassification, it comes as no surprise that the AMS's strategy of dealing with these errors was equally negligible in detail and sophistication. Pressed on the consequences of misclassification, the AMS simply reiterated that no profiling system would ever be perfect, and that it was the responsibility of the decision makers (i.e., caseworkers) to detect and correct erroneous classification. As a *human corrective*, they should both be able to rely on the system's output as a support tool to make their assessments more efficient, but also distrust it enough to question its recommendations. Neither did the AMS ever provide details on *how* exactly caseworkers should evaluate the automated classifications, nor did they resolve the self-evident contradiction between this additional responsibility put on the caseworkers and the claim that the system would somehow ease their burden. Seeing as correcting an error involves additional work for a caseworker in the form of documentation and justification for overriding the system's **CAM** with their **BAM**, it seems quite plausible that misclassifications by the system resulting in structural discrimination against subpopulations of jobseekers would remain either undetected or unaddressed.

5.4.3 System Transparency

In the fall of 2018, when the AMS first announced the system and Synthesis Research GmbH published the first report **[DOK_1]** supposedly documenting the system, the **AMAS** case study stood out from other examples of algorithmic systems in public administration due to the fact that the system's developers not only made multiple verbal commitments towards transparency, but actually followed through by providing some form of documentation. That initial impression, however, proved to be misleading, as the published document raised more questions than it answered. Stylistically, the document seemed to be targeting an expert audience, detailing technical aspects of logistic regression, the model and variables, as well as the evaluation of the system's performance in the form of precision rates discussed in the previous Section **5.4.2**. The fact that the actual calculation of a jobseeker's **IC** value was a comparatively simple ratio and not the result of the logistic regression at all was completely missing from the document. This omission resulted in the misleading impression that **AMAS** was using sophisticated machine learning methodologies for its predictions. Specific information about how the four subpopulations were sequestered depending on the availability of data were missing, as was a more detailed look at the system's precision rates on different levels of granularity. Besides the fact that the system would not be rolled out nationwide immediately, but that the first year would be an evaluation period during which the system would be gradually made available to select AMS branches, no information of *how* exactly the system and its impacts would be evaluated was provided either. Finally, no information was given at all about the concrete operationalization of the system as part of the consultation process beyond the repeated insistence that the AMS caseworkers would be the ones making the final decisions.

The initial claims of a commitment to transparency quickly faltered in the following

weeks, as more and more NGOs and stakeholder representative groups sought to attain further information about the system. Responses by the AMS were mostly limited to those mandated by law (e.g., parliamentary inquiries [PARL_1, PARL_2] or requests for information by the [GBA NOTES_10]), and were never completed in less than a month. As public pressure to respond to criticism rose, the AMS's response was mostly centred around a media offensive by AMS CEO Johannes Kopf, who responded to critics personally in newspaper opinion pieces, his personal blog and in panels at public events. The sporadic, piecemeal and fragmented nature of these responses made a consistent understanding of the inner workings of the system highly difficult. Access to some of the official responses to inquiries posed towards the AMS was only possible through backchannels, as the inquiring institutions had to consider the legality of sharing the information they had received. For the general public, a confusing barrage of contradicting opinions and assessments meant that a well-founded understanding of the system's goals, methods and operationalization was nigh impossible to attain.

For domain experts and independent researchers, the situation looked similarly challenging. Contradicting statements abound, the research leading to our first publication [3] was characterized by a constant weighing of the reliability of our various sources. Even after receiving the trove of documents, reconstructing the technical and operational aspects of the system was a challenging task. While the quantity of documents improved through this disclosure, the quality and coherence of information presented therein did not. Repeated text fragments and formulations often left out the most salient details or were (seemingly on purpose) obfuscated through organisationally specific abbreviations and technical jargon. Out of all the documents we received, only one [DOK_2, pp.20] explained the process of arriving at a given [IC] value based on the constellation, and it did so in purely textual form, without either an example calculation or a formulaic, mathematical definition, on 3 pages. By comparison, the part of the same document explicating the use of logistic regression analysis to estimate the impact of single variable assignments and to calculate the key performance indicators of the model in terms of accuracy and precision stretched out over 32 pages, included detailed statistics, mathematical definitions, and concrete case examples as well. While no explanation for this imbalance was provided, the overly vague and limited description of the [IC] value calculation process gave the impression of an attempt to conceal the simplicity and banality of what is, arguably, the system's core functionality: the prediction of a jobseeker's chances.

For those stakeholders evidently affected most directly—the caseworkers and jobseekers—this lack of *system-level transparency* was equally problematic. The timeline we reconstructed after receiving the trove of documents for our in-depth case study shows that, in late 2018 and early 2019, when the system was already being rolled out to pilot branches of the AMS for evaluation, neither the handbook, other documentation, nor a consistent concept for caseworker training existed. In fact, the KP-1 directive [RICHT_1] governing the new classification process and including the amendments for the use of the system was not even finalized until December of 2019, more than a year after the first branches were given access to the system. The official, internal handbook [HAND_1]

was finalized in November 2019, shortly before the KP-1 directive. During the same year, the AMS assigned regional branch experts as support contacts for caseworkers, to answer questions about the new system. These experts received more detailed information at a train-the-trainers, one-day workshop [SCHU_1] on March 16th 2020, about one and a half years after the system had been rolled out to the first pilot branches. How these experts were selected and how their training would adequately qualify them for this task remains unanswered in the documentation we received. At the same time, it is unclear if these experts would also be tasked with answering questions of jobseekers, or if their responsibility was solely to the caseworkers.

Most of the internal documentation including the handbook, internal training materials and presentations, are dated November 2019 to March 2020, about five months before the [DSB] ruling forbade the use of the system. The materials are comparably homogenous, covering justifications and strategic reasoning for the use of the system, some very limited technical information about the model and calculation methods, and more detailed information about the caseworker interface and the rules governing its use. Interestingly, some materials seem to pre-empt internal criticism or worries, and some even contain a response to external critique. For instance, in a handout for external trainers [SCHU_5], the authors insist that

“AMAS is no automated decision making system, no self-learning data-training and there is no behavioural recording (click behaviour)!”⁵²

[SCHU_5, p.2]

The last part of this statement is particularly interesting, as the idea that the system could be used to track caseworkers behaviour or be used to monitor or evaluate *caseworker performance*, as opposed to *jobseeker chances* never featured in the public discourse.

The same document [SCHU_5] also contains the most detailed description of the system’s accuracy available to the caseworkers, mentioning the precision values of 80% for the short-term, and 85% for the long-term indicator, as well as giving a quick example:

“That means that 80 out of 100 jobseekers, for whom the system predicts a 66% likelihood of at least 3 months of employment in the following 7 months, will attain this goal, and at most 20 will not.”⁵³

[SCHU_5, p.5]

⁵²Orig. “AMAS ist kein Entscheidungsautomatismus, kein selbstlernendes Datentraining und es gibt keine Verhaltensaufzeichnungen (Klickverhalten)!”

⁵³Orig. “Das heißt, dass mindestens 80 von 100 Arbeitslosen, für die das Programm zu Beginn ihrer Arbeitslosigkeit eine mindestens 66%ige Wahrscheinlichkeit errechnet, in den nächsten 7 Monaten mindestens 3 in nicht geförderter Beschäftigung zu verbringen, dieses Ziel auch schaffen und maximal 20 von 100 nicht.”

This quote is particularly remarkable as it contains the most details of any of the training materials regarding how reliable the system's predictions are. All other materials mention, at most, the average value of 85% precision, or omit this aspect entirely. None of the training materials discuss different precisions for different (sub-)populations, or explain any kind of strategy by which caseworkers should detect and pre-empt misclassifications based on these discrepancies.

In terms of the specific calculations leading to the IC values of individual jobseekers, the available training materials are equally vague. While some materials like the handbook [HAND_1], slides for information workshops [SCHU_7], and handouts [SCHU_2] explain the model variables and—to varying degrees of accuracy—their potential values, no clear description of the calculation for the IC values is given. The most salient description of how the system arrives at its predictions is the very general statement “*The AMAS system calculates how other persons in the past fared in a similar situation.*”⁵⁴ [SCHU_2].

Throughout the materials, two contradictory themes as to the relevance of the system's output for the individual can be identified. On the one hand, the system is identified (correctly) as a statistical profiling system that deals in probabilities, likelihoods and prognosis, and “*does not make statements about the individual [jobseekers]*”⁵⁵ [SCHU_7, p.23]. On the other hand, the descriptions of variables, the IC values, and categorization are consistently formulated as directly related to the individual jobseekers: the system's outputs are supposed to indicate “*their*” chances of finding a job, and “*their*” prospects on the labour market, and the model is based on “*their personal attributes*”. The user interfaces (see Figure 5.5) follows this same narrative, as it situates the predicted chances and categorizations clearly within the context of the individual case as well. Of those two themes, the latter is much more pronounced throughout the various training materials. Even in the KP-1 directive governing the caseworker-jobseeker consultation process, this perspective is embedded within the instructions to consider the system's outputs as the “*foundation*” upon which the assessment of an individual jobseeker's chances should rest [RICHT_1, p.14]. Consequently, it seems likely that caseworkers, in practice, will most likely not be able to make the fine distinction between the two competing narratives “*AMAS shows observed chances of similar jobseekers in the past*” vs. “*AMAS predicts this jobseeker's chances*”.

5.4.4 Ex-Post Explainability

The second aspect of algorithmic transparency (see Section 2.3.1 for the distinction) relevant in the context of AMAS pertains the possibility of explaining concrete system outputs during its everyday use. As made evident by the training materials (e.g., [SCHU_7, RICHT_2]), the AMS was concerned about how their caseworkers would respond to questions by jobseekers about the system and their classification. The AMS's

⁵⁴Orig. “*Das Arbeitsmarktchancen-Assistenzsystem berechnet also, wie es anderen Personen in der Vergangenheit in einer ähnlichen Situation gegangen ist.*”

⁵⁵Orig. “*trifft keine Aussage über das Individuum.*”

strategy to address these concerns focused on a number of predefined statements the caseworkers could use to explain the system to jobseekers.

The internal document titled “*Answering client questions*” [RICHT_2] in particular is the guideline that lists answers caseworkers should provide for common questions, including general questions about the system (system-level transparency) as well as questions about the jobseeker’s classification and chances (ex-post explainability). The overall strategy of dealing with jobseeker’s questions outlined in the document asks caseworkers to answer *honestly* and *respectfully*, avoid the use of institutional jargon, and try to avoid focusing on the classification or the IC values, but rather redirect the focus towards the jobseeker’s “*chances and opportunities*”⁵⁶ [RICHT_2, p.1]

General questions⁵⁷ include:

- Why does the AMS use AMAS?
- How does it work?
- Which attributes are being used for the calculations?
- Where are you taking this data from?
- Why are you allowed to use this data?

Overall, the suggested answers to these questions echo the internal communication, including the previously cited description “*The AMAS system calculates how other persons in the past fared in a similar situation.*”. Questions about the variables list personal attributes (e.g., Gender, Age, Citizenship, and so forth) and their potential values in detail, but become vague for the personal labour history. The Regional labour market variable is mentioned as well, but not explained in detail either.

Questions about the classification include:

- What does the system think my chances are? Can I see them?
- Can the classification be changed?
- Why do you think my chances are like this?

Regarding the first question, caseworkers should freely share the values and classification, as well as provide the definitions of the short and long-term integration criterium and the three classification groups. The suggested answer for the second question involves an

⁵⁶Orig. “*Chancen und Möglichkeiten*”

⁵⁷The original questions in German were translated by the author, but—in the spirit of brevity—not all are replicated here verbatim in the original language.

explanation of how often and under which circumstances the system’s calculations may change (e.g., due to changes in the underlying data such as the age group), but also the fact that the AMAS-generated results cannot be changed manually. The second part of the suggested answer then clarifies the role of the caseworker and their power to “*make an additional assessment, in case there are doubts about the system’s classifications or if [they] assess their situation differently.*”⁵⁸. Interestingly, the specific language suggested here does not explicate that the caseworker’s assessment overrules the AMAS assessment, only that it is an additional classification that “*counts more*” than the automated one.

The answer to the question “*Why do you think my chances are like this?*”⁵⁹ reveals an interesting defensive strategy reminiscent of the ‘Computer says no’ sketch discussed in the introduction to Section 2.4: “*This is not my personal assessment, but the result of a probability calculation.*”⁶⁰. This answer’s clear shift of responsibility away from the caseworker and towards the underlying mathematical calculation seems strategic in its attempt to emphasize the supposedly neutral, objective probabilities generated by the system, and may well be seen as deliberate exploitation of the human tendency towards *automation bias* [235, 236, 237]. Supporting this suspicion is the second part of the suggested answer, which points towards the explanatory text fragments or “*Segmentzusatzinformationen*”:

*“I will happily show you the 2 most important criteria, that influenced your assessment the most!”*⁶¹

[RICHT_2, p.3]

In terms of *ex-post explainability*, these explanatory text fragments⁶² are the only information available for both caseworker and jobseeker that directly connect their personal variable assignments to the outcome. At this point, it is worth considering the limitations of this functionality:

1. Only two or, in exceptional cases, three sentences are available
2. Only those jobseekers classified as *low* or *high* receive these explanations
3. Only jobseekers with a complete employment history receive explanations

In terms of absolute and relative numbers, for the year of 2017 as evaluated by Synthesis Research GmbH in [BER_7, pp.17-20], this means that out of roughly 728000

⁵⁸Orig. “[...] *ich kann eine zusätzliche Einstufung machen, wenn es Zweifel an der vom Computer errechneten Einstufung gibt oder ich Ihre Situation anders einschätze.*”

⁵⁹Orig. “*Warum schätzen Sie meine Chancen so ein?*”, emphasis and translation by the author.

⁶⁰Orig. “*Das ist nicht meine persönliche Einschätzung, sondern das Ergebnis einer Wahrscheinlichkeitsrechnung.*”

⁶¹Orig. “*Ich kann Ihnen gerne die 2 wichtigsten Kriterien sagen, die für Ihre Einschätzung am stärksten gewirkt haben!*”

⁶²For a more detailed discussion of these texts, see Section 5.3.4.1 and Appendix A.4

jobseekers, only about 307000 or just over 42% would have received explanations. The remaining 420000 jobseekers were either classified as *medium* chances, or were part of the subpopulations with incomplete data.

Besides the fact that these overall numbers show clearly that this explanation functionality is missing entirely for a majority of jobseekers, even those that do receive explanations have to contend with vague and potentially misleading or downright stereotyping explanations, as discussed in detail in Section 5.3.4.1. The claim made by the suggested answer that these texts would explicate the “2 most important criteria” [RICHT_2, p.3] in and of itself is misleading as well, as the conditions or *criteria* governing which texts are shown are not part of the texts themselves, and not always deducible from the explanations. For example, the text fragment “*You have not taken part in any supportive measures offered by the AMS in recent years.*” depends on both the fact that no measures have been claimed *and* the fact that the jobseeker has had at least one period of unemployment with the AMS lasting longer than 180 days. These problems are exacerbated by the fact that the rules governing *which texts are shown to whom* are, as far as we were able to determine from the documents, only communicated to caseworkers as part of the internal technical handbook [HAND_1, p.18]. It is unclear how much caseworkers were making use of this document, seeing as it was only made available towards the end of 2019, about a year after the pilot branches started using the system, and as its language and structure suggest a different target audience than the caseworkers themselves.

5.5 Chapter Summary

The case study of the AMAS system reveals a truly worrying amount of critical issues and challenges. Through an in-depth document analysis supplemented by qualitative interview data, this case study disentangled the complicated history of the AMAS system, and its embeddedness within the larger context of a semi-autonomous governmentally funded organisation. The concrete implementation of the system was shown to be the direct consequence of the larger trend towards *New Public Management* of public services, following the neo-liberal tenets of the *activation paradigm* through *profiling the unemployed*.

As previously [3, 4] noted, the AMS’s original goals for the system—increased *efficiency* in service, improved *efficacy* of measures granted and resources spent, and reduced impact of *personal biases* of caseworkers—must, under the circumstances, be considered as failed. Given the curious tensions and contradictions in narrative between reducing the burden placed on caseworkers by providing them with a *decision support system*, while at the same time adding significant responsibilities and procedures they are expected to fulfil, it seems implausible at best to claim the use of this system would result in a more efficient caseworker-jobseeker consultation process. If at all, the caseworker’s efforts for assessing jobseeker’s chances on the labour market and deciding how best to support them would be shifted towards their new responsibilities to act as a *human corrective* for an automated classification system, with surprisingly few tools or little agency to fulfil

this role. In addition, the controversial nature of the system and the public discourse its announcement has triggered meant that caseworkers would have to spend additional time discussing and explaining the system and its outputs to the jobseekers assigned to them.

From the AMS's standpoint, **AMAS** was supposed to ensure that they spent the scarce resources they have where they had the most effect, thus increasing the *efficacy* of measures spent. The instrument of choice to reach this goal—the classification and subsequent restrictions to granting costly measures for some of the jobseekers—seems mismatched with the original goal. As the AMS keeps insisting that the caseworkers have the *real* expertise on assessing jobseekers chances, and that **AMAS** as a support system was simply designed to speed up this process, the *real* role of the system would be dangerously close to a convenient scapegoat for unpopular and ethically challenging decisions. If the caseworkers' final say in terms of the classification and potentially available measures is to be taken seriously, and no *rubber-stamping* of supposedly non-binding autonomous classifications were to occur, the caseworkers would still carry the burden of having to take the moral responsibility of their decisions when deciding against granting jobseekers certain resources or measures. Conversely, in the more likely case of caseworkers trusting the system (in the spirit of *efficiency*) and accepting the system's classification *as is*, it seems quite plausible that they would also defer their personal responsibility towards the system - 'computer says no'. Finally, the system's instrumentalization as a cost-cutting measure betrays the seemingly positive aspirations of *efficacy*. In order to improve the impact of supportive measures granted to jobseekers, a number of other approaches could be considered, from a system that evaluates the positive impact of granted measures, to one that directly recommends fitting measures to jobseekers so they can better evaluate their options. The fact that the concrete, resource intensive measures that **AMAS** supposedly should help distribute more effectively only feature in the most coarse-grained form in the model calls the potential for reaching the goal of increased *efficacy* further into question.

For the final goal, the framing of **AMAS** as a countermeasure to human bias and discrimination seems absurd, given the numerous issues of complex, intersectional, and cumulative biases embedded in the data model, the error rates, and the operationalization of the system. The AMS's own admission that it is indeed the human caseworker who must take up the responsibility to correct the system's misclassifications raises the question of who is supposed to control whom in this socio-technical assemblage. In its current state, the system seems more likely to add new, algorithmically embedded biases into the mix, rather than counteracting existing human bias.

The questionable success of the system in relation to these three goals notwithstanding, the concrete issues introduced by the use of **AMAS** should be considered highly problematic at least, and prohibitive to the its deployment at most. The potential for misclassification, computer-supported structural discrimination of disenfranchised jobseeker groups and populations, and the concrete effects these practices could have for individual jobseekers struggling to find employment are significant. Whether or not these challenges could be addressed through socio-technical means at all is difficult to determine, but what

is clear is that the existing levels of system transparency and ex-post explainability functionalities are wholly insufficient to counter the potential for harmful discrimination. While, generally, a causal connection between a transparent system and an accountable one is not plausible, as I discuss in detail in Chapter 2, in the case of [AMAS](#), the lack of *transparency* and *ex-post explainability* is certainly prohibitive to achieve a reasonable level of accountability. To explicate this claim, the following chapter makes use of the proposed [A³ framework](#) for (micro)-accountability to analyse what currently is possible and what would be required to hold [AMAS](#) to account.

5.6 Chapter Conclusions

In the larger scope of this dissertation, the [AMAS](#) case study highlights some concerning trends in the use of semi-automated systems with potentially wide-ranging, but also *immediate* consequences that lead to the following conclusions.

An Immediate Impact

In contrast to EnerCoach, the need for and concerns about the accountability of the system are immediately obvious. [AMAS](#), were it in use today, would directly affect both caseworkers and jobseekers in their interactions, and—claims to the contrary by the AMS notwithstanding—would be integral to the decision making process of caseworkers *by design*. This immediacy of impact underscores the previous arguments made in Section [2.4.3](#) towards a need for *micro-accountability* in addition to *macro-accountability* processes. For the affected jobseekers, holding *AMAS* and the caseworker (as the operationalizing actor in the assemblage) to account is not a matter of demanding *justice* for the overall conduct of the AMS or *deterrence* of future unjust conduct, but the very immediate need to *protect* themselves against (at least potentially) unfair treatment and a *personal injustice*. As stated before, neither of these two scopes of accountability should be seen as mutually exclusive, and improving one may well have positive impacts on the other, and vice versa. Both scopes of accountability, however, are not equal in how much attention and focus they have received in the past from academic scholars and developers of algorithmic systems alike: while macro-accountability processes have been studied, considered and even implemented for algorithmic systems, micro-accountability processes (by this or any other name) have not. Discussing the dangers of runaway, unchecked automated decision making, of bias, discrimination and injustice propagated by algorithmic systems, we tend to look at statistical evidence of bias and discrimination, and offer macro-accountability as solution in the form of pre-emptive audits, ethics guidelines and compliance procedures. But for the immediate question of how an affected person could be assured of their fair treatment and their agency to hold the system to account, the furthest we have ventured is the acknowledgement of lacking *ex-post explainability* of algorithmic systems. Thus, the discussion of a controversial system like [AMAS](#) offers the opportunity to investigate more clearly how *micro-accountability* processes might be designed to address these concerns in all their immediacy.

Shifts in Power and Agency

The increased need for micro-accountability as a *protective* measure in [AMAS](#) stems from the shifts in power and agency as a consequence of introducing the system. Considering [AMAS](#) as a socio-technical assemblage highlights the complex web of inter-dependent and distributed agencies. On the caseworker's side, their choice to grant a specific measure to a jobseeker despite their classification now requires a precise set of actions in concert with both the technical components of [AMAS](#) as well as AMS internal rules and guidelines. Considering the question of moral responsibility for making or omitting such a choice, these interdependencies seem almost purpose-built to counteract the caseworker's *sovereign* agency, as described by Krause [\[222\]](#). That is not to say that caseworker's decisions would be, were the system to be used, free of moral agency or responsibility: after all, they still retain their *reflexive* and *norm-sensitive subjectivity*. But at the very least, it makes their moral responsibility more akin to *partial complicity* than *full culpability* [\[222, p.316\]](#), thus making it easier for them to justify their own choices in connection to the system's suggestions. Here, the claim repeated *ad nauseam* by the AMS that any decisions made will still be the result of a caseworker's independent choice falls apart entirely. By introducing the non-human components of the assemblage, and enforcing an interaction between caseworker and [AMAS](#) as a matter of procedure, their agency implicitly becomes distributed across a variety of actors, including the system itself and those setting the policies governing said procedures.

For the jobseeker, navigating the already complicated assemblage of the AMS, including its services and requirements, becomes even more of a seemingly unsurmountable challenge. To many, the addition of another agentic component influencing the outcome of their meeting with a caseworker may simply go unnoticed, given the already overwhelming set of rules and procedures governing the process of receiving unemployment benefits and gaining support in finding a new job. The impact on their agency, however, is far greater than may be immediately obvious. Before the introduction of [AMAS](#), the caseworker-jobseeker relationship was already characterized by the (questionable) tenets of the *activation paradigm*. Thus, jobseekers could, in order to convince a human caseworker to grant them certain resources like job trainings, argue that they were undertaking the "right" steps in order to minimize their presumed "burden" on society, and that these measures would indeed help them do so. After the introduction of [AMAS](#), however, that argument would now need to be extended to include plausible reasons why the system's assessment was false as well. Making such an argument, of course, may well present an unsurmountable challenge to most, if not all, jobseekers in that situation. After all, without a prior, detailed understanding of [AMAS](#) in terms of system transparency, and without immediate *ex-post explanations* after the profiling assessment was made, making those arguments would be essentially impossible. To claim any reasonable person could retain any sense of *self-efficacy* of agency [\[217\]](#) in light of these limitations seems, of course, patently absurd. Consequently, the system's introduction directly affects the jobseeker's agency in a complex socio-technical assemblage further, leaving them struggling to demand accountability or stand up for themselves.

Awareness, Willingness and Participation

Following these observations, the primary answer to [SRQ2.2](#)—*What actions can system stakeholders take to improve the accountability of their systems?*—seems to point towards their *awareness* and *willingness* to consider all affected stakeholders and their needs as a minimum requirement, before any other attempts at devising measures to improve the situation play a role. Utilizing theoretical standpoints such as assemblage thinking or models of human agency and moral responsibility can, as illustrated above, help raise the awareness for the needs of all affected humans. To avoid a narrow focus on those stakeholders that fit the overarching narratives (such as *efficiency* or *effectivity*) requires a willingness and understanding of broad algorithmic accountability for all stakeholders as a *virtue*. In Bovens and Schillemans [205] words, *meaningful* accountability requires a shift from *defensive* towards *deliberative* accountability. To this end, a broader and common societal discussion of what constitutes acceptable levels of automation and the virtues of transparent and accountable systems may lead to a shift in policy and, perhaps, better regulation. Ideally, developers and decision makers of algorithmic systems—such as the [AMS](#) and Synthesis Research GmbH then may start to direct their attention towards answering the question of *how* to improve their system’s algorithmic accountability, rather than finding reasons *why it is not necessary* to do so or arguing *why they have already done enough*.

In terms of more concrete answers to [SRQ2.1](#) and [SRQ2.3](#), [AMAS](#) may also serve more as a warning example than a guiding one. In terms of the system design and methodologies, the exclusion of jobseekers from the design process stands out as particularly problematic. It comes as no surprise to any scholar in [HCI](#) or [CSCW](#) that the needs of a primary stakeholder group are not reflected in the final system design if that group was kept at arms length for the entire development process; to assume knowledge of those needs from the standpoint of a system developer or decision maker is one of the most fundamental design fallacies identified in the field of Design Studies as well. The implementation of the ex-post explainability feature in the form of explanatory text fragments displays the inadequacy of such non-inclusive design processes particularly well: if anything, the texts seem to be an instance of the aforementioned *defensive accountability* intended to stop further inquiry by the jobseeker, rather than providing informational value to them. For the caseworkers, despite claims of inclusion in the form of workshops, the outcome seems similarly detached from their needs. Between the additional responsibilities as a “*human corrective*” and the need to provide answers to jobseeker questions they simply might not be able to give, the benefits of the system to caseworkers are rather slim. Consequentially, it seems highly unlikely at best that the caseworkers themselves were substantially responsible for the resulting system design, implementation or operationalization.

Considering potential concrete measures to improve the situation, the case study unearthed precious little potential for easy solutions or quick fixes. The system’s overall goals and the subsequent implementation as an [ADS](#) system present fundamental contradictions to the needs of the stakeholders directly affected by its use. Thus, a first step must be a more in-depth analysis of those needs regardless of the larger goals of the

system. In the absence of a willingness to involve those stakeholders directly or allow outside researchers access to them for the purpose of analysis, analytic frameworks such as the [A³ framework](#) presented in the following chapter may offer the only plausible way forward towards answering [SRQ2.3](#) more concretely.

The A³ Framework

The following chapter synthesizes the learnings derived from the case studies of the EnerCoach energy accounting system in Chapter 4 on the one hand, and the AMAS system in Chapter 5 on the other in the form of a comparative case study. Based on these learnings, I present the Algorithmic Accountability Agency Framework (A³ framework) as a tool and analytic lens to evaluate the two case studies in regards to their potential for successful algorithmic accountability processes.

Starting with a short discussion of the comparability of these case studies in Section 6.1, I then explicate some underlying preconditions and assumptions for the use of the framework in Section 6.2.1. Following that, I describe the A³ framework's procedural accountability model and its guiding questions in Section 6.2.2. To showcase the framework's capabilities, I then apply it to both of the case studies and evaluate the results, and draw conclusions on the methodological approaches taken in the design and implementation of the case studies and their impact on the potential of these systems to support successful algorithmic accountability processes in Section 6.2.3. To conclude the discussion of the A³ framework and to close the chapter in Section 6.2.4, I evaluate the framework in context with other frameworks for (algorithmic) accountability, situate the framework adjacent to theoretical work in HCI in the form of the *Human-Artefact Model* and discuss the frameworks potential for use with AIAs and algorithmic audits.

The first introduction of the A³ framework was published previously in [6]¹; due to the limitations imposed by the journal venue, this publication documents only an abridged, first iteration of the framework's development process. This chapter remedies these limitations by significantly expanding on this previous publication, and by extending the discussion to include both theoretical foundations, a practical application of the framework, and the evaluation in context with other frameworks, theories and methodologies.

¹Publication title: "*The Agency of the Forum*"

6.1 Comparability of Case Studies

Considering the descriptions of the case studies as presented in the previous chapters, one might conclude that EnerCoach and AMAS are simply too different to be plausibly comparable, the *a priori* considerations justifying the choice of case studies presented in Section 3.1.1 notwithstanding. Given the methodological differences in how the case studies were conducted and the choice of comparative methodology described in Section 3.4, I have previously stressed the need for a well-crafted rationale of comparability for such cases. In the following sections, I thus present such a rationale in the form of a short description of different dimensions of comparability, outline differences and similarities, and posit my arguments for why, in fact, the two examples are indeed well worth comparing to generate insights into algorithmic accountability. Rounding out this rationale is an overall reflection on the applicability of both case studies in the context of this dissertation and their possible contribution to the larger field of CAS.

6.1.1 Differences between the Case Studies

On a surface level, EnerCoach and AMAS present as substantially different based on their context and field of use, their application, and to a lesser extent, technologies and approaches. First and foremost, the two systems' core functionality and goals are, naturally, quite different. EnerCoach, as an energy accounting system, aims to provide a *retrospective*, calculated insight into the sustainability performance of participating communities and organisations. Through the set of complex reports, it transforms various data inputs (e.g., energy consumption data, energy mix data, climate data) into standardized aggregated outputs, with the target audience being the communities themselves, as well as (external) energy consultants and auditors. AMAS, on the other hand, makes *predictive* assessments about a given, individual jobseeker's chances of finding employment in the future. It bases these predictions on a variety of past data points that are supplied by the AMS as the intermediary entity interacting with the system on behalf of and without direct interaction by the jobseekers. Although the AMS claims the resulting information to be a useful feedback to the jobseeker as well, there is no doubt that the primary target audience for the system's output is the AMS caseworker, who operationalizes the system's categorization as part of the bureaucratic consultation process between caseworkers and jobseekers.

Taking a close look at the stakeholders, differences in scope slide into focus. EnerCoach's stakeholders are essentially groups of humans that share an organisational affiliation, and interact with the system as representatives of that organisation. For instance, both a building facility manager tasked with data entry and the community administrator tasked with generating the reports and maintaining the community's energy mix base data contribute to the same entity within the system—the community—and share various levels of responsibility for the results and outputs. Likewise, the normative assessments made by the system impact the community as a whole, rather than single individuals. While scenarios in which individual humans must accept (political) responsibility and blame

for a negative sustainability performance of a community may exist, the system itself cannot be used to assign said responsibility: its assessment happens on the larger scope of communities, buildings and even regions rather than on an individual level. AMAS, by comparison, is directly assessing individual human beings, and is operationalized as an intermediary or support systems for and between two humans. As the jobseekers themselves have no *individual agency* to affect the system's output directly (other than the remote possibility of, perhaps, lying about data points the caseworker collects from them), it is primarily the caseworker who has the agency to operationalize the results into further action outside the system's sphere of influence. While it may be possible to consider the *constellation* a jobseeker finds themselves in as a kind of group of individuals contributing equally to the system's result, this view is purely abstract insofar as any collaboration between members of this constellation is almost impossible, since they will almost never know or have means of contacting each other. Consequently, members of a constellation share neither blame nor responsibility for the system's predictions, while being fundamentally, individually and directly affected by these outputs at the same time.

The organisational and political context in which both systems are embedded are different as well. EnerCoach was, until recently, provided as a free tool and offered to communities on a voluntary basis. Its design and implementation were determined by the EnerCoach Working Group, as well as entities in the administrative sphere of Switzerland's political and governance apparatus. Their shared interest in the correctness, trustworthiness and usefulness of the EnerCoach tool stems from the a foundational belief that combatting climate change requires both descriptive and normative assessments of communities' performance in order to take appropriate measures to improve energy and greenhouse gas emission footprints for the greater good of Switzerland as a whole. For AMAS, as a profiling system of the unemployed, the fundamental assumption underlying its usefulness is the streamlining of resource allocation and maximising the impact of supportive measures for jobseekers in light of budgetary scarcity. Although the political 'spin' given in public statements by the AMS points to the individual jobseeker as a passive recipient of these services, the system's design is clearly aimed at collective improvements over a large number of people. This larger goal is made particularly evident by the nonchalant approach to potential misclassifications of tens of thousands of people per year, the consequences of which are clearly weighed against the overall budget savings and quantitative key performance indicators such as the number of jobseekers the AMS can claim to have helped reintegrate into the labour force.

Finally, on a technological level, the specific implementation of EnerCoach and AMAS show different approaches as well. Where EnerCoach's complex aggregations, rule-based assessments and highly detailed reports operate on a comparably small amount of highly heterogeneous input data, AMAS does the opposite by calculating extremely simple ratios on a very limited, highly homogenized set of data points stemming from a comparably large dataset.

6.1.2 Similarities between the Case Studies

Contrasting the *differences* between EnerCoach and AMAS outlined in the previous section are some *larger similarities* that, from the perspective of CAS, make a comparison between the systems worthwhile.

Starting with the primary focus of this dissertation, both systems have been shown to have significant deficiencies in terms of transparency and ex-post explainability, resulting in particular challenges in terms of algorithmic accountability. These deficits in transparency and explainability are, in both cases, the result of a certain level of *technical sophistication*. For EnerCoach, the compound nature of the aggregations, the various influences of system-defined constants (such as the energy carrier factors), externally defined climate data, and user inputs, yield a complex, multifaceted problem and solution space, where the sheer number of influencing factors all but guarantees that no two reports look exactly the same. For AMAS, the same issues of complexity apply not to the underlying data model, but to the number of data points and the implicit normalization and homogenization steps necessary to curate this large dataset of more than a million jobseekers and potential constellations. The result is the same: individual system stakeholders, be it energy auditors, community managers, caseworkers or jobseekers, find themselves confronted with an automated and algorithmically produced output that they may not be able to explain or trace back to its inputs, but are asked to trust at the same time. The fact that neither of the two systems are, at their core, AI/ML applications that would present their own, intrinsic and methodologically generated, challenges to *transparency* and *ex-post explainability*, arguably makes the two case studies even more interesting. For once, the source of these deficits lies not in the opaque nature of technology, but in the specific manifestations of the socio-technical assemblage that leads to their complexity. Even though each technical step of calculation could, *in theory*, be plausibly done by hand and could indeed causally explain the link between inputs and outputs, such an approach is still infeasible *in practice*.

Holding the system *accountable* under these circumstances is as difficult as it is necessary. For the stakeholders of the EnerCoach system, accountability is an integral part of the primary mission the tool is supposed to support: providing an account of energy practices that holds up to scrutiny by energy auditors. Consequently, all stakeholders directly interacting with the system have a vested interest in holding the system to account for its outputs, in order to be able to trust these outputs to take the role of the account in the larger energy auditing accountability process. For AMAS, a jobseeker's immediate future may be determined by the system's output, and the negotiations between caseworker and jobseeker to arrive at a shared, agreed upon assessment of the jobseeker's chances on the labour market are embedded in the consultation process as a mandatory step. When the system delivers, in the best case, a starting point for these negotiations, and at worst, an assessment that the caseworker may insist on as the final verdict, it is obviously in the best interest of the jobseeker to hold the system accountable for its conduct. In both cases, the system's complexity and deficits in *ex-post explainability* may significantly hinder the potential for a satisfactory completion of such an accountability process.

Another aspect the systems share is the heterogeneous nature of the *complex assemblage of stakeholders*, particularly in regards to their varying degrees of algorithmic literacy and domain knowledge. EnerCoach stakeholders range from those with very little domain knowledge, tasked with data entry or administrative jobs, to domain experts in energy accounting. On this same spectrum, some stakeholders (e.g., the EnerCoach hotline) may have significant levels of algorithmic literacy and a good grasp of the general automation and calculation approaches implemented in the system, potentially being able to complete some of the system's automated aggregation and calculation steps manually and on paper to assess the plausibility of report outputs. Others may not have the same levels of understanding about the inner workings of the system, and must rely on other resources and collaboration to determine the plausibility of results. For AMAS, a similar picture emerges. While caseworkers can be considered domain experts for the AMS internal procedures, its corporate language and abbreviations, and the general concepts underlying the profiling and classification process, their knowledge about the concrete implementation of the system and the intricacies of statistical errors, misclassifications and embedded (socio-)technical bias will vary based on their personal qualifications. Jobseekers, as the stakeholders primarily affected by the system's classification, are representative as a cross-section of the Austrian populace, albeit with a bias towards certain professions and educational backgrounds. Consequently, neither domain knowledge on AMS internal procedures, profiling and classification approaches, nor statistical or technical knowledge about **AMAS** can be reasonably expected of them. For the question of *algorithmic accountability*, this similarity of a diverse set of stakeholders with varying, but limited algorithmic literacy dictates that accountability processes face serious limitations depending on the individual *actors, fora*, and required *account*.

Building on Section **2.4.3**, the differences in scope I propose for *macro-* and *micro-accountability* are equally applicable to both of the case studies. To summarize my arguments made previously, the overall scope of (public) accountability as characterized by Bovens [22] mostly plays out on the macro-level of governmental institutions, states, the judicial system and interaction with the general public, and falls short when considering the specific and immediate interactions between humans and algorithmic systems. Following the recent focus on *human-centricity* and *digital humanism* in technology development and assessment, I argued for an increased focus on *micro-accountability* as the accountability process that plays out between individual users, operators, those affected by an algorithmic system and the system itself. Both case studies are excellent examples for this multiplicity of the term algorithmic accountability, since both perspectives are applicable to them equally, as I will explicate in the following paragraphs.

Starting with the larger scope of *macro-accountability*, both EnerCoach and **AMAS** face scrutiny by the various institutional fora, be it *political, legal, administrative, professional* or *social*. EnerCoach is bound by an accountability towards its financiers entrenched in the political landscape of Switzerland through the EnerCoach Working Group, the EnergyCities program and the larger European context of the European Energy Awards. The reports and normative assessments the system provides have to fulfil the standards

of energy accounting and sustainability set by these governing bodies. The larger *actors* that may be called to account include the EnerCoach Working Group (as the stakeholder that set the policies) and WIENFLUSS (as the company implementing them in the form of code). Similarly, the community of energy auditors or energy consultants represents a type of *professional* forum that holds the system and its community users to account for their conduct. Finally, in the grandest sense of accountability, the underlying assumption that the use of EnerCoach as an energy accounting tool should lead to more sustainable energy practices of communities may face scrutiny by the general public as a *social* forum; the system in its entire socio-technical assemblage thus would be held to account as to whether it succeeds in this mission. AMAS, on the other hand, faces similar requirements towards large-scale, public accountability. As the AMS itself is financed by the federal government of Austria, it must be accountable to its board of directors and the delegates of the chambers of labour, commerce and industry represented therein as well as the Austrian government, judicial system and the professional community of labour market services in the European Union. This accountability extends across all its strategic and procedural decisions, including [AMAS](#) as a specific implementation of labour market policies. Finally, as the heated and controversial discourse following AMAS's introduction illustrated so clearly, the many different voices of the general public demand accountability for the AMS's conduct in creating and utilizing this tool: from NGOs like *epicenter.works*, advocacy groups for the unemployed, the [GBA](#), to the scientists speaking out about their concerns regarding the system, many different public fora have already demonstrated the power and importance of *social accountability* in the case of AMAS. For both case studies, EnerCoach and [AMAS](#) alike, the taxonomy of accountability presented by Bovens is well applicable and highly relevant, and their various accountability processes fit well within the model of the *actor-forum* relationship.

Simultaneously, both case studies also serve well as illustrations of the demand and necessity of *micro-accountability*. For EnerCoach, the various examples of user interactions I have presented in Chapter [4](#) aimed at tracing erroneous or implausible results, are clearly instances of *micro-accountability* processes in their minute enactment, involving not institutions or the general public, but single individuals representing either forum or actor, going through the process of demanding an account with the possibility of imposing consequences. AMAS, on the other hand, requires the same kind of *micro-accountability*, with even more clearly defined *fora* and *actors*: after all, the discussion and justification of the resulting classification is nothing else than the jobseeker, as the *forum*, demanding an account from the caseworker, as the *actor*, about their conduct in relation to the system's [CAM](#) and [BAM](#) classifications. In both cases, the success or failure of this micro-accountability process can be assessed, and is influenced by various factors, as I will explicate in the following sections.

A final point on the question of comparability of the two case studies from a methodological standpoint relates to the generalization of the proposed [A³ framework](#). As all frameworks, not unlike algorithmic systems, model reality through abstractions and generalizations, one of the core qualitative measurements of a framework is its applicability to a wide

range of instances. Many of the existing, proposed frameworks related to accountability in the academic context of [CAS](#) are narrower in scope (e.g., [196](#), [354](#)), which may limit their usefulness. As one of the challenges of algorithmic accountability lies in its *wicked* nature, a framework that claims to be a useful contribution must also be applicable to a variety of algorithmic systems in different contexts, scopes, disciplines, built upon a variety of underlying technologies and operationalized in various settings. To this end, the differences and similarities between the two case studies as outlined in this and the previous section should prove a valuable point about the usefulness of the proposed [A³ framework](#).

6.1.3 Applicability in the Context of CAS

As any, even cursory, search through social media platforms, news outlets and magazines shows, a significant bias exists in the public discourse of *which* algorithmic systems and technologies we are talking about: engagement strongly favours those technologies complying with an algorithmic imaginary [333](#) of bleeding-edge, highly complex [AI/ML](#) systems. The light in which these systems are being presented often suggests that they are, hyperbolically speaking, either the cause of the impending doom of human society, or its saviours, depending on the flavour and inclination of the source. The nature of this discussion, particularly in the context of the attention economy of social media and the modern online news landscape, comes as no surprise to those familiar with the academic discourse surrounding algorithms, which reflects a similar bias, albeit sans hyperbolic language. Case in point, the focus of accepted papers at this year’s ACM FAccT conference [355](#) shows that an overwhelming majority of publications references [AI/ML](#) technologies, or even subsumes “algorithms” under these technologies.

The fact that these imaginaries seem to function as a reliable trigger to elicit strong emotions in the public discourse can not be detached from the trajectories of academic research and development. Given the nature of co-production [88](#) of technology, a strong engagement with specific imaginaries of algorithms, such as public controversies like the one surrounding the [AMAS](#) system, also impact the potential for funding and the chances for high-profile publications in the academic sphere, incentivizing scientists to direct their attention where both money and prestige are to be found. These underlying issues notwithstanding, there are, of course, plausible arguments to be made for why it is important to study these complex systems and technologies from a scientific standpoint. As I have argued myself in Section [2.3.3](#), the intrinsic qualities of [AI/ML](#) technologies present unique challenges towards transparency and explainability that other, more simple systems do not. This intrinsic challenge becomes even more acute when considering the aforementioned zeitgeist of presenting such complex and inscrutable systems as solutions to an increasing number of social problems and challenges. Critical voices [138](#), [129](#) rightfully point to a lack of accountability of these systems precisely because of the intrinsic challenges prohibiting ex-post explainability, and addressing these issues are thus, arguably, one of the necessary preconditions to holding [AI/ML](#) systems to account.

Considering these arguments, the choice of case studies in this dissertation deserve another look. Given the seemingly simple nature of the underlying technologies—neither EnerCoach nor [AMAS](#) make use of technologies that could be, by any stretch, considered artificial intelligence—one might ask the question: What insights from such case studies might be applicable and useful contributions to a field primarily preoccupied with [AI/ML](#) technologies? In other words: Why waste our time on simple systems, if the biggest technical challenges seemingly lie elsewhere?

To answer these (pointedly formulated, but nonetheless relevant) questions, a more nuanced look at the algorithmic landscape of real-world applications is required. The current state of public and academic discourse on algorithms and algorithmic systems simply does not adequately reflect the diversity, heterogeneity and variety of actual algorithmic systems “in the wild”, particularly when considering the broad conceptualizations of the term as presented in Chapter [3.2.3](#). Regardless of how prevalent these technologies may be in these discourses, the digital transformation of society is built not on the back of high-profile facial recognition systems, AI-based predictive policing or criminal risk assessment, but on the everyday algorithmic systems that both permeate and shape the environments we live in. Following much less sophisticated business logics, rule-based or very simple statistical calculations, these systems nonetheless exert their influence on us in a myriad of ways, from automatically reminding us to pay our taxes or renew our public transport cards, generating bills for the services we use, to allowing us to digitally prove our identity as citizens with our local governments in order to request a new passport, driver’s licence, or register for social services. So ubiquitous are these algorithmic technologies that they, indeed, make up part of the *digital infrastructure* orchestrating our daily lives. These algorithmic systems thus exemplify the “*Banality of Infrastructure*” [\[356\]](#), as Nikhil Anand formulates it most succinctly: Not invisible per se, but too “*boring*” [\[242\]](#) to garner our attention, it is all too easy to overlook the power they may exert—and their potential for injustice. Even more concerning than such unintentional omission is the possibility to intentionally exploit this seeming *banality*. The repeated insistence by the [AMS](#) that [AMAS](#) was *not* an AI-system and the false claim that, consequently, the criticisms posed against it in the public discourse did not apply, is an example of such a deliberate attempt at downplaying the complexity of the system as a defensive stance to avoid further scrutiny. In the end, whether or not these simpler systems exercise the same *levels* of power and agency as their more sophisticated counterparts may well be a matter of discussion; the fact that their *ubiquity*, their *prevalence* and even their *potential* impact makes them deserving of our attention, however, should not.

The complexity of algorithmic systems at the utmost bleeding edge of what computer science promises as the technologies of *tomorrow*, I argue, is a *red herring* distracting us from paying attention to some of the still unsolved challenges presented by the technologies in use *today*. Both case studies exemplify this gap in knowledge well: Neither EnerCoach nor [AMAS](#) are founded on intrinsically inscrutable technologies, yet were lacking in ex-post explainability features nonetheless. While one may argue that, in the case of [AMAS](#), this

gap may be founded on the lack of attention this issue received during development, the fundamentally dissatisfactory attempt to provide explanatory text fragments (see Section 5.3.4.1) also illustrates how difficult explanations for comparably simple correlation-based statistical profiling can be. In the EnerCoach case, by contrast, all good intentions and willingness to provide better *transparency* and *ex-post explainability* for the stakeholders revealed the challenging nature of tailoring these features towards diverse target groups of varying levels of algorithmic literacy and domain knowledge. Furthermore, a major challenge was determining which supportive measures might best fit within the complex interplay between the various components of the assemblage, including the technical components themselves, the EnerCoach Hotline, energy consultants, auditors and end users, as well as the practical reality of energy accounting processes. In summary, neither of these cases had simple, formulaic or previously established and agreed upon solutions to address issues of transparency and ex-explainability, regardless of the (comparably) simple and perhaps “*boring*” nature of these systems.

Considering the core topic of this dissertation, algorithmic accountability, the arguments made above apply as well. These fundamental issues of *transparency* and *ex-post explainability* directly affect the potential for both *micro-* and *macro-accountability* processes by supporting or limiting the ways in which an *account* can be rendered. This insight applies to both the difficulties of explaining inherently complex **AI/ML** and seemingly simpler, less sophisticated systems like EnerCoach or **AMAS**. Consider, for a moment, the hypothetical assumption that significant progress in research on **XAI** as well as *ex-post explainability* in general were to provide definitive solutions to these challenges, however unlikely that scenario may be. Then, our gaps in knowledge on the nature of the algorithmic accountability process, the lack of best practices for various contexts of application, and the lack of more holistic evaluation frameworks for algorithmic accountability would still leave us wanting. In other words, as I have formulated it before in Section 2.7, if we do not know what to do with these explanations, it makes no sense to concentrate all our attention on solving the, arguably harder, challenges of **XAI** before addressing the more fundamental questions posed by the algorithmic accountability process. Put even more boldly: let us learn how to walk before attempting to run.

Finally, we must consider what similarities and overlaps exist between simple and complex technologies in order to reconcile the tensions outlined above. As many of the challenges involving accountability processes stem from the characteristics of algorithmic systems as socio-technical assemblages, approaches to address them may well be applicable to both simple and complex technologies alike. For instance, the learnings we can derive from studying **AMAS** regarding the limitations its operationalization places on micro-accountability processes would (at the very least partially) be applicable to another system using, for instance, *random decision forest* or *k-nearest-neighbour* machine learning approaches to achieve the same result. Case in point: despite the shift in understanding of the underlying technologies of **AMAS** as not, in fact, based on logistic regression for the classification of jobseekers, but rather the simple ratios of “*constellations*”, most of the critical issues related to accountability and transparency of the system remained

the same between our first and subsequent analysis after receiving the complete set of documents.

Similarly, the use of *participatory design* methodologies in the case of EnerCoach provided suggestions and insights that transcend the specific output in the form of visualizations or annotation features. After all, the value in including stakeholders in the design process lies not just in what results these processes yield, but also in the empowerment and increase in agency of those involved, and the spotlight such processes can shine on where the true opacities of a given algorithmic assemblage lie, regardless of the underlying technologies. For the participants of the EnerCoach design workshop, for instance, the implementation specifics of the database queries governing the plausibility framework for report generation remained as opaque as the implementation of machine learning classifier to detect similar data entry anomalies would have been. Yet, in concert with a developer, they could still develop solutions and socio-technical measures to address the resulting issues of *ex-post explainability*.

In summary, the prevalent focus on **AI/ML** should not distract us from the simpler technologies and algorithmic systems already in use as part of our digital infrastructure, however “*boring*” or “*banal*” they may seem. Neither can advances in **XAI** or similar fields alone provide answers to the more fundamental questions we must address to move towards more accountable algorithmic systems, nor are the insights we can glean from the study of less technically sophisticated systems necessarily inapplicable to **AI/ML**-based systems. Indeed, as those simpler and less inscrutable systems already carry the potential for significant, and sometimes harmful, impacts on society, avoiding the highly complex and abstract issues that inscrutable systems raise may be an easier and more promising way forward to figure out the fundamentals of algorithmic accountability and transparency. Consequentially, both the EnerCoach and **AMAS** case studies offer important contributions to the field of **CAS** in general, and may even provide valuable insights into the applicability of **XAI** approaches for real-world examples of larger algorithmic assemblages. As a first step, however, they are certainly a fitting foundation to synthesize generalizable insights into the **A³ framework** as an assessment and evaluation tool.

6.2 The A³ Framework

The **Algorithmic Accountability Agency Framework (A³ framework)** is an assessment tool to evaluate the accountability mechanisms and processes of a given algorithmic system. It is meant to be used as an analytic lens for the socio-technical assemblage of the system, and its primary function is to highlight a system’s existing or future potential to support or hinder human agency in the accountability process.

6.2.1 Preconditions and Assumptions

Building on the theoretical foundation of human agency as *emergent interactive agency*, the proposed [A³ framework](#) relies on the following concrete assumptions and preconditions:

Prior socio-technical analysis The application of the framework requires a prior analysis of the algorithmic system in question, including its overall goals and underlying technologies, stakeholders, and operationalization. A detailed technical analysis about the inner workings—*system-level transparency*—is beneficial, but not necessarily required. In line with this dissertation’s approach, a holistic approach to this analysis akin to *algorithmic ethnography* [21] is recommended.

Willing participation In order for any accountability process to occur, its participant actors and fora must be generally willing to participate in this process. While the framework’s procedural conceptualization of the accountability process and its open question format help guide the evaluation towards identifying supporting and hindering factors to this process, individual participants may still be unwilling to participate. This refusal may take many forms and have multiple personal or social reasons. Among those, the characteristics of the *obligation* of the actor towards accountability constitutes a primary factor, but may not be the only one when considering *micro-accountability* processes in particular. While a refusal to participate can be seen as an exercise of personal agency in and of itself, an assessment of the reasons would transcend the scope of this framework, and indeed this dissertation. Therefore, the framework assumes at least the general willingness to engage in accountability processes as *actor* or *forum*.

Limits to human agency The [A³ framework](#) rests on the assumption that human agency as part of the accountability process can be supported or curtailed by a number of *social*, *technical* or *socio-technical* factors. Identifying these factors thus must be the first step in any attempt to improve the situation and increase the potential for a successful accountability process. For instance, limitations to human agency in the realm of the *social* could include limited literacy or domain knowledge, whereas *technical* limitations could be imposed due to system capabilities or functionalities (or lack thereof). *Socio-Technical* limitations are the result of socio-technical factors such as a purposeful policy of obfuscation or intentional opacity that is reflected in the system’s technical implementation as well. Depending on their origin, these factors may be considered part of the *socio-structural environment* or the *personal internal factors* of the agents, as explicated in Section 2.4.4 on Bandura’s *Social Cognitive Theory*. Considering the algorithmic system from different theoretical standpoints—including, but not limited to viewing it as *socio-technical system*, *socio-technical assemblage* or *actor-network*—can help identify additional factors influencing human agency. Depending on the context of application, some of these conceptualizations of algorithms may fit better than others, thus the *prior socio-technical analysis* should consider which of these perspectives are most applicable.

Assessment and emergent solutions The goal of the framework is twofold: first, the framework is meant to provide a structured view on accountability processes, and offer guiding questions that shine a spotlight on where human agency to successfully complete this process is being limited. Secondly, by breaking up this process into its constituent parts, potential solutions to address some of these specific issues will emerge, which can subsequently be evaluated as part of a design or requirements analysis process. This design process must be tailored to the case in question, and is not part of the [A³ framework](#) itself.

Improving human agency requires nuanced approaches Assuming the limitations to human agency in the accountability process can be identified, addressing them will require careful consideration to avoid either technical solutionism or undue burdens on human participants. Just as the limiting factors to human agency may span the entire range between the *social* and the *technical*, so must the measures to address these issues be attuned to the underlying problems.

As a final point, the framework purposefully does not offer quantitative assessments of accountability. Consequently, using the framework to determine a definitive, quantitative ranking of different systems is not possible. The reasoning behind this choice rests on the understanding of *algorithmic accountability* as a *wicked problem* as explicated in Section [2.5](#). To provide a quantitative assessment tool would suggest that instances of wicked problems are necessarily quantitatively comparable (which, by definition, they are not), and would also encourage a fundamentally reductionist view of accountability processes in algorithmic systems. Given the complex nature of algorithmic systems as socio-technical assemblages, the framework instead encourages the use of critical reflection and qualitative inquiry to tackle the issue of algorithmic accountability.

6.2.2 Procedural Accountability Model & Guiding Questions

The [A³ framework](#) structures processes of algorithmic macro- and micro-accountability with a special focus on the dimension of *agency* of the involved *actor* and *forum*. As such, it is a procedural adaptation and extension of the taxonomy for *public accountability* as a *mechanism* described by Bovens [\[22, 156\]](#), which I have discussed in some detail in Section [2.5](#).

The model describes the accountability process as the bi-directional exchange between the *actor* and the *forum*. Brandsma and Schillemans [\[196\]](#) proposed to structure this into the three phases of *information gathering*, *deliberation* and *consequences*. Through the lens of *human agency*, these phases reveal four types of (inter-)actions, each requiring that the respective *forum* and *actor* utilize their agency: (1) *Requesting Information*, (2) *Providing Account*, (3) *Imposing Consequences* and (4) *Effecting Change*. Figure [6.1](#) illustrates this process.

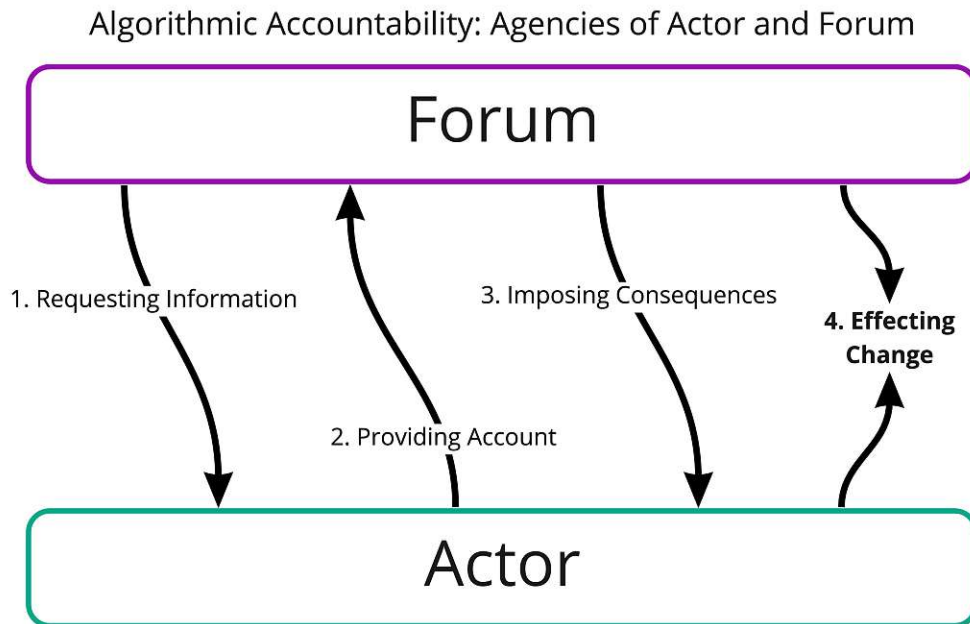


Figure 6.1: Procedural model of the accountability process for the A³ framework.

Requesting Information

Starting with the first step, the *forum* must have the agency to *request information* of the *actor*. Depending on the identity of the *forum*—be that an individual human, an NGO or civil society watchdog organisation, a scientific institution, journalists, a legal, administrative or executive governmental institution, or simply the general public—the *mode* of agency may be *personal*, *proxy* or *collective* agency [24]. In practice, the preconditions for that agency are similar: to be able to request information, the *forum* must (1) be aware of the existence of the system, (2) be able to identify who the *actors* are and how to contact them, and (3) have the requisite means to make their request for information. Many factors can negate these preconditions before the process has even started. Having knowledge of the existence of the algorithmic system in question is not always guaranteed, as the numerous challenges to *algorithmic transparency* detailed in Section 2.3.2 show. Identifying the actor(s) or, in other words, *knowing who to ask*, can be a challenge in and of itself, as Nissenbaum [136, 227] so vividly describes as the *many-hands problem*. With the number of people involved in conceptualizing, designing, implementing, maintaining, and utilizing an algorithmic system, finding the appropriate actor may well prove prohibitive enough that the *forum* may lose their *belief of self-efficacy* of their action, and choose not to follow through with their request. Even if they do make their request, but pick the wrong actor, they may find themselves in a never-ending chain of referrals from one involved party to the next. Finally, formulating

and then relaying that request for information can be its own challenge. Similar to figuring out *who* to ask, *knowing what to ask* may require levels of *algorithmic literacy* that exceed the forum’s available knowledge. Not all algorithmic systems provide as simple and clearly identifiable outputs as **AMAS** does (e.g., a single classification). For EnerCoach, for instance, many requests for clarification of report results were triggered by what energy consultant “D” described as a “*gut feeling*” that something was wrong: In order to then request further information, they had to transform that gut feeling into a concrete question, collect and provide additional information on what they were expecting vs. what the system provided, and so forth. For those *fora* that cannot provide such a level of detail—e.g., members of a minority being targeted by a biased predictive policing system—this precondition may preclude their agency to act as a forum entirely.

Providing the Account

Having successfully navigated these issues, the *actor* receives the request, and has to *provide the account* as a justification for their conduct involving the algorithmic system in question. As described above, unwilling actors may exercise their agency to refuse to provide this account, or refer the request to another, perhaps more suitable *actor*, effectively skirting their responsibility. The reasons for such a referral may be manifold, and can range from legitimate doubts about their abilities to provide the requested information to using the many-hands problem as an excuse to avoid personal scrutiny. Presuming the *actor* is generally willing to participate in the process, as previously determined as a general precondition for the use of the **A³ framework** in Section **6.2.1**, they still may not have the agency to do so. Here, an entire catalogue of factors may influence their ability to deliver an account: legal and procedural restrictions as a consequence of intentional opacity or secrecy, unclear organisational responsibilities, a lack of time and resources, or a (perceived) lack of knowledge, and any combination of these factors, could be the reasons even a motivated, well-meaning *actor* simply cannot comply with the request for information they received.

The nuanced, manifold issues of algorithmic transparency and ex-post explainability play a particularly important role as well: depending on the system, the questions a *forum* may have asked could be mismatched to the *actor*, or simply the result of misunderstanding or lack of (domain) knowledge; they may be unanswerable due to the inherent complexity of the system (e.g., for machine learning or AI applications), or the answer may be possible, but meaningless to the *forum*. Naturally, the *algorithmic literacy* of the *actor* influences their ability to provide an account, as well as the quality and detail of that account as well. For instance, a jobseeker asking a caseworker a detailed question about the statistical likelihood that they have been misclassified will most likely not result in a satisfactory account if they neither have the statistical background knowledge nor access to the actual precision rates for the jobseeker’s subpopulation and constellation.

Considering these numerous challenges, a successful completion of this step may often only be plausible when considering *proxy* or *collective* agency, as opposed to the *personal* agency of an individual *actor*. Depending on the conduct in question, a meaningful

account could require input from a variety of actors, such as system operators, developers or the individuals in charge of setting the policy that the system manifests.

Imposing Consequences

The third step of the process—*imposing consequences*—is beset with the most significant and simultaneously, complex limitations to individual agency. The ability to demand any kind of consequences for the *actor* is most obviously impacted by hegemonial power imbalances between the *forum* and the *actor*. While examples of more equal power balances in *forum-actor* accountability relationships do exist (e.g., EnerCoach Working Group members demanding changes to the system from the developers as part of their product warranty), in most cases of large-scale algorithmic systems, the *forum* has little implicit agency to make these demands. Even in the rare cases where an suitable account was given, and this account documents illegal conduct (e.g., discriminatory practices) in the legal sense, the *forum* may be limited *proxy* agency by taking legal action through an attorney acting on their behalf. The applicability of this particular type of legal consequences is complicated by the general lack of regulation governing algorithms, the myriad of grey areas a given actor's conduct may fall in, and the time and financial resources required to engage in lengthy litigation. Given the fact that the most impactful algorithmic systems are also often complex and expensive, the organisations responsible for their creation and/or conduct are often large enough to have the means to avoid legal consequences.

This means that both individual and smaller collective fora, such as civil watchdog groups, may only exercise their agency to impose consequences by other, more commonly available means. This may include the publication of the account and their findings, organising collective action and protests, or otherwise elevating the hitherto individual accountability process to a social accountability process by forcing a public discourse about their systems to occur. Although these forms of consequences—be it public outrage, grass roots protests, or negative media coverage—can certainly be powerful exercises of *collective* agency [357], an individual's *belief of self-efficacy* is challenging to maintain considering how slim the chances of success for starting such a movement can seem.

As argued in Section 2.4.3, the introduction of *micro-accountability* also necessitates rethinking the nature of plausible consequences in these processes. In addition to the potential to impose consequences traditionally part of macro-accountability processes through exercising *proxy*- or *collective* agency, or even escalating the procedure to a macro-accountability forum, new and more limited forms of individual consequences may be more fitting for the scope of the system. These could include the option for an individual *forum* to opt out of the use of the system, demand the deletion of personal data held by a data provider, or to escalate a (semi-)automated decision making process to a human instead.

Effecting Change

The question of human agency to impose consequences on the actors leads directly to the last and final step in a successful accountability process: the agency to *effect change*. It is worth noting that this step is not explicitly part of either Bovens' or Brandsma and Schillemans' [22, 196] definitions of accountability. It follows, however, implicitly from the underlying intention to make algorithmic systems more accountable: After all, consequences that do not carry at least the potential for positive, and ideally sustainable changes are meaningless for all but the most proximate stakeholders involved. At the same time, just because the consequences imposed in the previous step *could* conceivably lead to such changes does not necessarily mean that they will, or even should. After deliberating the account given by the *actor*, the *forum* may well conclude that the conduct should be deemed unacceptable, and demand consequences that would indeed effect changes to future conduct. However, fundamental disagreements between *forum* and *actor* on whether or not said conduct was justified may still occur. Binns [151] provides an illustrative example with their hypothetical case of an algorithmic credit-scoring system:

“For instance, the bank might justify their decision by reference to the prior successes of the machine learning techniques they used to train their system; or the scientific rigour involved in the development of their psychometric test used to derive the credit score; or, more fancifully, divine providence. The loan applicant might reject such justifications for various reasons; for instance, on account of their being sceptical about the machine learning technique, the scientific method, or the existence of an interventionist God, respectively.”

[151, p.2]

Binns later explicates such conflicts as the result of irreconcilable “[d]ifferences of opinion about the normative standards” [151, p.6], and although an accountability process leading to an end result of, put plainly, “agreeing to disagree” will most likely not be particularly satisfying to the *forum*, it nevertheless could be considered a completed accountability process. However, if the limiting factor for effecting change is not such a disagreement, but rather the *a priori* limited agency of the *actor* to implement said changes, an accountability process may be futile from the start. In other words, if the *actor* has no ability to influence any kind of change, from small changes like adapting their future conduct or the behaviour of the system, to extreme changes like the scrapping of the entire system, the *virtue* of accountability must be called into question in this case entirely.

Unsurprisingly, there may be a great number of reasons why the actor's agency to effect such changes may be limited, starting with their own, limited power within the larger organisational context in which the system in question is being utilized, to prohibitive costs or even technical reasons that make the required adaptations simply unfeasible. An example of such limited agency would be the many cases [358] in which algorithmic systems have proven incapable of processing users with unusual names. Often, these issues

are the result of technical limitations or wrongful assumptions by the system’s developers, and constitute either a form of technical or pre-existing bias. The widespread nature of this problem of so-called *edge-cases* even inspired the famously on-point “xkcd” comic “*Exploits of a Mom*” depicted in Figure 6.2. Even though the *actor* in an accountability process involving such a system error may well concede in their account that this conduct is not justifiable, adapting the system may require significant efforts in order to cover all possible edge cases. Consequently, as the examples of such cases show, companies often decide against such adaptations as a trade-off between a small number of dissatisfied users and significant costs for fixing the error.



Figure 6.2: XKCD comic titled “*Exploits of a Mom*”² illustrating the challenges of processing unusual user names [359].

Whether or not an accountability process that reveals such unjustifiable conduct can lead to actual changes thus depends on the *socio-structural context*. The agency to effect such changes may not lie with the *actor* in the accountability process, but other actors in the domains of organisational or political decision-making or algorithmic governance. For some problems, a broader societal discourse that transcends the scope of the accountability process may be the only plausible course of action to effect change.

Guiding Questions

This outline of the accountability process and the numerous limiting factors impacting human agency illustrate the fragile nature of the entire process. If just one of the steps fails due to the limited agency of its actors, neither a satisfying nor even *satisficing* [241] conclusion to the accountability process will be reached. Due to this fragility, it is paramount to assess the potential for human agency throughout these steps as early as possible, and identify limiting factors and enabling resources. When considering potential measures designed to improve a system’s accountability, they must be assessed in light of their impact on human agency as well: designing accountability measures that fail to support human agency, or even circumvent it, will always carry the danger of making things worse, if only by introducing another level of opacity to the system.

²Image credit: Randall Munroe / xkcd, licensed under (CC BY-NC 2.5)

| Step | Guiding Questions | Agent |
|---------------------------|--|---------------|
| 1. Requesting Information | <p>Which factors may restrict the ability of the <i>forum</i> to request information?</p> <p>How can the <i>forum</i> be made aware of the existence of the system? What information or domain knowledge does the <i>forum</i> require to formulate their inquiry? How can the <i>forum</i> identify the <i>actor</i>? What channels of communication can the <i>forum</i> use to communicate with the <i>actor</i>?</p> | Forum |
| 2. Providing Account | <p>Which factors may restrict the ability of the <i>actor</i> to provide an account?</p> <p>How can the <i>actor</i> respond to an inquiry? What information or domain knowledge does the <i>actor</i> require to provide an account? What tools does the <i>actor</i> have at their disposal to collate the required information? What channels of communication can the <i>actor</i> use to communicate with the <i>forum</i>?</p> | Actor |
| 3. Imposing Consequences | <p>Which factors may restrict the ability of the <i>forum</i> to impose consequences?</p> <p>What information does the <i>forum</i> require to comprehend and assess the justification? What options does the <i>forum</i> have to impose consequences? What possibilities for change do these consequences provide?</p> | Forum |
| 4. Effecting Change | <p>Which factors may restrict the ability of the <i>actor</i> / <i>forum</i> to effect change?</p> <p>What influence can the <i>forum</i> exercise on the algorithmic system? What influence can the <i>actor</i> exercise on the algorithmic system? What <i>other stakeholders</i> are required to effect change? What technical, social and procedural limitations may shape the changes required?</p> | Forum / Actor |

Figure 6.3: A³ framework: guiding questions for each step and agent of the accountability process.

To support this process of assessment, the A³ framework offers a set of *guiding questions* for each of the steps. Applying the questions to either real or hypothetical accountability process should help in both gaining insight on the shortcomings and hindering factors for the process, as well as serve as a starting point for developing additional measures, technical features, supportive resources or adaptations to the socio-technical operationalization of the system that can improve the situation. Figure 6.3 lists these questions for each of the four steps in the accountability process.

As mentioned in Section 6.2.1, before the framework can be applied to a concrete accountability process, such a process and its requisite *actors* and *fora* have to be identified. To this end, a comprehensive stakeholder analysis and, ideally, a detailed case study of the system in question is necessary in order to identify plausible *actor–forum* combinations. Furthermore, multiple different types of accountability processes, depending on the nature of *actors*, *forum*, *conduct* and *obligation* may be possible, as Bovens' [22] taxonomy shows³. Finally, as Wieringa [23] points out, accountability processes may occur *ex-ante*, *in medias res* and *ex-post* in the lifecycle of an algorithmic system. Considering the diversity of algorithmic systems and their socio-technical assemblage, including the domain, technologies, context of application, and stakeholders, prescribing the *right* methodologies for all possible systems for this necessary prior case study and stakeholder analysis is simply impossible. Consequently, to avoid narrowing the potential scope of application for the A³ framework, the framework itself does not cover this prior research.

³For an overview of these types of accountability processes, see Figure 2.5 in Section 2.4.1

The methodologies used in the case studies for this dissertation (see Chapter 3) may, however, serve—at least—as a starting point and, perhaps, even best-practice examples for future case studies of algorithmic systems, as they have been proven to deliver the required results to apply the A³ framework.

Based on the information gathered in prior case studies, and after selecting a real or hypothetical accountability process, it may be prudent to explicate the concrete accountability *scenario* [360, 361] and create *personas* [362, 363] for the *actor* and *forum* prior to applying the guiding questions. Both *scenarios* and *personas* are well established methodologies in HCI, user and design studies, and CSCW, and adapting them for CAS is certainly an appropriate approach to provide a solid foundation for the use of the A³ framework.

The guiding questions for each phase of the accountability process always include a single, *primary question*, and a set of additional *secondary questions*. The *primary questions* are formulated in such a way as to elicit answers focusing on the overall restrictions and limitations that the system’s implementation and context of use place on both *actor* and *forum* agency. The broad nature of the primary questions is purposeful insofar as the answers can be collected in different forms and by various methods, from brainstorming keywords, checklists, short paragraphs or even longer design fiction [364] approaches. The conscious decision not to limit the results by insisting on quantitative evaluations should be seen as a commitment to the applicability of the A³ framework in a wide range of contexts and algorithmic systems. Instead, the resulting qualitative data should be as rich and detailed as contextually appropriate, and as concise and concrete as possible depending on the knowledge and access to the system. By allowing this wide range of answers, the framework retains usefulness even in those cases where a severe lack of transparency (intentional or otherwise) may prohibit a researcher from applying other, more domain-specific and fine-grained accountability frameworks.

To still allow a more granular analysis of *actor* and *forum* agency throughout the process, the *secondary questions* point towards more specific aspects of each phase. Depending on the case study, their relevance or applicability may be more limited than the primary questions. Where applicable, however, these secondary questions will spotlight the common challenges in the accountability process as outlined above, and elicit responses aimed at generating potential measures to support human agency in each step as well. To this end, these questions are formulated in a positive way, e.g., by asking what channels of communication are available to the *forum*, or the concrete ways in which an *actor* can exercise their agency to respond to the inquiry for an account. It is worth noting that the *process* of answering these questions is as important as the *answers* themselves and may even provide additional insights that the answers alone do not. For instance, if the question “*How can the forum identify the actor?*” is particularly difficult to answer, or if the answer comes with significant conditional requirements (e.g., requirements towards domain knowledge of the forum), this realization by itself may be a valuable indication of systemic deficits. Similarly, in case the A³ framework is used as a purely retroactive assessment tool, a given question may yield a strictly negative answer: For example,

“The actor has no influence to effect any kind of changes, and also does not know which other actors could do so” could be an answer to the question “What influence can the actor exercise on the algorithmic system?”. In this case, a serious and *a priori* prohibitive issue has been identified, which calls into question whether a successful conclusion to this accountability process is even possible in any case.

6.2.3 Applying the Framework

In the following sections, I exemplify the usefulness of the **A³ framework** by applying it to selected *accountability process scenarios* for each of the two case studies *EnerCoach* and **AMAS** respectively. Each application focuses on a specific instance of *micro-accountability* in the context of the algorithmic systems. The reasons for focusing on *micro-accountability* over *macro-accountability* for this example are twofold: First, micro-accountability is significantly under-represented in the current state of research, as I previously described in Section 2.4.3. Thus, this section also serves as a contribution to the field of **CAS** as an example of in-depth analyses of micro-accountability processes and their importance for the overall accountability of a system. Secondly, many of the existing frameworks only focus on the macro-accountability (for a more detailed look at the state of the art in this respect, see Section 6.2.4). Consequently, this focus on micro-accountability provides a showcase of the **A³ framework**'s capabilities to support the analysis of these micro-accountability processes as well.

Each case study analysis starts with a short textual vignette explicating a hypothetical, yet concrete scenario in which a micro-accountability process is triggered. Accompanying this scenarios are short introductions of the *personas* that take the role of the *actor* and *forum*. Both *scenarios* and *personas* are based on either real or at least plausible accountability processes derived from the prior analysis of the case studies. In the next step, the four phases of the accountability process scenarios are iterated, and answers to the guiding questions are given in a textual form where applicable. Each scenario ends with a summary assessment of whether or not the accountability process could plausibly be completed successfully, and what the result may have been.

6.2.3.1 EnerCoach

Accountability processes in EnerCoach most commonly occur in the context of verification and plausibilisation of report results, in which stakeholders such as energy consultants or auditors try to determine if implausible results for a given report and community are (1) correct and simply provided them with an unexpected insight into the community's energy practice or performance, (2) the result of data entry errors or other mistakes made by a user of the system, or (3) the result of a software or implementation error in the calculation routines for the report in question. The system's trustworthiness is a crucial factor in its success: communities must be able to trust the accuracy of reports, since they directly impact a community's ability to be certified as an EnergyCity, an important label with potential political and financial ramifications for both small and large communities. Although the system's designers, operators, as well as the EnerCoach hotline staffers are

not immediately legally liable for the correctness of the results, *reputational concerns* can be considered what Buhmann et al. call “*a pragmatic necessity and a normative obligation*” [365, p.272] towards accountability. In other words, what is at stake for EnerCoach is not necessarily legal compliance, but the contribution the EnerCoach system makes towards the larger fight against climate change and the promotion of sustainable energy practices by participating communities. The following scenario illustrates this *pragmatic necessity* and the resulting accountability process.

Scenario *Implausible Report Results*

“K is an energy consultant for the region of Val de Bagnes, Switzerland. She has just completed the data entry for new electricity and gas mix data for the year 2021, and generated the energy certificate report for Le Châble, one of the communities in the region. Looking at the result, she discovers that the water consumption keyfigure performance classification has improved by 4 steps from a ‘D’ rating to an ‘A’ rating since 2020. While she is not intimately familiar with the community’s water consumption practices, she knows the improvement to be significant enough to potentially raise suspicion in the upcoming audit process, and decides to investigate whether or not this is a legitimate result, the consequence of a calculation or software error, or a mistake made during data entry.”

Forum *Ms. K, Energy Consultant*

“K has been an energy consultant for more than a decade, and has significant domain knowledge in the field of energy accounting. She knows about the general process by which keyfigures and performance classifications are calculated, but has no technical knowledge about the EnerCoach system internals. In her work as a consultant, she uses EnerCoach as well as a number of other energy accounting tools to prepare yearly reports that will be assessed by energy auditors to certify her clients’ sustainable energy practices as part of the EnergyCity program. She is aware of the fact that she is liable for any deliberate attempts to manipulate the reports to show a better performance than factually correct.”

Actor *Mr. S, EnerCoach Hotline Staffer*

“S has recently completed his training as a hotline staffer for the company providing EnerCoach support to users of Romandie, the French-speaking part of Switzerland. He has been working with users of the EnerCoach system for about two months, and has a good grasp of the most common issues and support requests. At this point, he is not familiar with all French communities using the system, but is aware that the larger communities and regions, including Val de Bagnes, have significant stake in the system’s correctness. He is eager to support the communities to ensure they continue using the EnerCoach system and do not decide to migrate to a competitor.”

Step 1: Requesting information

Having identified the *scenario*, *forum* and *actor* in this process, the analysis of the resulting accountability process starts with the *agency of the forum* to make the request for an account. Considering K, the persona in question, has significant domain knowledge, but limited technical knowledge about the system’s internals, the most obvious hindering factors to express her request are centred around both *system-level transparency* and *ex-post explainability* of the implausible result. Additional factors may include limited time and resources to invest in this issue: if she is operating under time constraints determined by her employer, the region of Val de Bagnes, she may even decide not to follow up with this particular issue at all.

Assuming she does decide to investigate, she may want to ensure that she cannot find the reason for the implausible result by herself as a first step, by screening the water consumption and building zone data. Limiting her chance at success may be the fact that, as an energy consultant, she did not enter this data herself, and the fact that the region as a whole has over 200 objects and about thrice as many water meters. Each building also features at least one building zone, which complicates the search for data entry mistakes to the point of searching for a needle in a haystack. Alternatively, not unlike what Sandvig et al. [324] suggest as field experiments to audit black-boxed algorithms, trying to change certain data points in one or more of the buildings in question may allow a verification by comparison: If the changes show either very little, or unexpected, results in the reports, this experiment may indicate that, indeed, a calculation error lies at the root of the problem.

If she cannot determine the culprit out of all these model entities, she needs to collect the implausible result (e.g., take a screenshot), provide as much information about the community in question and why she suspects an error in the system, and make her request for clarification to the *actor*. Finally, in order to send off her request, she must be aware of the existence of the EnerCoach hotline (an information the system itself, as well as the training she received when she first started using the system, conveyed to her), and choose how to contact them. In her case, writing an email may seem the most

promising approach, as the complex nature of the underlying data makes communicating her request through the phone more challenging and error-prone.

Taking a closer look at the *secondary questions* for this phase, we can see immediately that, in this case, (1) K is obviously aware of the algorithmic system she is using, (2) has the requisite domain knowledge to both identify the issue and formulate the request for support, and (3) most likely has knowledge of the *actor* S and (4) how to contact them.

Step 2: Providing the account

As S, the EnerCoach hotline staffer and *actor* in this scenario, receives the support request, he has multiple courses of actions. As he is less familiar with the system, but may have colleagues that have more experience, he may immediately defer the question to them, and relegate his agency from *personal* to *proxy* agency. Otherwise, a number of factors influence his ability to answer the question and determine whether or not the results are, in fact, the result of a system error or a data entry error, or if they are correct and simply reflect an edge case of circumstances unfamiliar to the *forum*. First, his domain knowledge about both energy accounting in general and the peculiarities of the EnerCoach system in particular could be the result of similar support request in the past, pointing him in the right direction to investigate certain frequent data entry mistakes that could explain the result. Second, if that approach yields no explanation, the system's lack of *ex-post explainability* features may make it impossible for him to trace the inputs to the (implausible) output, and he may have to contact a different *actor*, namely WIENFLUSS's technical staff, to investigate the problem. As an EnerCoach hotline staffer, it can be assumed that he is aware that WIENFLUSS is available for support requests and is also the primary contact for suspected bug reports. Limiting his agency to make that request, however, may be the fact that WIENFLUSS is operating on a scarce support budget as well, and that his superiors might encourage him to only make such requests as a last resort or only if there is strong evidence of an actual software error.

In summary, S's agency to provide an account is limited by their domain knowledge, their familiarity with the specific community in question, potential economic factors such as the cost of escalating support requests towards WIENFLUSS, and finally, the system's available tools to trace the result back to its inputs. Answering the secondary questions thus reveals that the administrative aspects of communicating with the *forum* or other *actors* are most likely not a problem, but that the lack of tools available to him and the requisite domain knowledge may well make answering the question impossible.

Step 3: Imposing consequences

Depending on the outcome of step 2, the *forum* may have received an answer by either the hotline staffer S themselves, one of their colleagues, or from WIENFLUSS. This answer may, essentially, point to one of three possible explanations: (1) the result is correct, (2) mistakes were made in the data entry, or (3) a software error was identified

that was the root cause of the wrong result. In the first case, K may be satisfied with the answer and conclude that no further consequences are necessary; she may take the explanation given by the actor as a learning experience that certain combinations of (correctly entered) data can result in such extreme changes between years. In the second case, she may want to contact the person responsible for the data entry, and demand an explanation for the mistake—in this case, the new request should be considered its own micro-accountability process deserving of analysis. Finally, in case the *account* she received points to an actual error in the system, her agency to impose consequences depends on her identity and influence with other *actors*. Assuming she is not seeking damages through legal action for her time spent identifying the error, the most likely consequence she might want to impose is one that leads to the correction of the error. Furthermore, the type of consequences she might consider appropriate depends on the severity of the error: small, isolated and rare issues might not warrant further action beyond notifying users of the problem, perhaps with a strategy to avoid it. Larger errors that could affect a greater number of users and communities might require fixing the problem altogether, and thus depend on both her and the *actor's* agency to effect changes as discussed in the next step.

Considering, for this step, only her ability to impose any kind of consequences, answering the secondary questions mostly point to her understanding of the account, how clearly the account identifies the issue as a software error, and whether or not the account provides a justification for not attempting to change the system as the requisite preconditions for her to impose consequences. In the unlikely case that K would seek to impose consequences on the person or organisation responsible for the error, her ability to do so is severely limited by her knowledge about the development process and the stakeholders involved in it. Determining a culprit in this case might not even be possible if the error stems from an early phase of development, as pinpointing whether or not the problem stems from the specifications or the implementation would require information that might simply not be available any more. This example also illustrates the difference between *moral responsibility* and *moral agency* as discussed in Section 2.4.5 particularly well: Determining the culprit would be necessary to assign blame (i.e., *moral responsibility*). Separating this issue from the fact that a *morally charged action* (the system's erroneous output) should be addressed and, ideally, corrected to avoid this behaviour in future thus means that the *ability to effect change* is independent from the *ability to determine culpability*, thus potentially reaching step four without having to determine moral responsibility at all.

Step 4: Effecting Change

Having determined that a software error was a the root of the problem, both the *forum* K and *actor* S could try to ensure that the error gets corrected and thus attempt to effect change as a consequence of the accountability process. Both of them have different levels of agency in this case.

As a long-time energy consultant, K may have direct contact to the EnerCoach Working

Group, and could lobby for the funds to correct the error. Depending on the gravity of the error and her domain knowledge of the impact this error may have on other communities, she may also pressure either WIENFLUSS directly to address the issue under software warranty and liability laws, or organize colleagues and other communities to utilize *collective agency* to demand that the problem is fixed. Either of these options require a more in-depth knowledge of the system and its stakeholders than most EnerCoach users, including most energy consultants or auditors, have. For S, as a hotline staffer, access to the EnerCoach Working Group and WIENFLUSS is much easier, as contact with these entities happens more regularly as part of his everyday job. Furthermore, S will have an easier time arguing for the necessity to fix the error. Through his knowledge of prior support requests, S may have proof how many communities would be affected by the error, and can make a better case to the EnerCoach Working Group or WIENFLUSS to expend the resources necessary to correct the system's behaviour. As a final consideration, both S and K could pool their influence into a form of *collective agency* to make their case together, improving the overall chances of success.

Finally, to answer the last of the secondary questions for step four, technical limitations may play a role in whether or not change is possible. As many complex algorithmic systems often suffer from *low cohesion* and *high coupling* [316], addressing the issue might be technically unfeasible without a significant redesign. Consequentially, both S and K's agency to effect change may be impacted by factors entirely outside their sphere of influence and, indeed, knowledge.

Summary Assessment

Explicating the four steps brings into focus the most pressing issues in this accountability process. First and foremost, low *system-level transparency* and lack of *ex-post explainability* features have been shown to significantly impact both the *forum* and *actor's* agency throughout the process. Determining whether or not the issue was the result of a software error or a data entry error may require the expertise of the system's developers, expanding and complicating the accountability process. Domain knowledge and algorithmic literacy of the *forum* and *actor* also play an important role, but can vary strongly between different potential *fora* (i.e., energy consultants, energy auditors, community users, building managers) and *actors* for this scenario.

Consequentially, improvements of ex-post explainability should be designed *for and with* the *actor* and the *forum*, to best adapt them to their needs and account for their levels of domain knowledge and algorithmic literacy. The scenario purposefully omitted the "*ground truth*" of the (potential) error to showcase the framework's usefulness not just for retroactive analysis, but specifically for hypothetical, future scenarios, in which the specifics of the underlying issue may not be known. For the sake of the argument, however, let us assume that the observed behaviour was neither a software error nor a data entry error, and that the explanation for the system's output may lie in an anomalous usage pattern of the community buildings in question. For instance, the community may have built a new wastewater treatment plant in 2021, which would have

necessitated turning off the water supply for some buildings in the region for a while and thus resulted in the anomalously low water consumption for the same year. The feature of contextual annotations implemented for the EnerCoach system as a consequence of the participatory design workshop (see Section 3.2.3 and Figure 4.16) could, in this case, improve the *forum*'s agency by adding an additional source of information that may have explained the anomalous results. The existence of an annotation explaining the low water consumptions could have either made the accountability process superfluous from the start, or at least offer a way of cross-referencing the account provided by the *actor* (i.e., S, the EnerCoach hotline staffer).

Finally, assessing the scenario through the A³ framework brought to light various factors reducing the chances for effecting change in the system in case of an error. The economic and technical limitations for such a change are certainly cause for concern, as they may prohibit addressing an issue with the system's behaviour even if it was identified as unjustified by both *forum* and *actor*. Improving the situation in regards to the potential for change would certainly be a complex undertaking, but possible solutions exist nonetheless. For instance, one might reserve a set amount of budget per year for fixing such software errors, or establish a well-documented process of how change requests and bugfixes can be made by users, and how they should be prioritized by the EnerCoach Working Group. Both of these suggestions are clearly situated in the social and socio-technical realm of potential measures to improve the system's accountability, showcasing the importance of avoiding a solely techno-centric problem-solving strategy that might, in this case, not produce the desired outcomes.

6.2.3.2 AMAS

In our previous work on AMAS [3, 4], we have already identified algorithmic accountability as problematic for a variety of reasons. Both *macro-* and *micro-accountability* are highly desirable for such a system. On the larger scale of macro-accountability, *political*, *legal*, *administrative* and *social accountability* are all relevant, and some of their requisite accountability processes have already taken place throughout the system's history. On the *political* level, the decision to create the system, its overall goals and aim, the funds necessary to create it, and its adherence to the larger labour market policies guiding the AMS's management are subject to scrutiny from Austria's governmental institutions (e.g., the various federal ministries or the Austrian parliament) as well as the AMS's internal governance in the form of its board of directors. *Legally*, the decision handed down by the DSB forbidding the use of the system due to its lack of legal foundation is an ongoing accountability process playing out in the federal administrative courts at the moment. In terms of *administrative* accountability, the internal departments of the AMS governing the operationalization of the system, would have required the AMS caseworkers to account for their conduct with AMAS, for instance by monitoring how often they chose to disagree with the system's assessment. Finally, *social accountability* encompasses the controversial public discourse that arose after the first announcement of the system, which includes the critique voiced by academic experts and researchers and

resulted, among others, in the research project that became the foundation for this case study.

An analysis of these macro-accountability processes as *public accountability* is certainly relevant and would spotlight, among others, shortcomings in terms of public communication and system-level transparency as hindering factors to a substantial and positive macro-accountability process. The importance of this perspective notwithstanding, the *micro-accountability* perspective is, arguably, even more pressing to address. First, the discussion of how jobseekers can express their agency to hold the system to account is crucial insofar as the system's impact on their lives would be immediate, and the consequences of a lack of accountability would likely be particularly problematic for those already affected by systemic disadvantages. Secondly, the slow nature of macro-accountability processes means that effecting change—be it adaptations to the system itself, the policies governing its operationalization, or even stopping its use as effected by the DSB's ruling—may have come too late for those jobseekers that already had to contend with the system's classification of their purported chances on the labour market. Finally, considering micro-accountability processes earlier in the system's development may have already established safeguards and changed the systems trajectory in a more socially acceptable direction, pre-empting some of the contentiousness of the larger macro-accountability processes.

Given these considerations, the following scenario is purposefully situated in the realm of micro-accountability as the process occurring between a *jobseeker* as the *forum* and a *caseworker* as the *actor*. The scenario is based on a hypothetical persona my colleagues and I created with the goal of illustrating the potential of the system for technical bias and subsequent discrimination as part of our case study [4]. To illustrate these issues and how these technical biases are embedded, a section of additional background information about the scenario was included the following description.

Scenario *Surprising Misclassification*

“Mrs. J is recently unemployed and has just arrived for her initial consultation with Ms. T, her caseworker at the AMS. After going through her biography and supplying the data points her caseworker asked of her, she is being informed by the caseworker that she was classified by the system as “CAMN” or having low chances on the labour market. She is very surprised by this assessment, and demands to know why the system assessed her as such, and wants to know what this assessment will mean for her future options with the AMS.”

Forum Mrs. J, Jobseeker

“Mrs. J is an Iranian citizen and migrated to Austria 3 years ago to study Computer Science at the TU Wien and finish her master’s in Data Science. During this time, she was legally employed at the university as a part-time assistant. She has just finished her degree, and is looking for a job in the IT sector as a data scientist. Her actual chances to find employment are very high: while her German language skills are still somewhat limited, she speaks English and Farsi natively, and most companies she had looked at require only English as a language skill. As she has recently married an Austrian citizen and has attained the requisite Visa granting her full access to the Austrian labour market, she is highly motivated to find a job and has already identified a number of fitting, open positions she plans to apply for.”

Actor Ms. T, AMS Caseworker

*“Ms. T has been working as an AMS caseworker for four years. During this time, she has adapted well to the numerous changes in policy and rules governing her work, and the newest addition of the **AMAS** system was described to her as a helpful tool to deal with her significant caseload. She is currently responsible for 154 cases, and has very little time to spend with each jobseeker. Consequently, she is grateful to be able to rely on the system’s assessments to speed up the consultation process.”*

Scenario Background *Technical Bias*

*“Mrs. J’s qualifications and real chances on the labour market are simply not covered by the system’s variables and coarse data points. Her constellation categorizes her as a woman with “migration background”, non-EU citizenship, living in the least favourable district in terms of the **Regional labour market variable**, seeking work in the “Service” sector, and her data is considered only partially valid, since she has only been working in Austria for three of the four required years. Many of the other members of her constellation indeed struggled to find a job in the previous four years, as most of them are not as highly educated and are also not looking for work in the booming IT sector. Based on this comparison, the system wrongly classifies her as having low chances.”*

Step 1: Requesting information

The first step, *requesting information*, already requires a number of preconditions to be fulfilled in order to occur. First and foremost, the jobseeker must be informed about the system’s existence and be able to distinguish between the caseworkers personal assessment

and the one generated by [AMAS](#). As the procedural rules for consultations include this information, caseworkers should inform the jobseekers of their classification, and that the assessment was performed by an algorithm. Secondly, the jobseekers agency to request this information is predicated on the fact that they are either particularly interested in their assessment, or more specifically, that they disagree with it and suspect a mistake was made. While many jobseekers may be surprised or disagree with the assessment made by the system, their level of algorithmic literacy may determine whether or not they fall prone to *automation bias* and accept the system’s authority over trusting their own assessment, prompting them not to initiate the accountability process at all. Finally, the intimidating nature of the consultation setting and the jobseekers dependency on the AMS for financial and other support may severely impact their their agency to demand such a justification, particularly if they are worried such a request might result in being treated worse.

In this scenario, Mrs. J, the jobseeker, (1) is aware of the systems existence, (2) happens to have domain knowledge in Data Science due to her educational background, and (3) is also suspicious of a classification that conflicts strongly with her own expectations and assessment. To answer the secondary questions, we also see that the context of this accountability process makes identifying Ms. T as the *actor* trivial, as the process plays out verbally as part of the consultation meeting. However, due to the potential language barrier between Mrs. J’s and her caseworker, her agency as a *forum* to formulate the request and voice doubts about her classification may be limited by either her German skills, or Ms. T’s English skills and willingness to communicate in English. In summary, the most relevant, limiting factors include knowledge about the existence of the system, the requisite domain knowledge required to make the initial request for an account, and potential communicative barriers between the jobseeker and caseworker.

Step 2: Providing the account

Having been asked about the system’s classification, the *actor* must provide an account and justify the system’s behaviour, i.e., the resulting [IC](#) values and classification. Their agency to do so is clearly limited by the system’s design, available features and documentation. Primarily, they can answer the jobseeker’s questions based on the the guideline document “*Answering client questions*” [\[RICHT_2\]](#), which provide only very broad information and address *system-level transparency* first and foremost. Based on this document, the caseworker as the *actor* can only provide a very limited account; depending on the caseworker’s experience, domain knowledge and whether or not they attended the AMS’s training workshops for the system, they may have additional understanding about the system they can add to the account.

For specific questions about the result—i.e., to satisfy requirements towards *ex-post explainability*—the caseworkers must turn to the *explanation texts* (see Section [5.3.4.1](#)). As I have argued previously, these texts are extremely reductive, and it is highly questionable whether or not they constitute an acceptable account for most jobseekers as a *forum*.

To exemplify this issue, let us consider this text fragment identifying *significant gaps in employment* as a ‘challenging factor’:

“*Your were only employed over limited periods of time during the last few years.*”

See Appendix [A.4](#)

Neither does the text given here define that “*limited periods*” refers specifically to fewer than 75% of days of employment during the previous 4 years, nor does it mention the fact that this text is only applied for jobseekers with a period of unemployment of 12 months or longer during their current case. While these additional pieces of information are theoretically available to the caseworkers in the [AMAS](#) handbook [\[HAND_1\]](#), it seems unlikely that they would have this information memorized or on hand for each of the explanatory text fragments.

A further limitation to the caseworker’s agency to provide a meaningful *account* for a specific result is the fact that the explanation texts are only available for jobseekers classified either as having *low* or *high* chances, and of those only for the jobseekers with *fully valid data*. In other words, for more than half or roughly 58% of all jobseekers as *fora* asking about a justification for their results, these explanation texts are simply not available, leaving the caseworkers without a basis upon which to formulate their account. This fact alone may impact their agency through affecting their *belief in self-efficacy* [\[24\]](#): since the primary interface detailing the jobseekers file does not clarify which subpopulation they belong to, and since the explanatory texts are only available in a separate interface, caseworkers may learn over time how limited the explanation system is, and thus avoid even offering explanations for the result if they do not believe they have the agency to give them.

The final, and arguably most severe factor limiting the caseworker’s agency to provide a meaningful *account*, are the time-restraints they face during their interaction with the jobseeker. The system is supposed to make the consultation process more *efficient*, and this goal was communicated to the caseworkers in every internal document, fact sheet, guideline or handbook concerned with [AMAS](#). Taking the time to look up additional information or even request assistance in providing the account as a form of utilizing *proxy agency* seems quite contradictory to the overarching mission narrative of speeding up the consultation process. Consequently, it stands to reason that spending additional time to provide a better account is simply not possible for most cases.

In our concrete scenario, Ms. T has no explanatory texts to work with, as Mrs. J is part of the “*partially valid assessable population with migration background*”. Her agency to provide an account, thus, is limited to the information in [\[RICHT_2\]](#), and addresses only *system-level transparency* questions, but offers no help in terms of *ex-post explainability*. Considering the information provided in this document, including the system description and list of variables used to create the constellation, it is not implausible that Mrs. J, as a

data scientist, may have specific questions that the caseworker simply will not be able to answer. In this concrete example, we can observe the challenges of a micro-accountability process where a mismatch between *forum* and *actor* domain knowledge and algorithmic literacy. While most cases of such a mismatch will be in favour of the *actor*, rather than the *forum*, in terms of prior knowledge, the unique combinations of actors and fora in the context of the **AMAS** system illustrates that the opposite situation can occur as well, with no less severe impact on the accountability process.

Step 3: Imposing consequences

Assuming both previous steps have been completed successfully, the agency of the forum to impose consequences on the actor requires a differentiated analysis based on the type of possible consequences available. Most immediately, a jobseeker dissatisfied with both the classification given by the system and the explanation provided by the actor might intend to have their classification changed. As described in Section 5.3.4, the system's automatically generated **CAM** value is unchangeable, leaving the change of the caseworker's assessment in the form of the **BAM** value as the only viable option. The consultation process does include this, but a power imbalance exists between caseworker and jobseeker. At best, in a discussion of the result, both agree on a new classification—but in the worst case, the caseworker makes their assessment and notes down the jobseekers dissent as part of their file. Consequently, the jobseeker's agency to impose this particular form of consequence is primarily determined by the caseworkers willingness to cooperate.

Beyond this immediate consequence of having their assessment adapted, the jobseeker may seek to lodge a complaint and escalate the issue further. In our case study, we could not determine clearly what options the jobseekers have for such an escalation, but some documents and statements indicate the existence of an AMS ombudsperson for jobseekers. This person would be tasked with mediating in all kinds of conflicts, not just those arising from the use of **AMAS**. Here, the jobseeker's agency is limited by their knowledge of this person's existence, and their willingness to go through the process of making the complaint. At the consultation, gaining this knowledge and stating their wish to make a complaint seems implausible for most cases, due to the jobseekers dependency on the caseworkers goodwill in future. Even if they do make their complaint, it is unclear what, if any, potential for lasting change they might have: after all, the system provides almost no information that they could use to support their case (e.g., to prove that they have been wrongly classified). Considering that roughly 15% of jobseekers would be misclassified per year, it is also completely absurd to assume that even a tenth of them could go through such an arbitration process, as that would require 7500 or more of such processes to play out.

For Mrs. J, the jobseeker in this scenario, it seems extremely unlikely that she might be able to impose these consequences. Although she is, indeed, a domain expert in Data Science, her limited language skills in German may already make communicating with the caseworker challenging; given the fact that there are no explanations available for her classification, she also has no information to support her assessment of a misclassification

beyond the fact that she has already identified some potential job opportunities and feels well-prepared for the interviews. Considering the secondary questions, we see that Mrs. J simply has no plausible account to work with, and thus would most likely have to assess the justification as insufficient. Depending on Ms. T's goodwill and understanding, she might still be able to argue for a change of her classification from *low* to *medium* by virtue of her prior, general knowledge about the fallibility of data science for individual assessments.

Step 4: Effecting Change

Finally, the possibility of effecting sustainable change in this micro-accountability process is curtailed by the same factors that limit the imposition of consequences as outlined in the previous step. If the caseworker in question is not willing to change the classification, the jobseekers face dwindling options for further action. Short of investing significant time and resources to initiate an arbitration process with the AMS's ombudsperson, and taking the risk of being labelled as a troublemaker by their caseworker, there is little a jobseeker can do to effect lasting change. Given the fact that the system's classification is continuously re-calculated for each milestone in their period of unemployment, the system also reinforces its decision with every consultation they have to attend. Considering the secondary question "*What influence can the forum exercise on the algorithmic system?*", we also see that the jobseekers have almost no impact on their own future classifications. Most of the variables governing their assignment to a subpopulation and constellation depend on factors outside their sphere of influence: neither could Mrs. J change her Age or Gender, her Employment history, Health impairments or Responsibilities of care. The only options to impact her constellation would be to either move to a location more favourably rated by the Regional labour market variable, or start looking for a job in the "Production" sector instead of in the "Service" sector. Seeing as the account she received from Ms. T in Step 2 offers no guidance on any of these factors, she has no *agency* to make such a choice in an informed way, even if she was willing to go to these lengths just to influence her classification. To truly initiate a process of change that transcends their own, personal case, jobseekers would face even more prohibitive obstacles, as they would have to enlist the help of other agents such as civil society watchdog organisations or NGOs willing to take their case in order to effect greater change.

From the point of view of the caseworker, effecting change is hindered by their own scarce resources. While internal processes to collect feedback on the system's usefulness for caseworkers may well exist, none of the documents analysed provide any detail on how these processes would be implemented. Adapting either the technical implementation of **AMAS** or the way the system is used would require massive efforts and would likely face steep opposition by AMS management and the political stakeholders governing it, given the investment made in the system. As the strong public opposition to the system showed, the AMS's willingness to respond to its critics with constructive changes is limited—in the end, *legal accountability* in the form of the **DSB**'s ruling was the only successful intervention in regards to the system. Short of an organised resistance of caseworkers

or more subversive tactics (e.g., consistently ignoring the system's output, or lengthy reclassification explanations taking up more time than plausibly saved by the use of the system), the potential for effecting lasting change can only be described as minimal.

Summary Assessment

The analysis of the four constituent parts of the accountability process with the help of the A³ framework foregrounds the significant shortcomings regarding micro-accountability of the AMAS system and its operationalization. Even for a scenario like this one, in which the jobseeker has a significant advantage over her peers in terms of domain knowledge and education, the system imposes technical and procedural limitations that make a positive outcome all but impossible. For less privileged jobseekers, even requesting information may already be an unattainable goal due to their lack of algorithmic literacy. Furthermore, the strict rules governing how the caseworkers must interact with the system, and the technical limitations of the explanation texts put the jobseekers at a significant disadvantage throughout the entire process.

For the caseworkers, the situation looks no less dire. Not only are they limited in their agency by the same technical limitations, exacerbated by their own limited algorithmic literacy and lacking documentation, their institutional mandate to make the consultation process faster and more efficient further curtails their chances to participate in this micro-accountability process in a meaningful way. Even for a motivated caseworker willing to engage in the process, the system offers preciously few resources to help them along the way: if the explanation texts are not available, the caseworkers are left to their own devices (e.g., their own experience or interpretation) to explain the system's behaviour. The fact that this reliance on caseworker experience is the source of precisely the personal biases that the system is supposed to eliminate is, indeed, not without irony.

While the hindering factors for step one and two of this process are already worryingly severe, the process collapses almost entirely when it comes to the question of consequences and change. The significant power imbalance between *actor* and *forum* dictates what limited actions jobseeker can take, and the political decision-making governing the system's implementation and operationalization restricts the potential for transforming this already precarious micro-accountability process into the larger context of macro-accountability.

As an additional point, it is worth noting that the specific scenario for this analysis is based on a previous persona and scenario we described in our final research project report [4]. In this scenario, we foresaw that the caseworker would be unwilling to reclassify Mrs. J from *low* to *high*, but would compromise by classifying her as having *medium* chances. Mrs. J, consequentially, was questioning her own assessment, and agreed to take an intensive German language course paid for by the AMS, to better her chances of finding a job. As a consequence, not only did she use AMS resources unnecessarily, but also delayed her job search because of the time investment necessary to complete the course. This also resulted in a higher cost to the federal government due to her longer reliance on

unemployment benefits, illustrating how the **AMAS** system might fail in its overall goals of more efficacy in money spent on supportive measures. These hypothetical consequences further underscore the potential value of a meaningful, successful and impactful (in terms of consequences and change) accountability process, in order to identify and remedy similar situations may determine the overall success or failure of the system as a whole.

6.2.3.3 Comparative Results

Looking at the two case study scenarios side-by-side as evaluated with the **A³ framework**, the guiding questions helped highlight the most crucial deficits throughout the four steps in the accountability processes.

For EnerCoach, we saw a lack of system transparency and ex-post explainability features, as well as limited algorithmic literacy as the most hindering factors in the first two steps, despite a generally high level of domain knowledge of both *actor* and *forum*. Additionally, the communication between *forum* and *actor* presented some additional, but not insurmountable, challenges that were primarily addressed through social measures (e.g., the training workshops) rather than technical ones. In the case of **AMAS**, algorithmic literacy also plays a role, but is entirely overshadowed by a lack of supporting documentation to provide *system-level transparency* and the absence of reliable technical features to guarantee *ex-post explainability* for the first two steps of the process. Furthermore, the **AMAS** socio-technical assemblage also suffers from severe power imbalances between *actor* and *forum*, with the forum being essentially dependent of the actor's goodwill to participate in a meaningful accountability process—a fact that stands in stark contrast to the **AMS**'s new public management narrative of being a service provider to its clients.

For the second part of the accountability process—imposing consequences and effecting change—both case studies exemplified the difficulties in identifying the relevant stakeholder groups and actors to impose these consequences on, and the considerable, perhaps even prohibitive, efforts it may take to effect lasting change in algorithmic systems. Both EnerCoach and **AMAS** are complex, socio-technical assemblages enacted, governed, influenced and utilized by a variety of stakeholders. The knowledge required to identify whom to address in order to adapt the system or remedy shortcomings represents a significant threshold separating those with the power to effect change from those that simply can not. The complex nature of the EnerCoach system makes adaptations potentially costly, and the limited funds available require careful triage of *which* problems to address first, and *how* to do so. Conversely, the political pressure for **AMAS** to succeed and the ideological underpinnings of the *activation paradigm* makes a micro-accountability process that results even in the slightest adaptation of the system or its operationalization almost impossible. Finally, while the analysis of both case studies shows the diversity of the various potential *fora*, it also shines the spotlight on the importance of considering the overall impact and potential harms for the affected groups. While a failed accountability process in the EnerCoach system may, overall, harm the success of the initiative to support sustainable energy practices of communities overall, individual stakeholders are unlikely to be directly affected by the system's shortcomings beyond wasting time and

resources. In contrast, for the jobseeker as the *forum*, their own personal future may be at stake, with multiple, possibly severe negative impacts on their lives as a consequence of the system’s assessment and the failure a subsequent micro-accountability process. Considering the fact that many jobseekers are members of particularly vulnerable groups and in danger of intersectional discrimination and cumulative disadvantages further underscores the importance of a plausible, functioning micro-accountability process to address their needs and concerns when and where it matters, instead of lengthy and drawn out macro-accountability processes.

Methodological evaluation of interventions

The final perspective emerging from the previous analysis concerns the question of how the shortcomings in these accountability processes revealed by the [A³ framework](#) *should* be addressed. In this regard, I posit, EnerCoach should be considered a positive example, and the [AMAS](#) system should serve as a showcase of misalignment between the various stakeholder interests and needs resulting in sub-par accountability capabilities.

EnerCoach’s severe challenges in terms of system-level transparency and ex-post explainability were addressed by and through a collaborative and participative effort, involving stakeholders affected by these shortcoming directly. Through this process, differences in domain knowledge, algorithmic literacy, and different needs and requirements for participating users, energy consultants, hotline staffers, the EnerCoach Working Group and even the system developers were made clear. Solutions were found through the application of design studies methodologies and a negotiation process between the involved parties, resulting in compromises that proved effective and useful additions to the system yet did not require unrealistic investment or implementation efforts. The implementation of the contextual annotations feature, in particular, directly addressed an entire class of similar explainability challenges that would make some micro-accountability processes obsolete entirely. Finally, the methodological approach to these interventions also helped reveal the non-technical options for supportive measures in the form of the process visualizations as an additional reference for the hotline staffers, thus avoiding a techno-deterministic or solutionist dead end that often carries the danger of implementing unusable or misguided features.

In contrast, the [AMAS](#) development process addressed these issues from an entirely different perspective. On the one hand, the involved stakeholders—namely, project managers at the AMS, and representatives of the implementing company Synthesis Research GmbH—identified the need for transparency in the form of ex-post explainability early on in the conceptualization and development process. On the other hand, the timeline of meeting notes and protocols ([\[PROT_1_JF\]](#) through [\[PROT_22_JF\]](#)) reveals that the explanation texts and the rule-based conditions system determining their applicability to a given jobseeker was never prioritized and addressed very late in the process, seemingly as an afterthought. Furthermore, the conceptualization and implementation of this feature was done solely with the *actor*, i.e., the caseworker in mind. These documents give no indication whatsoever that the agency of the *forum* to

request or process this account played any significant role in the design of this feature. As the analysis through the [A³ framework](#) showed, however, even the usefulness for the *actor* in this accountability process is highly doubtful, given the technical and procedural limitations of this feature. Many of the hindrances to both *actor* and *forum* brought to light by this analysis would, most likely, have been blatantly obvious to any of the involved stakeholders, had they been more closely involved in the development process. It seems entirely plausible that any caseworker could have identified how problematic it is that the explanations would be available for less than half of the jobseekers; similarly, even a small, representative sample of jobseekers invited to a mock consultation in which caseworkers explain the system's functionality and results would have revealed the obvious deficits in *system-level transparency* and *ex-post explainability* in light of their limited domain knowledge, limited algorithmic literacy and immense pressure of completing this process within the available time frame of roughly 15 to 20 minutes.

In summary, the measures designed to address system aspects related to the accountability processes enacted therein are the result of very different processes and stakeholder involvement. While the participative methodologies used for the EnerCoach case study elevated both *actor* and *forum* to that of a *critical audience* [28] and thus improved their agency, the technologically and procedurally determined measures for the [AMAS](#) system were aimed almost entirely at the *actor*, and—failing to include them in the process—proved to be insufficient even for them, let alone the *forum*.

6.2.4 Evaluation in Context with Other Frameworks

Given the rising concern about accountability deficits in algorithmic systems voiced by academic scholars and practitioners alike, it comes as no surprise that various other frameworks have been proposed to assess and evaluate algorithmic systems and technologies in terms of their accountability. To evaluate the usefulness of the [A³ framework](#) proposed in this dissertation, the following section situates it within the larger context of accountability frameworks and argues for its relevance and advantages.

Frameworks for Accountability as Virtue or Mechanism

As outlined before, two general interpretations of accountability can be discerned: Accountability as a *virtue* and as a *mechanism* [156]. While an academic discourse on the former is certainly an important contribution to our overall understanding of the issue and can help elucidate the variety of meanings the term can take on for different contexts (e.g., [155, 366]), the resulting insights contribute very little in terms of actionable, concrete suggestions for the accountability process as a relational construct between *actor* and *forum*. Frameworks and guidelines touching upon accountability as a virtue are becoming increasingly ubiquitous (e.g., [168, 167] as an example of *professional accountability*, but remain largely toothless [169] in light of their need to be broadly and generally applicable throughout a professional domain.

For the latter approach of accountability as a *mechanism* that follows a specific, relational process, more appropriate examples of frameworks exist. The *Accountability Cube* previously mentioned in Section 2.4.2.3 as proposed by Brandsma and Schillemans is the most relevant comparable approach, and served as a contributing conceptualization for the *A³ framework* as well. As a *quantitative assessment tool*, specifically geared towards a ranked comparison of different systems, the *Accountability Cube* offers excellent insights into deficits and problematic aspects of the process, but is lacking in its ability to reveal the causal contributing factors for this assessment. In other words, it is a useful tool to identify *where* the deficits of accountability processes lie, but offers little to explain *why* these deficits manifest. Furthermore, the original application domain for the *Accountability Cube* lies outside the realm of *algorithmic accountability*, and puts the focus on “[...] empirically informed normative judgments on the state of accountability of, for instance, networks, public bodies, or international organizations such as the EU.” [196, p.2] While this definition does not necessarily exclude (semi-)automated decision-making processes, it also does not specifically cater to the requirements and contextual idiosyncrasies of accountability processes involving algorithmic systems.

Macroscopic Algorithmic Accountability Frameworks

In the specific realm of *algorithmic accountability*, Tagiou et al. [354] present a more fine-grained assessment framework focused on the two dimensions of *organisational* and *algorithmic* issues related to accountability. On the organisational side, their framework considers accountability performance in terms of *responsibility*, *explainability*, *auditability*, *accuracy* and *fairness*; on the algorithmic side, they evaluate the disclosure of the *algorithmic presence*, its *data*, *model*, *inferencing* characteristics and *performance*. For each of these aspects, they offer indicative questions, whose qualitative evaluation should lead to a quantitative assessments. The framework’s guiding questions cover some of the relevant, overall aspects related to transparency, explainability, moral responsibility and technical aspects of the systems, but are somewhat confusingly formulated: some questions seem to be clearly addressed at representatives of the organisation employing the algorithm, while others are more general in nature and may be answered from an outside observers perspective as well. Finally, as their approach makes no use of the procedural conceptualization of accountability as presented by Bovens [22], the applicability of the framework to evaluate the concrete accountability processes that may result from the use of the system is extremely limited. This seems intentional, as the authors describe their framework’s value as that of “*check-lists providing a set of best-practices to organizations in order to cater for accountable algorithmic systems at an early stage of their creation.*” [354, p.1]. In summary, while Tagiou et al. present a useful self-assessment tool, the prioritisation of quantitative over qualitative evaluation and the vague definition of accountability, combined with the somewhat convoluted nature of their evaluation criteria limit its usefulness for a holistic, external evaluation that includes stakeholders outside the responsible organisation.

What both of these frameworks exemplify is the larger trend towards the macroscopic perspective on accountability as a process or characteristic of organisations and their policies. In a similar vein, Busuioc [367] identifies largely systemic challenges to algorithmic accountability specific to AI technologies, such as prevalent *information asymmetries*, *inherent opaqueness* of deep learning techniques or the *behavioural effects* of algorithm outputs on human decision-making, but provides no assessment or evaluation framework. Although they acknowledge Bovens' [22] conceptual framing of the accountability process, they explicate their perspective as *systemic* and “[...] do not aim to zoom in on specific microlevel (actor–forum) relationships, but rather set out to capture a bird’s-eye view of the challenges that arise along the three phases.” [367, p.828]. Similarly, Buhmann et al. [365] provide a framework aimed at managing algorithmic accountability. Their framework considers the three dimensions of *reputational concerns*, *engagement strategies* and *discourse principles*, and prescribes four *discourse-ethical principles*—*participation*, *comprehension*, *multivocality*, and *responsiveness*—as guidance to improve algorithmic accountability. Of these four, *participation* highlights the importance of providing all affected stakeholders access to a *forum* as deliberative, institutional setting in which the accountability process takes place. *Comprehension* describes the precondition of having the requisite access to information in order to take part in a meaningful discussion, and *multivocality* expands on this by demanding that not just all stakeholders can participate in an informed way, but are given the chance to be heard and their arguments considered equally. Finally, *responsiveness* corresponds most closely to the question of consequences and effecting change: after all, as Buhmann et al. formulate it, “all three [prior] principles are meaningless if the different concerns and suggestions [...] cannot influence actual recommendations or decision as a result of the discourse” [365, p.275].

Both of these contributions serve well as examples of organisation-level frameworks to assess and conceptualize algorithmic accountability challenges and solutions on the macroscopic level of institutional processes. The suggestions for interventions they offer to improve the situation are more concrete than the previously described quantitative assessment frameworks, but are simply not applicable to the immediate and specific micro-accountability processes that occur whenever humans interact with algorithmic systems directly. This is not surprising, as Buhmann et al. conceptualize accountability mechanisms as “*institutional arrangement*” [365, p.269] rather than socio-technical measures to support human-centred relational accountability processes.

Frameworks for Legal Accountability of Algorithms

Finally, a number of contributions towards framing algorithmic accountability in a *legal* sense exist. As a consequence of the differences in legal systems across the world, these only offer limited applicability beyond the confines of the national stages they are designed for, and a comparative discussion of their approaches transcends both the scope of this dissertation and, indeed, my own expertise. Just to name an example (admittedly, only from my perspective as a layperson in the domain of law and jurisprudence): Engstrom and Ho [368] take a closer look on algorithmic accountability in the administrative state

and suggest oversight boards as a concrete measure to improve the (legal) situation specifically in the context of the United States of America. As a matter of course, legal frameworks play an important role in algorithmic macro-accountability, and the omission of a deeper analysis in this dissertation should not be interpreted as an argument against their necessity. However, for the specific challenges that stakeholders of algorithmic systems face, their applicability may vary greatly depending on the context of use. While algorithmic technologies used in governmental or administrative contexts can surely profit from the introduction of such frameworks, many private enterprises may struggle to incorporate processes and guidelines aimed at the bureaucratic apparatus of a state or government. Finally, legal frameworks also must necessarily remain on the macroscopic level of accountability in order to retain their wide applicability; as such, they can offer only very little specific suggestions to help organisations implement measures to support micro-accountability processes.

Advantages of the A³ Framework

Considering these observations on the state of the art of modern accountability frameworks, the [A³ framework](#) offers a number of distinct advantages. First and foremost, the conceptual lens of *agency*—be it *individual*, *proxy* or *collective*—allows for a wide applicability to both macro- and micro-accountability processes. While the scenarios presented previously highlight micro-accountability processes specifically to underscore their importance, the framework’s design is equally suited to assess macro-accountability processes between *collective fora* and *actors*. To explicate this claim, simply consider the process of *social accountability* for the [AMAS](#) case study exemplified by our own research on the topic in light of the [A³ framework](#)’s guiding questions. Here, the limited, contradictory and misleading initial documentation provided by Synthesis Research GmbH as a form of system-level transparency would immediately be identified as a hindering factor prohibiting our research team as the *forum* from making informed requests for information.

Secondly, the [A³ framework](#) purposefully avoids quantitative evaluation in favour of qualitative approaches. This approach provides an advantage over quantitative frameworks insofar as it supports not only the identification of deficiencies in the accountability process, but also implicitly encourages critical thought processes that offer a chance at discovering possible measures to improve both *forum* and *actor* agency throughout the process. This critical discourse is both fostered and given structure by the open nature of the questions over the course of the four steps in the accountability process. Thus, the framework supports a discourse between all stakeholders of the system, be they internal or external to the parent organisation, and the clear and comprehensible nature of the questions will allow people of various levels of organisational and domain knowledge as well as those with limited algorithmic literacy to contribute to the discussion in a participative way.

Finally, the [A³ framework](#) should not be seen as a replacement for other quantitative or macroscopic frameworks, but rather as a complementary resource: a widely applicable,

analytic lens that encourages a critical exploration of a given algorithmic system’s accountability processes.

6.2.4.1 Human Agency and Activity Theoretical HCI

Building on Bandura’s *Social Cognitive Theory* and modelling algorithmic accountability processes through the lens of human *emergent interactive agency* presents certain overlaps, similarities and contradictions with established theoretical foundations in HCI, specifically Bødker and Klokmoose’s [369] *Human-Artifact Model*⁴ based on *activity theory*. To better situate the theoretical approach of the A³ framework specifically, and this dissertation in general, within and adjacent to HCI, the following excursus provides a very brief introduction to the *Human-Artifact Model*, outlines similarities and differences between the theoretical foundations for the A³ framework and *activity theoretical* approaches, and discusses the applicability of this model to the algorithmic accountability process.

Bødker and Klokmoose’s *Human-Artifact Model* [369] is founded in a *dialectical* (as opposed to *causal*) understanding of the interaction between *users* and *artefact ecologies*, i.e., the interactive computing devices, interfaces and (digital) technologies that *mediate* user’s activities. Building on Leont’ev’s [370, 371] hierarchy of human *activity*, *action* and *operation* to answer the guiding questions of (1) *why* humans are motivated to perform an *activity*, (2) *what* the *actions* determined by their conscious goals are, and (3) *how* these actions can be performed as a set of *operations*.

Bødker and Klokmoose position their model as a theoretical framework to respond to the need for a generalizable conceptualization of these interactions. Although they recognize the value of ethnomethodological approaches—they credit, for instance, Suchman’s [372] contributions to HCI as “*eye-opener to many in the field*” [369, p.317]—they also point to the importance of a theoretical framework in HCI as a means to “*continuously avoid going back to specific, detailed accounts of particular cases*” (Bødker and Klokmoose [369, p.318], summarizing [373]).

Central to the application of activity theory to HCI is the understanding of the human-artefact relationship not as a *subject-object* or *subject-subject* relationship, but rather as a *subject-mediator* relationship, or, as Bødker and Klokmoose put it most succinctly, “[*to see the computer as something that the user acts through, on other objects or with other subjects*”⁵ [369, p.321]. In this view, the ideal artefact is one that becomes a “*functional organ*” [369, p.324] for a given activity, e.g., cutlery used by humans in the activity of eating food, and is, at this point, considered part of human beings. From the opposite perspective, less-than-ideal artefacts can cause “[*b*]reakdowns [...] due to either insufficient capacities or possibilities in the artefact, or lack of available action possibilities, either culturally or in the individual repertoire of action possibilities” [369,

⁴Bødker and Klokmoose’s model uses the American English spelling of ‘artefact’, which is reproduced here to retain the accurate reference to their work. Throughout this section, ‘artefact’ as the British English variant of the term, is used interchangeably when not directly referencing or quoting their work.

⁵Emphasis by original authors.

ibd.], effectively restricting or prohibiting the successful completion of the overall *activity*, individual *action* or specific *operation*. Combining these two characteristics, artefacts, on the one hand, are shaped by or—as Leon’tev [371] formulates it—*crystallize* activity, and, on the other hand, shape human activity themselves.

| | | |
|--------------|--|--|
| Why? | Motivational aspects | Motivational orientation |
| What? | Instrumental aspects | Goal orientation |
| How? | Operational aspects – Handling aspects | Operational orientation – Learned Handling |
| | – Adaptive aspects | – Adaptation |
| | Artifact | Human |

Figure 6.4: Bødker and Klokrose’s illustration of the *Human-Artifact Model*, characterizing the human-artefact relationship for the three levels of *activity*, *action* and *operation* or the *why*, *what* and *how* of human activities. [369, p.333]

Bødker and Klokrose [369, pp. 333-337] then synthesize this characterisation of computing artefacts and their relationship to human activity into their *Human Artefact Model* illustrated in Figure 6.4. Each of the fields in their model represents a specific perspective of analysis for a given artefact-mediated activity, and can be considered separately. Starting from the top, the *motivational aspects* of the artefact include the reasons why humans would choose to use a specific artefact over another, whereas the *motivational orientation*

on the human side refers to the reasons why humans want to perform the activity overall. The second level of activity—the *action*, or the *what*—specifically describes the action to be taken based on the overall goal of the activity. The third layer describes the specific operations necessary to achieve the goal on the artefact side, and the necessary knowledge or *operational orientation* to perform these operations on the human side, as well as how well the artefact is adapted to the operations, and which adaptation humans have to conform to in order to operate through the artefact.

By applying the model to specific activities, a side-by-side comparison of different artefacts as functional organs for the same activity can help identify advantages and disadvantages of one artefact over the other. Likewise, for a single artefact and activity, mismatches between the left and right side of the model highlight potential points of *breakdown*, where expectations, goals or operational aspects do not align between human and artefact. Bødker and Klokmoose [369, pp. 336] also point to the versatility of the model, depending on the starting point of an analysis either on the human or artefact side: If the analysis starts with the human side, the model helps structure human practice along the levels of activity in order to match a current or future artefact to said practice. Complementarily, if the analysis starts on the artefact side, the particular shortcomings of an artefact for a variety of human activities, actions or operations can be identified, and may inform an ongoing design process for said artefact.

Applicability to the A³ Framework

The *Human-Artifact Model* introduces a number of concepts that overlap with the theoretical foundations of this dissertation in general, and the **A³ framework** in particular. Directly considering algorithmic systems as artefacts, however, would be contradictory with Bødker and Klokmoose’s own characterization of artefacts from a *tools perspective*, as opposed to a *systems perspective*:

“In many ways the perspective of this article has its roots in the tool perspective, [...]. The tool perspective was introduced to address quality of the mediated interaction between the human users and their materials and products. This perspective was introduced as a contrast to the systems perspective, where human users were addressed as components of larger systems.”

[369, p.365]

Although algorithmic systems as conceptualized in Chapter 2 share some attributes with artefacts as characterized by Bødker and Klokmoose—namely their *multiplicity*, *heterogeneity*, and their *reciprocal relationship* to human behaviour—they do not fit well within the scope of *mediating* human activity, goals and actions, and operations. As complex socio-technical assemblages, algorithmic systems often transcend human activities, or even replace them entirely. Contrary to individual, physical computing artefacts, they are often collaboratively and simultaneously used and interacted with

by different stakeholders (including, but not limited to, “users”). Algorithmic systems manifest in the physical world as well, but they do so *through* computing artefacts in Bødker and Klokmoose’s sense, i.e., physical devices like computers, laptops, tablets, smart phones or even specialized and purpose-built equipment. The interfaces themselves, through which humans can interact with these algorithmic systems, similarly fit the description of computing artefacts. Nonetheless, one and the same user interface, even on the same computing device, but combined with different algorithmic systems in the background might yield very different results and impact human behaviour in vastly different ways. To simply subsume an entire algorithmic system—including its user interfaces, the data it operates on, and its technical implementation—under the term “computing artefact” would be quite reductionist. Such an approach would also severely limit the scope of analysis for many of the issues discussed in this dissertation, including the wicked challenges of *transparency* and *accountability*. Consequentially, the terminological and conceptual discussion in Chapter 2 purposefully does not consider algorithmic systems as a whole as artefacts in the sense of *activity theoretical HCI*.

Considering the accountability process as outlined by the *A³ framework*, Bødker and Klokmoose’s model only covers limited aspects, as it presupposes both a human-artefact interaction and a specific activity. The complexity of the process, the various individual actors in the case of *micro-accountability*, and even more so, the *collective actors* involved in *macro-accountability* processes make the model ill-suited as an analytic tool. Accountability processes are not a single, clearly delineated, goal-driven human *activity*, but rather a complex collection of interwoven, sometimes even contradictory, human *activities* that, taken as a whole, determine the success or failure of algorithmic accountability. While the *computing artefacts* certainly play a role in the actor’s and forum’s *emergent interactive agency*, they are but one influential factor of many, as explicated in Section 2.4.4. Human agency—as an analytic lens—thus both includes and transcends Bødker and Klokmoose’s perspectives on human-artefact interaction.

These limitations to the applicability of the *Human Activity Model* to algorithmic systems as a whole, and the accountability process specifically notwithstanding the model can still offer valuable insights into *micro-accountability* processes as outlined in the *A³ framework*. By integrating the model into each of the four phases and thus considering the specific tasks that *actors* and *fora* have to achieve as a potentially artefact-mediated *activity*, answering the guiding questions of the *A³ framework* from the perspective of a (hypothetical or observable) human-artefact interaction may be a promising approach. Such an analysis may follow either or both of the perspectives suggested by Bødker and Klokmoose: starting with the *artefact side* can help evaluate what hindrances and limitations the algorithmic system’s interface and interaction design may impose on the *actor* or *forum*, and starting with the *human side* can support a design process for an algorithmic system (and its interfaces) *in statu nascendi*.

Additionally, the *Human Activity Model* covers an aspect specifically excluded from the *A³ framework* and its underlying theoretical lens of human agency: the motivational aspect. As explicated in Section 6.2.1, the willingness to participate is considered a

precondition for a successful accountability process. Besides intrinsic motivations to do so, it stands to reason that the interactive manifestation of the algorithmic system (i.e., its user interface and the capabilities thereof) can have a significant impact on both the actor's and forum's motivation to engage in the activities necessary for a successful accountability process. Utilizing the *Human-Artifact Model* to design systems that better embody these *motivational aspects* thus can increase the overall likelihood that successful accountability processes occur. As such, Bødker and Klokrose's model can help bridge the gap between human agency on the one side, and human motivation to utilize said agency on the other.

Finally, for the special case of *artificial accountability* introduced in Section 2.4.5, the *Human-Artifact Model* is even more directly applicable, as the accountability process plays out between a (human) *forum* and a computing artefact in Bødker and Klokrose's sense representing the algorithmic system as an actor. Depending on the algorithmic system and the user, the model could be applied to a number of activities related to the accountability process, and both help identify shortcomings of existing interfaces and support design processes for future implementations of artificial accountability processes.

6.2.4.2 A³ and Algorithm Audits / Impact Assessments

As an analytic lens, the *A³ framework* can be helpful in a research context as an evaluation and assessment tool. Following this claim, practitioners in *Algorithmic Impact Assessments (AIAs)* or algorithm audits may want to consider the use of the framework. To explicate where the *A³ framework* fits within the larger movements towards *AIAs* and algorithm auditing, this section takes a closer look at each and considers the framework's applicability and the requisite preconditions for its use.

First and foremost, the fields of algorithm auditing and impact assessment are engaged in similar struggles with terminological overlaps as *CAS* are in regards to the term *algorithm* as a whole. To help structure the following analysis, I refer to a helpful report [374] published by the Ada Lovelace Institute⁶ based on a literature survey, in which they offer a taxonomy for these two terms (see Figure 6.5 for an illustration of this taxonomy).

Based on their report, *algorithm audits* can either refer to *bias audits* as a specific and non-comprehensive study methodology targeting algorithmic systems suspected of biased and discriminatory behaviour, or *regulatory inspection* as a broad approach utilizing a variety of tools and methods to assess a systems compliance to regulation or norms [374, p.3]. By contrast, *AIAs* are distinguished by their time of application in an algorithmic system's lifecycle: *algorithmic risk assessment* evaluate potential societal impacts before the system is deployed, while an *algorithmic impact evaluation* does so after it has been put to use [374, p.4].

⁶See <https://www.adalovelaceinstitute.org/>

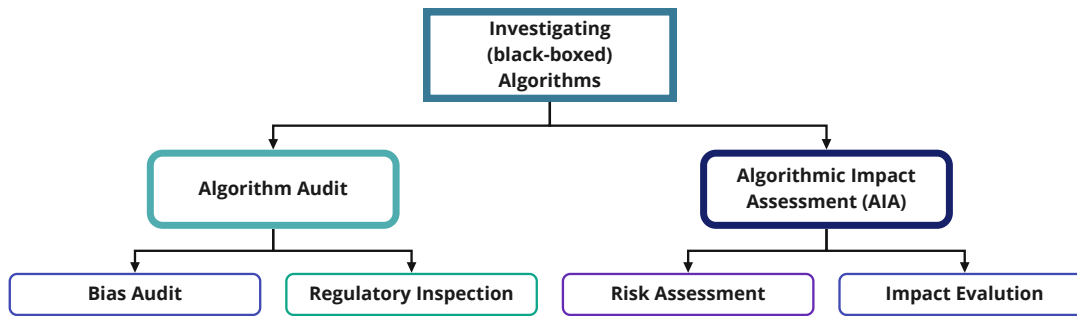


Figure 6.5: Illustration of a taxonomy of methods to investigate algorithms based on a report by the Ada Lovelace Institute [374].

Bias audits

Bias audits were originally introduced by Sandvig et al. [324] simply as *algorithm audits* and are a direct response to the inscrutability of (intentionally) black-boxed algorithmic systems. Primarily aimed at internet platforms (e.g., YouTube, Google Search, or Facebook/Meta) whose internal algorithms are simultaneously wielding exorbitant power and are also shrouded in purposeful secrecy, *algorithm audits* are (adversarial⁷) research methodologies to discover systemic discrimination. With the exception of the design for a *code audit*, they propose a variety of designs (e.g., *scraping audits* or *sock puppet audits*) to infer a system’s behaviour and causally connect it to discriminatory patterns without gaining direct access to the system’s internals. By varying inputs between a test and control group—for instance by performing the same action with two sets of inputs differing only in a single variable, e.g., the gender or last name of a user—and observing the difference in output, these algorithm audits allow a closer inspection of the underlying algorithmic system in a more technical sense than used throughout this dissertation. Given the continued examples for problematic and discriminatory practices of large platform providers with potentially wide-ranging negative and disparate impacts, the popularity of these audit studies comes as no surprise [375, 376, 377, 378, 379, 380, 381, 382]. However, while these methodologies are a vital tool to prove discriminatory practices and their potentially disparate impacts, they are limited in their ability to provide explanations for the observed behaviour, or, as Seaver puts it: “*What they cannot do is explain conclusively how that disparate impact came about.*” [21, p.5]. In his critique, Seaver goes even further and points to the fact that—the laudable intentions of uncovering the secret inner workings of such systems notwithstanding—these audit approaches “*are part of a set of coordinated practices through which algorithms become understood as, and remain, secret.*” [21, p.5]. In other words, audit studies implicitly respond to opacity in

⁷Adversarial is used here in the sense of opposition to suspected wrongdoing, i.e., discriminatory practices that the actor is trying to knowingly keep secret. Sandvig et al. concede that their research designs are “*fundamentally both anti-capitalist and pro-competition*” [324, p.18], as those trying to prove wrongdoing have—at least in the context of a capitalist society—historically been its competitors more often than governmental institutions.

algorithmic systems by treating the inner workings as *unknowable*, and thus contribute to a larger societal narrative of algorithmic technologies as *inscrutable*. This argument somewhat stands in contrast to Sandvig et al.'s [324] original proposal for a shift towards *accountability by auditing*, specifically by addressing the regulatory limitations that are burdening researchers in the United States (e.g., the Computer Fraud and Abuse Act of 1986 (CFAA)). Seaver's criticism notwithstanding, recent regulatory successes in the United States directly related to Sandvig et al.'s work in the form of U.S. Supreme Court rulings on the CFAA and changes to policy in the prosecution of researchers under this law seem to tilt the argument in their favour.

Considering the applicability of the A³ framework for *algorithm audits*, two perspectives are possible: utilizing the framework to assess the algorithmic accountability process enacted *by* these audit studies as a form of *social* or *professional accountability*, and utilizing the framework itself as part of an audit study. For the former, the A³ framework may provide a useful lens to assess the various legal and regulatory, but also socio-technical limitations placed on the collective agency of the research community to perform such studies. Considering, for instance, the automated scraping of an algorithmic system's publicly available Application Programming Interface (API) as a very specific type of a request for information, and the resulting aggregated responses as a type of *account*, would allow the application of the framework's guiding questions to the process. Such an analysis would spotlight, for instance, any intentional attempts by the *actor* (i.e., the platform operator of an algorithmic system) to curtail such automated collection of data as a hindering factor, but also help in identifying the regulatory limitations that may put a researcher in legal jeopardy when employing such methods. In contrast, for the latter perspective on using the A³ framework as part of an audit study itself, its applicability is indeed limited by the implicitly secret nature of the systems in question. As the framework itself is strictly qualitative in nature, its contribution to prove discriminatory practices would be limited to, at most, pointing out disparate *forum* agency for different groups of users when demanding justification for algorithmic behaviour that has been previously identified as discriminatory (e.g., by using other *algorithm audit* study methodologies).

Regulatory Inspection

Regulatory inspection takes its roots from the common language understanding of the term *audit* (e.g., financial or legal auditing), and describes a process of evaluation by regulating authorities to assess the compliance of a given algorithmic system and its operationalization to various legal standards and regulatory requirements. As compliance with these standards is not optional, regulation prescribing a mandatory audit in the sense of a regulatory inspection can be used to compel an otherwise unwilling organisation to allow a specific forum (i.e., the *auditors*) access to information detailing the inner workings of an algorithmic system. Consequently, these inspections would represent a type of *legal* or *administrative accountability*, depending on the nature of the forum and actor: a governmental auditing body inspecting a private enterprises' compliance would be considered *legal accountability*, while an internal inspection of an algorithmic system

employed by the government itself could be considered *administrative accountability*. The methodologies for such an audit can be an eclectic collection of approaches to cover the system in its entirety, or specific to the requirements mandated by the relevant regulation domain. In practice, few examples for such regulatory inspections of algorithmic systems currently exist, not least due to the difficulty in providing a coherent and widely applicable framework for such audits. This challenge is illustrated by the case study example given by the Ada Lovelace Institute’s report [374] detailing the UK Information Commissioner’s Office efforts to develop an Auditing Framework for AI technologies: Based on their draft, they recommend a variety of methodologies, including “*identifying and assessing trade offs, bias auditing, explanation and training, and documentation of decision making including legal, organisational, technical and security considerations*” [374, p.14].

The draft [383] also considers *accountability and governance* as one of the areas in need of assessment as part of such regulatory inspection. In this sense, the **A³ framework** could be used as part of a larger set of methodological approaches to evaluate both micro- and macro-accountability processes embedded in the algorithmic system in question. To correspond to the regulations governing such an inspection, however, adapting the guiding questions may be necessary to highlight where, specifically, the algorithmic system may infringe on the *actor’s* and *forum’s* agency to adhere to requisite regulatory requirements.

Algorithmic Impact Assessments

A³IAs—both in the sense of *a priori risk assessments* and *in medias res* or *ex-post impact evaluations*—are holistic analyses of the potential or observed societal impact of an algorithmic system. Originating from similar assessments in environmental regulation or, more recently, data and privacy preserving regulation such as the EU’s **GDPR** [170], these assessments have been, to date, most prevalent in the context of the public sector [374, p.15]. In terms of access to internal information about the algorithmic system in question, the broad scope of such an assessment in terms of large-scale societal impacts almost always entails full knowledge and access to a system’s composition and behaviour; in other words, high standards of *system-level transparency* and *ex-post explainability* at least for the party executing the assessment is required. For an *a priori*, forward-looking risk assessment that accompanies the system’s design process, this access is easier to attain than when retrospectively analysing the impact of a system. This imbalance may be offset by the inherent limitation of predictive studies and forecasting methodologies, as a retro-active *impact evaluation* can rely on empiric observations of real events as the information sources for their assessment.

For both of these types of **A³IAs**, the **A³ framework** offers a useful perspective and contribution to the specific assessment of accountability capabilities of the system in question. In fact, the application of the framework to the **AMAS** system as presented in the previous sections arguably constitutes such a contribution to the larger impact assessments⁸ performed as part of our research projects [3, 4]. Depending on the type of

⁸While it is important to note that, given the scarce resources and limited access we had during our

desired outcome, the [A³ framework](#) may not produce the needed results if, for instance, a quantitative assessment is required. This limitation notwithstanding, using the framework as an exploratory device to highlight aspects of an algorithmic system pertaining to algorithmic accountability that would warrant further investigation is certainly plausible and recommended.

6.3 Chapter Summary

In this chapter, I presented the [Algorithmic Accountability Agency Framework \(A³ framework\)](#) as a contribution towards the ongoing efforts to assess and improve algorithmic systems in terms of their algorithmic accountability.

In accordance with the methodological considerations discussed in Section [3.4](#) on the *comparative case study* of both the EnerCoach and [AMAS](#) algorithmic systems, I have presented a rationale for the comparability of the two case studies, discussing both differences and similarities in terms of their requisite dimensions of comparability. In addition, I provided a more thorough discussion of both case studies' applicability to the field of [CAS](#), given the overall emphasis on complex, [AI/ML](#)-based algorithmic system and the comparably simple nature of the case studies in terms of their underlying technical sophistication.

Building on the theoretical foundation of Bandura's *Social Cognitive Theory* for *emergent interactive agency*, the framework operationalizes the synthesized learnings from both case studies into an *analytic lens* for the relational processes of *macro-* and *micro-accountability* occurring between a *forum* and an *actor*. In applying the framework to prototypical micro-accountability processes for each of the case studies, the framework's capabilities to help identify and assess the factors enabling or hindering both the *forum's* and the *actor's* agency were demonstrated. Based on these results, I also presented a comparative, qualitative assessment of methodological attempts to improve both system's potential for algorithmic accountability, and described the benefits of participative and inclusive design processes involving multiple stakeholders for the resulting measures and overall algorithmic accountability of the systems.

Finally, I compared the [A³ framework](#) to prior proposals for accountability frameworks, contextualized it adjacent to activity theoretical approaches in the form of Bødker and Klokmoose's *Human-Artifact Model*, and evaluated the potential applicability of the framework as part of various types of *algorithm audits* and [AIAs](#).

study, the outcome should not be seen as a comprehensive technology assessment study (in the sense of the term used in [STS](#) and related disciplines), some of the characteristics of such studies featured in our approach as well.

6.4 Chapter Conclusions

In addition to the detailed description and analysis of the [A³ framework](#) and its application to the case studies as summarized above, a few more general conclusions are worthy of consideration at this point.

Addressing the Research Questions through the [A³ framework](#)

First and foremost, in [SRQ3](#) and its subsequent parts, I posed the overall question of what guidelines to improve algorithmic accountability for future algorithmic systems could be developed. Admittedly, the [A³ framework](#) as an assessment tool does not constitute such a set of guidelines *per se*. Considering the insights gained through the theoretical considerations in Chapter 2, the case studies, and the application of the [A³ framework](#) itself to those case studies, however, it became clear that creating a set of universally applicable guidelines that, at the same time, remain concrete enough to offer any practical use is simply impossible.

The [A³ framework](#) and its foundational analytic lens of *human agency* (see [SRQ3.2](#)) represents the next best thing: a framework that, when applied to a prospective algorithmic system or an existing one, will provide the insights necessary for a guided, accountability-conscious implementation of future systems in that application context. Answering [SRQ3.1](#)—*How can accountability requirements be formulated and adapted based on the application context?*—then becomes trivial: By offering a process of either exploratory and scenario-based—or retrospective and empirically founded—evaluation that yields the context-specific requirements for algorithmic *micro-* and *macro-accountability* processes.

In summary, the [A³ framework](#) may not directly represent a specific set of guidelines, but something rather more useful: a guiding framework ([SRQ3.3](#)) capable of producing the insights necessary to identify the deficits and suggest material solutions for algorithmic accountability in a wide range of contexts.

Applying the [A³ framework](#) to Real-World Processes

Second, a note regarding the epistemological validity of an analysis based on the [A³ framework](#) and its guiding questions. In presenting the framework, the primary goal of the application is that of an exemplary showcase. To illustrate the multiplicity of possible dimensions for both enabling and hindering factors of human agency in accountability processes, it was necessary to deviate from real interactions captured, for instance, in interviews or through observation. However, the framework itself presents no such limitations: it is entirely plausible to utilize it for specific, real-world instances of *micro-* or *macro-accountability* processes that have been recorded through the various established methods of behavioural studies, user studies and so forth. If we had been given access for the case study of [AMAS](#) to observe caseworkers as they interact with jobseekers and the system, the framework could have easily been applied to help make sense of the

interaction through the lens of agency. Reformulating the secondary guiding questions to directly address the *forum* and *actor* afterwards would have been the logical next step to collect empirically validated insights into the real limits of their potential actions in the process.

In light of this approach to use direct forms of empiric data collection to employ the [A³ framework](#) in a more discursive manner, considering the other aspects of Bandura's [SCT](#)-based model of *triadic reciprocal causation* for human agency becomes a possibility. As it stands, the application of the [A³ framework](#) through the use of scenarios could not account well for either *internal personal factors* (e.g., *cognitive*, *affective* and *biological* events) or the *human behaviour* of individual *actors* or *fora*. Including these considerations, however, could be a worthwhile endeavour to better characterize, at least anecdotally, which internal reasons actors or fora may have had for not engaging in a productive accountability process, or which perceived behaviours of their counterpoint impacted their own agency to participate in the process. Particularly for the relevant question of *motivation* as an example of an *internal personal factor*, including Bødker and Klokrose's *Human-Artifact Model* [\[369\]](#) could add some particularly pertinent perspectives. Given the complexity of human psychology and behaviour, considering these aspects on a speculative basis based on scenarios and personas seemed rather imprudent, hence the omission of such considerations in the application of the [A³ framework](#) to the two case studies.

Simplicity and Complexity, Versatility and Specificity

Third, and on the topic of simplistic versus complex frameworks: by and large, a framework's scope of applicability rises with its generality, at the cost of specificity to the case in question. Naturally, a different framework founded on the same concept of spotlighting human agency could be devised with a much larger set of more specific guiding questions, tailored to domain-specific groups of *fora* and *actors*. The benefits of such a specialization are evident: the more specific the guiding questions, the smaller the risk of overlooking certain (domain specific) aspects. The current approach as presented in this chapter, however, prioritizes *versatility* over *specificity* based on two considerations. Primarily, this more simplistic framework puts the focus on the *universal nature* of human agency in human-computer interaction, which governs what is possible for humans to do independently of which application domain or, indeed, task they set out to accomplish. Secondly, the insistence on qualitative inquiry over quantitative measurements encourages a deeper, more critical and nuanced look at the four steps of the accountability process than a check-list style or Likert-scale based evaluation framework would. By doing so, and by recommending the use of the framework in collaborative, exploratory research settings, it is my hope that the resulting analysis will be more characteristically representing the algorithmic system and its assemblage—as suggested by Seaver [\[21\]](#)—*as culture*.

At the same time, we must acknowledge the importance of accessible and intuitive tools for assessing algorithmic accountability. As a consequence of the ubiquity of algorithmic systems, their often seemingly “banal” nature combined with the significant power

they have to shape and impact human lives, complex and highly specific assessment methodologies may only ever be used by the domain experts, academic researchers or highly trained auditors. Such an audience, I argue, will simply not suffice to address the many different algorithmic systems permeating our world as they are continuously co-produced and transformed by an enormous—and growing!—number of stakeholders. If we are to improve algorithmic accountability *at scale*, we must offer easily and widely applicable tools and conceptualizations to a broader audience, from students in Computer Science, [STS](#) and the larger Social Sciences and Humanities, to software engineers and other professionals in many technical and non-technical fields, to politicians, administrators, bureaucrats and policymakers. The core tenet of *human-centricity*, the simple and open nature of the guiding questions, and the structured approach to micro- and macro-accountability processes thus makes the [A³ framework](#) much better suited for a wider audience than more specialized, quantitative or complex other frameworks for algorithmic accountability.

The A³ Framework and Moral Responsibility

Fourth, the [A³ framework](#) as presented and applied to the case studies purposefully omits questions of *moral responsibility*, which may seem, at first glance, counter-intuitive in light of the discussion on *moral responsibility* and *artificial accountability* in Section [2.4.5](#). While accountability processes can lead to an assessment of culpability in a moral sense, and some types of *macro-accountability* processes (such as, for instance, legal accountability) are designed to determine blame and issue consequences as sanctions or punishment, the broader scope of accountability (both in its micro and macroscopic instances) transcends these narrower goals. Particularly in the context of micro-accountability, however, assuming that rendering an account and imposing consequences would include a judgement by a *forum of one* on such matters of moral responsibility runs counter some of the most fundamental paradigms of a modern understanding of justice. Furthermore, designing an assessment framework that would be capable of evaluating how well an algorithmic assemblage supports such a process of assigning responsibility would, by definition, require to be contextually adapted for specific systems of legal culpability and moral responsibility, as I discussed in more detail in Section [6.2.4](#) as well. Following the arguments on *versatility* over *specificity* made above, such specialization would make the [A³ framework](#) also less applicable and useful in the broader contexts for which it was designed. As a final argument, including dimensions of moral responsibility also would categorically preclude the (potential) future use of the [A³ framework](#) for an assessment of *artificial accountability*: even though we may consider the technical components of an algorithmic assemblage as capable of moral agency, we do reject the notion of machinic moral responsibility on the grounds they lack *reflexivity*, *norm-sensitive subjectivity*, *sovereignty* and *intentionality* [\[222, 223\]](#).

Yet Another Framework?

Finally, a concession: As a scholar working in a variety of inter-disciplinary fields that are, arguably, already more than saturated with any number of guidelines, frameworks, or other abstractions of complex, often *wicked* problems, I am painfully aware of the inherent irony in presenting *yet another* guiding framework for algorithmic accountability (or, for that matter, *agency, transparency, literacy*, or any number of other previously identified challenges in [CAS](#), [HCI](#) or [CSCW](#)). At a certain point, as scientists and researchers, we must concede that the act of *finding* and *evaluating* the *right* framework out of the many available options may present a larger challenge than even *applying* that framework to a future case study. This reality creates a strange dynamic not unlike the attention economy of *surveillance capitalism*, as Zuboff [74] so provocatively termed it, where the success of such a framework is determined less by its usefulness and more by how well it is marketed or distributed. Consequentially, many of these approaches may never be applied by anyone but the authors, and—considering the institutional pressures of evolving research foci that force us to move on to the next critical topic or crucial issue in the field immediately—many of them may already be doomed to oblivion on the day of publication.

At this point, I can make no claims that the [A³ framework](#) will not, possibly, face the same future of irrelevance. However bleak this outlook may seem, I do believe that the fundamental approach of considering abstract, *wicked* problems from a human-centric perspective and as concrete processes playing out in and limited by the realm of human agency can stand on its own merits, regardless of the success of practical applications such as the [A³ framework](#). Perhaps even just inspiring other scholars to adopt a similar approach, to adapt the framework to their needs and to other problems, or simply to consider the arduous methodologies of qualitative inquiry as worthwhile for future endeavours constitutes its own form of success outside of the prevalently captured metrics of ‘citations’ and ‘impact factors’.

CHAPTER 7

Conclusio

Algorithmic Accountability is, and remains, a *wicked problem* for many of the algorithmic systems permeating almost all aspects of modern society. In this dissertation, I ventured to tackle this problem from both theoretical and conceptual, as well as very practical perspectives.

Throughout the previous chapters, I explored the different ways we can conceptualize what *algorithms* and *algorithmic systems* are, how they interact with humans and the world around them, and the frictions, tensions, challenges and issues that arise from their existence. As with all *wicked* problems, a definitive formulation of the problems of algorithmic accountability and its related challenges of *transparency*, *bias* and *discrimination*, will remain elusive. The theoretical foundations upon which this dissertation builds its arguments are, by necessity and not just by choice, eclectic and diverse, including *socio-technical systems* and post-humanist models such as *assemblage thinking* or ANT to describe algorithmic systems, to further cross-disciplinary concepts such as Bovens' work on *public accountability*, Bandura's concept of human *emergent interactive agency* or discussions of *moral responsibility* and *non-human agency*. Similarly, the practical study of algorithmic systems in the form of the two case studies required a similar methodological diversity, including various qualitative methods of inquiry, *auto-ethnographic* approaches, *participatory design* and *critical document analysis*.

The necessity for such theoretical and methodological eclecticism is a consequence of the subject matter at hand: Algorithmic systems are diverse, fluid, *ontogenetic* and *heterogeneous* socio-technical assemblages that, should we attempt to analyze them in their entirety and not just as technical artefacts alone, decidedly defy mono-disciplinary approaches. Thus, as I argue and demonstrate in this dissertation, the need to look beyond the narrow confines of single academic disciplines in order to grasp the breadth and variety of all that algorithmic systems are.

Such inter- and cross-disciplinary analysis led to a better understanding of the problem of algorithmic accountability. The need to adapt Bovens' well-established and foundational work on *public accountability* to different challenges and requirements posed by algorithmic systems, primarily in terms of their scope and the immediacy of their impacts on the world, led to the introduction of *micro-accountability* as a counterpoint to more traditional, but slow-paced, macro-accountability processes. As a consequence of the closer look at individual human contributions to these smaller accountability processes, the analytic lens of *human agency* as a more specific perspective of human-centricity emerged. In other words: by reducing the collective, macro-accountability (“*We need to hold this algorithm to account!*”) to the individual, micro-accountability process of “*Tell me why this system made this decision about me!*”, the requirements for algorithmic accountability include the questions of how our human agency is shaped, limited or improved by the algorithmic system in question.

The practical application of these theoretical foundations to two case studies—the EnerCoach energy accounting system and the AMAS profiling system for the unemployed—was the logical next step to further explore how these different, sometimes contradictory, concepts and theories can help us understand the relationship between *algorithmic accountability* on the one side, and *algorithmic transparency*, *ex-post explanations*, *literacy* and *domain knowledge*, as well as *bias* and *discrimination* on the other. Overall, the insights gleaned from these case studies—from the value of participatory design methodologies or the dangers of a lack of inclusivity, to the limitations a lack of transparency and explainability poses to the stakeholders and the various socio-technical strategies they developed to circumvent the issue—underscored the deeply interconnected nature of these socio-technical assemblages and their human, non-human, and immaterial components.

Working towards the overall goal of more generalizable insights as well as concrete, practical tools to work towards higher standards of algorithmic accountability, the diverse nature of the case studies offered the opportunity of sharpening our perspective towards the common nature of accountability processes in the form of *human agency*. Despite the different, underlying technologies, context of application, and vastly different needs and requirements of their respective stakeholders, one unifying question remained after the comparative case study: *How can we maximise the potential of human beings, individually or collectively, to act in their roles as actors or fora and conclude a meaningful accountability process?*

This dissertation provides an answer to this question in the form of the Algorithmic Accountability Agency Framework (A³ framework). With the diversity and heterogeneity of not just the case studies, but the overall landscape of real-world algorithmic systems, in mind, the A³ framework offers not a set of concrete guidelines, but a structured process for the evaluation and assessment of algorithmic accountability processes. In other words, in a bid for versatility and accessibility, the A³ framework does not offer a predetermined list of *do's and don'ts*, but empowers practitioners by offering instructions of how to arrive at their own list—tailored to the context, stakeholders, technologies of the algorithmic system in question. This specificity and context-sensitivity of the A³ framework, as

showcased through applying it to the two different case studies, also allows it to support the ideation of possible improvements and concrete measures specific to the system and its *micro-* or *macro-accountability* processes.

Versatile and accessible tools like the [A³ framework](#) are an important contribution towards more accountable algorithmic systems, in all their complexity, diversity and heterogeneity. Its procedural nature ensures its wide applicability to a variety of systems and contexts. The fact that it is simple and accessible and encourages critical reflection over checklists and recommendations, makes it a powerful tool beyond the realm of academia and scientific research as well. To address the challenge of algorithmic accountability, after all, raising awareness for the importance of algorithmic accountability as a *virtue* must go hand in hand with providing practitioners, system developers and decision makers with the tools to implement algorithmic accountability as a *mechanism* as well.

Finally, in light of the categorical impossibility of providing definitive and universal solutions to the *wicked* problems of algorithmic accountability and transparency, this dissertation offers the best possible alternative: a solid and holistic set of theoretical and conceptual approaches to analyze concrete algorithmic systems, methodological suggestions that may serve as best practice examples for analytical or interventionist case studies, and the [A³ framework](#) as a concrete tool to assess and evaluate algorithmic accountability processes and generate material solutions.

To explicate this claim and summarize the detailed findings of this dissertation, this chapter offers detailed conclusions on the contributions made by this dissertation to the field of [CAS](#) and beyond. In the subsequent sections, I also discuss the limitations of this work, and present an outlook towards promising future avenues of inquiry into *algorithmic accountability* and its related issues.

7.1 Summary Contributions

The contributions of this dissertation to the field of [CAS](#) are *theoretical* and *conceptual*, as well as *practical* in nature. In the following paragraphs, I outline these contributions and discuss some of the larger implications that follow from the more detailed results presented in the previous chapters of this dissertation.

7.1.1 Theoretical Contributions

The theoretical and conceptual discussion of the terms *algorithm* and *algorithmic system* presented in Chapter [2](#) represents a summary of approaches from various academic disciplines and fields. While the observations made in that chapter are directly founded on prior work from these disciplines, not all of the theoretical underpinnings had been applied to algorithms and algorithmic systems before. What is certainly a novel contribution is the structured, side-by-side presentation and comparison of these different conceptualizations, as well as the integration of cross-disciplinary theories. Taken together with the following,

overarching theoretical conclusions, these contributions directly address [SRQ1](#) and its supplemental questions.

Embracing Complementary and Conflicting Perspectives

When juxtaposing the different perspectives on algorithms discussed in Chapter [2](#), it is paramount to note their *complementary* nature, as opposed to a *competing* one. Neither is one perspective inherently more accurate than others, nor should they be considered as ranked by their universal usefulness. Instead, the different perspectives can all serve a different purpose when considering algorithmic systems, and should be chosen to *complement* each other as different views on the subject matter. After all, as I discuss in Section [2.1.3](#), algorithms are *multiples*, that is to say, *many things at once*. It follows that, in order to characterize and understand them well, multiple theoretical models will be required as well.

With this assertion, however, comes the challenge of reconciling the conflicting nature of various models. As discussed in more detail in Section [4.7](#), not all theoretical models and approaches complement each other harmoniously. For instance, considering algorithmic systems as complex socio-technical assemblages offers a lot of utility, for instance in the form of *non-human agency* as a conceptual tool to account for characteristics and processes within algorithmic assemblages that can not be directly traced to human action. At the same time, the *post-humanist* roots of these approaches—in their deeply sceptical view of human sovereign subjectivity—make it difficult to reconcile assemblage thinking with questions of moral responsibility and human agency. It is precisely this conceptual struggle of reconciliation, however, that makes these contradictions so *productive* and leads to the emergence of new insights: Forced to consider and weigh these various perspectives, their merits and shortcomings, conceptual focus and gaps, we end up with a more nuanced and rich characterization of the subject matter at hand.

These various perspectives each emphasize different aspects of algorithmic systems. *Technical definitions* highlight the most tangible and precise behaviours of algorithmic systems, albeit detached from their context of use and larger societal scope they are embedded in. A perspective of algorithms as *socio-technical systems* spotlights the interrelatedness between algorithms and their context of use, in relation to an enterprise, and the way they get used, abused, and mis- and re-appropriated by the people that use them. *Assemblage thinking* and [ANT](#) teaches us about the hybrid forms of *human and non-human agency* shaping algorithmic systems, and their fundamentally *relational* and *heterogeneous* nature. They both also highlight the fact that algorithms do not *represent* reality, but *interpret*, *mould* and *abstract* it, and in doing so impact both themselves and their figurations as much as the outside world they operate on and in. These commonalities aside, both *ANT* and *assemblage thinking* offer slightly different perspectives on the nature of change in algorithmic systems: [ANT](#) helps us remember the *fluid* and *ontogenetic* nature of algorithmic systems as ever-changing, ever evolving entities, and *assemblage thinking* assists in pinpointing specific points of rupture or change as *discrete events* in the lifecycle of the system. In terms of the way technologies like algorithms

come into being, the concept of *co-production* forces us to abandon *techno-deterministic* or *socio-deterministic* points of view in favour of a *hybrid* and *reciprocal* relationship of influence between society and technology. Finally, the *functional* perspective on algorithmic systems removes much of the clutter and distractions inherent in these previous conceptualizations, by putting the focus solely on what specific algorithmic systems *do in the world* and how they are instrumentalized by various stakeholders.

The Shift in Power and Agency

One characteristic effect of algorithmic systems emerging from nearly all of these different conceptualizations, I argued in Section 2.7, is the shift in power and agency caused by an overall, societal trend towards automation. As algorithmic systems become the primary infrastructure supporting and enacting automation, a gradual shift from human agency towards algorithmic agency accompanies this process. Against this backdrop of shifting power hegemonies, the manifold lofty promises of algorithmic technologies must be scrutinized with careful scepticism and critical reflection. Whether or not increases in automation truly will free humankind from the shackles of tedious, meaningless, repetitive and possible harmful tasks may be a matter of discussion, but the fact that handing over the metaphorical reigns to algorithmic systems requires new and more reliable processes of algorithmic accountability is hardly disputable. *Bias and discrimination* are already a primary focus of ongoing research in CAS, and represent just one of the reasons why we need to be able to hold algorithmic systems to account. Friedman and Nissenbaum's [102] warnings, written more than a quarter of a century ago, have lost none of their relevance. Case study after case study provide ample evidence as to how algorithmic systems can embody various *pre-existing*, *emergent* and *technical* biases, and how they can subsequently reproduce and enact them in a discriminatory manner. The AMAS case study, for instance, is a particularly illustrative example as it shows the presence of each of these types of biases, although this should not be seen as representative for the larger landscape of algorithmic systems, since—thankfully—not all case studies on algorithmic bias exhibit such a worrying diversity of discriminatory potential. What unites many instances of algorithmic bias, however, is the fact that *automation bias*, i.e., the human tendency to ascribe objectivity to the decisions made by automated systems, can further exacerbate the discriminatory impacts of bias in computer systems. The importance and impact of these non-technical types of bias notwithstanding, many of the scientific efforts to combat bias still happen on a purely technical level (e.g., quantitative assessments of biased datasets, or automated bias reduction for machine learning algorithms) and thus, by definition of the various conceptualizations of algorithmic systems, can—at best—only offer part of the answer to these multi-faceted, complex, socio-technical problems.

This potential for bias and discrimination inherent in the use of algorithmic systems also serves as one of the strongest arguments for the need for better *algorithmic transparency*: if we aim to detect and improve biased and discriminatory systems, we must be able to first recognize that they are being used, and second, be able to investigate and analyse their inner workings. The shift in agency diagnosed above, however, complicates this

challenge further: limited human agency to act as part of socio-technical assemblages often implies a limited agency to identify *where* algorithmic technologies are deployed, to scrutinize *how* the non-human components of these assemblages act, and *why* they do so. *Algorithmic transparency* then, by itself, already is a wicked problem, much like *algorithmic accountability*. Providing a unified and generally applicable definition of when a system should be considered *transparent* thus becomes a rather impossible goal.

The Many Facets of Transparency

To tackle this complex conceptual problem, we can consider the two different aspects of transparency—*system-level transparency* and *ex-post explainability*—as different endpoints of the same spectrum. Such a perspective emphasizes the need for different, complementary measures to approach the issue: Transparency challenges such as *intentional opacity* or *remedial incomprehensibility*, *unintentional opacity* (e.g., due to a lack of algorithmic literacy or domain knowledge of the intended audience), and the *inherent complexity* of certain algorithmic technologies further highlight the need for very different, contextually appropriate solutions that can only be achieved through a comprehensive analysis of stakeholders and their needs. For some systems, however, we may have to consider whether or not such a solution may even be possible. The limitations of certain *inscrutable* techniques such as [Artificial Intelligence/Machine Learning \(AI/ML\)](#) poses some rather difficult questions. First and foremost, I would argue, we must take a hard and unflinching look at whether or not employing these techniques can even be considered ethically acceptable, in particular for those contexts in which human health, well-being, autonomy or integrity are at stake. In absence of simple, definitive answers to this, admittedly, controversial and complex ethical question, tackling the problem solely through technical means as exemplified by much of the work in [XAI](#) simply is the wrong approach.

Although the field of [XAI](#) is receiving a disproportionate amount of attention at the moment, as I argue in Section [6.1.3](#), there still are many significant gaps in our understanding of fundamental aspects of both transparency and explainability, as well as the question of how they relate to the accountability of algorithmic systems. Before we know what to do with them, investing into ever more difficult to attain and abstract concepts of explanations offers little practical use for the problems at hand. On the contrary, we must renew our interest in studying the seemingly banal and “*boring*” [\[356, 242\]](#), but ubiquitous and no less powerful algorithmic systems that are embedded in the fabric of our digital infrastructure today. Studying systems like the two case studies in this dissertation, EnerCoach and [AMAS](#), then promises fundamental insights without the challenges of intrinsic complexity and inscrutability in the way.

Finally, relating *system-level transparency* and *ex-post explainability* to *accountability*, we can see how, in principle, a lack of transparency may prohibit *successful* accountability processes—which does not necessarily preclude *unsuccessful* accountability processes from happening regardless. At the same time, we know that neither full *transparency*

nor excellent *ex-post explainability* automatically guarantee a successful *accountability relationship* due to the range of other factors influencing the viability of such a process.

The Nature of Algorithmic Accountability

When considering what these other factors may be, answering what exactly algorithmic accountability *is* becomes the necessary prerequisite for addressing these issues. *Algorithmic accountability* as an adaptation of *public accountability*, as I have argued alongside many other scholars in the field, indeed provides a useful foundation for further inquiry. Bovens' [22] body of work underscores the *procedural* and *relational* nature of accountability, and his taxonomy allows differentiating between the various needs and requirements which different types of *actors*, *fora*, *obligations* and *accounts* introduce to this process.

Appropriating these concepts for *public accountability* towards an accountability of algorithmic systems also confirms the need for different types of *ex-ante*, *in medias res* and *ex-post* algorithmic accountability processes, as introduced by Wieringa [23]. Thus, algorithmic accountability becomes applicable throughout the various lifecycle phases of an algorithmic system, from inception, design, implementation towards application and, finally, decommission. This point is particularly important, as many of the existing accountability measures for *public accountability* are implicitly linked with (moral and legal) responsibility, and consequently only come into play after something has gone wrong—in which case the (potentially avoidable) damage has already been done. While this approach may be acceptable for the slow-moving pace of *public accountability*, the scope and large-scale impacts of algorithmic systems make the omission of pre-emptive accountability unacceptable in light of their potential for harm. In other words, the speed at which automated technologies operate creates a tension between their capacity for damage and our ability to react through traditional means of large-scale accountability processes. To put it even more bluntly: *If an algorithm can easily discriminate against millions of people in a matter of seconds, and we can only hold its socio-technical assemblage to account in a matter of months or even years, we are, indeed, in very deep trouble.*

A Shift in Scope Towards Micro-Accountability

The introduction of *micro-accountability* as the counterpoint to the more traditional *macro-accountability* processes described by Bovens [22] responds directly to this challenge. *Micro-accountability* provides us with the necessary perspective to interrogate a system's capabilities on the smaller scope of real-world, concrete and individual interactions between human and non-human actors. Addressing the shortcomings of slow *macro-accountability* processes directly, the power of *micro-accountability* lies in its *immediacy*: as humans are affected by or working with algorithmic systems, they react directly and immediately by identifying an actor to address and demanding an account. In the most

minimal case of a single individual initiating this process, such an act constitutes them as a *forum of one*.

This shift in scope opens up a whole range of possible avenues of inquiry. For one, the question of what kind of power a *forum of one* may have to impose meaningful consequences forces us to reconsider the nature of consequences themselves. No longer primarily associated with negative judgements in the form of sanctions or punishment for the purpose of *deterrence*, the agency to impose *consequences* may serve the goal of *self-protection* from harm for individual fora. Thus, such consequences may include the ability to opt out from being subjected to a system's decision making process, to demand an alternative process with a human counterpart, or simply the ability to escalate the accountability process to a suitable *macro-accountability* forum and halt further processing until the conclusion of the larger process. In the hypothetical *micro-accountability* process for the AMAS case study occurring between jobseeker and caseworker, the agency to impose these consequences could be an important safeguard: being able to opt out of the AMAS process for future meetings, for instance, would most likely suffice for many jobseekers as a satisfying solution to an accountability process.

The current lack of attention to these *micro-accountability* processes does not, in any way, mean that they do not exist or that they are not being initiated. It does, however, mean that they often lead to dissatisfactory conclusions due to a lack of agency of affected stakeholders, actors and forum alike. To ensure meaningful micro-accountability, we have to purposefully design algorithmic systems to offer the necessary features, capabilities and processes to empower affected stakeholders to participate. In this way, the shift to a meaningful *micro-accountability* also requires a paradigm shift from *defensive* to *deliberative accountability* as a *virtue* [205]. In order to incentivize this shift, the benefits of successful micro-accountability processes should be emphasized clearly: more than just another form of compliance, these processes can help *actors* ensure higher customer satisfaction due to less friction and unresolved disputes, provide valuable feedback on the needs of users and the challenges stakeholders of algorithmic systems face, and finally, they can serve as a preventative measure to avoid larger, slower and often more costly high-level *macro-accountability* processes. In other words: if we can plausibly explain to decision makers in the private sector that they will save money and reduce the risk of costly litigation by pre-emptively designing meaningful micro-accountability processes, their willingness to invest into such measures will rise.

Finally, from an academic standpoint and for analytic purposes, *micro-accountability* may even be considered as a *synecdoche* for accountability processes as a whole: if a system already shows a significant lack of *micro-accountability*, we should take this observation as the proverbial *canary in coal mine* pointing us to larger, structural issues with the *macro-accountability* of the system.

Synthesizing Theory and Method into an Analytic Lens

For algorithmic accountability processes in general, but even more so when considering *micro-accountability* processes from a human-centric perspective, we inevitably must ask ourselves what theoretical models are best suited to characterize the human ability to engage with such a process. After all, designing such a process with and for humans requires an understanding of the factors that (positively or negatively) impact their agency. The theoretical conceptualizations (such as [ANT](#) or *assemblage thinking*) that we turned to before do indeed offer us fundamental theories on the nature of agency, including the helpful extension of the concept to non-human actors. However, their insistence on the distributed nature of agency and their scepticism towards human subjectivity makes them less-than-useful when trying to describe what, precisely, characterizes human agency in an algorithmic accountability process. Reconciling this contradiction—humans exercise, arguably, a different quality of agency than non-human actors in the same assemblage—thus requires us to reject the more radical positions of [ANT](#) on *exclusively mediated* agency in this case. Assemblage thinking, on the other hand, integrates better with other theories of agency through accepting *intrinsic qualities* of its constituent parts—hence the inclusion of Bandura’s [SCT](#) and the model of *triadic reciprocal causation*¹.

Thus, the theoretical contributions of the proposed [A³ framework](#) itself rest in its specific perspective on accountability processes as seen through the analytic lens of human *emergent interactive agency*. As a cross-disciplinary concept, this perspective cuts through the complexity and heterogeneity of both micro- and macro-accountability, and lets us approach these wicked challenges from a fresh and unified perspective. By relating this approach previous theoretical work in [HCI](#), including Bødker and Klokmoose’s *Human-Artifact Model*, the framework’s potential to integrate prior conceptual models and utilize their perspective in a meaningful way was demonstrated as well. Consequently, the [A³ framework](#) also serves as a showcase for the value of integrating inter-disciplinary insights into the study of socio-technical challenges such as *algorithmic accountability*.

From an ever wider perspective, this dissertation also introduces Bandura’s [SCT](#) and *human emergent agency* as a concept to the larger debates on human-centricity in technology development—including, for instance, the “*Digital Humanism*” movement. Doing so, I posit, is an important reminder for us that, if these debates are to be more than just theoretical or philosophical lip-service to abstract ideals and virtues, we must look closely at the specific, individual and direct ways in which we can make sure humans retain their ability to navigate these complex socio-technical assemblages. The contradictions between post-humanist and humanist models and approaches, as admittedly difficult to reconcile as they may be, should serve not as the end point, but rather as the productive foundation for discussions and conclusions that lead to concrete solutions. Human agency in algorithmic systems or assemblages can offer the unifying lynchpin around which an empowering human-centric agenda for algorithmic accountability may be formed.

¹For a more detailed justification on this theoretical choice, kindly refer to Section [2.4.4](#).

Moral Responsibility and Moral Agency

Micro-accountability processes will never, at any rate, entirely replace the need for well-defined *macro-accountability*. They can, however, help address the specific gaps in accountability introduced by the appropriation of public accountability to algorithmic systems. Beyond the issue of *immediacy* as discussed above, the concept of *micro-accountability* also offers a plausible approach to address the looming spectre of accountability in fully automated systems.

The excursus on the relation between moral responsibility and moral agency for human and non-human actors towards the end of Chapter 2 outlines the philosophical challenges we face when we consider the technical components of algorithmic systems as part of a *socio-technical assemblage* that is involved in accountability processes. As algorithmic systems find more and more widespread use, human-to-human accountability processes will inevitably become infeasible due to the scale at which these systems operate. This holds particularly true for macro-accountability processes, as no forum will be able to deliberate the rising number of accounts of conduct of algorithmic systems and their processes. But even micro-accountability processes between individual human *fora* and *actors* will not be able to provide meaningful accountability in many cases. Thus, moving towards a perspective in which non-human actors—such as the technical components of algorithmic systems—are seen as having *moral agency without moral responsibility*, as suggested by Floridi and Sanders [158], may simply become a pragmatic necessity. These developments may lead to what I introduced as *artificial accountability*: limited and, ideally, context-specifically designed micro-accountability processes between *human fora* and the technical components of algorithmic systems as *non-human actors*. Learning on established theoretical and practical work in HCI, such as the *Human-Artifact Model*, will be key in shaping these novel types of accountability processes to ensure they embody *human-centric* values and can lead to satisfactory outcomes.

Wicked Problems and Satisficing Solutions

Analogous to *algorithmic transparency*, the wicked nature of *algorithmic accountability* poses unique challenges for the study and improvement of algorithmic systems. Understanding these problems as *wicked* introduces some important methodological implications as well. First and foremost, it highlights the value of and—arguably—necessity for qualitative methods as the primary approach to assessing algorithmic systems' accountability and transparency capabilities. Only through qualitative and interpretive inquiry can we make a plausible *causal* determination of factors hindering or supporting algorithmic accountability processes of any scope. While there are certainly many use cases for quantitative methodologies—for instance for the comparative evaluation of a large numbers of case studies—they are wholly unsuitable to help us understand algorithms as socio-technical assemblages in their fluid, heterogeneous and ontogenetic nature. Alternative approaches, including ethnomethodologies or participatory design, are decidedly better suited to create the rich descriptions necessary for such analyses. Furthermore, the diversity of

contexts in which algorithmic assemblages manifest also demand a diversity of methods, forcing us to transcend traditional methodological and disciplinary boundaries, if we are to gain the most relevant insights. Without a doubt, the analysis of, e.g., a criminal recidivism risk assessment system like [COMPAS](#) will require different theoretical and methodological approaches than studying the algorithms behind targeted advertisement systems. An inter-disciplinary, sometimes even eclectic and perhaps seemingly conflicting combination of theories, methods and approaches to studying algorithms, I posit, thus is the foundation on which any attempt at a *satisficing* solution to *algorithmic accountability* and *transparency* must rest.

7.1.2 Practical Contributions

In practical terms, addressing both [SRQ2](#) and [SRQ3](#) and their supplemental questions, the results of the two case studies and the comparative case study leading to the [A³ framework](#) offer the following insights and contributions to the field.

The Value of Participation

The EnerCoach case study yielded important methodological observations on the importance of participatory design methodologies. First, involving stakeholders in the design and implementation of transparency measures clearly helped elevate them to a *critical audience* [28], as they carefully balanced the needs of various stakeholder groups with differing levels of domain knowledge and technical literacy with their own needs and assumptions. As such, applied participatory design methods fundamentally transform the participants' relationship with the socio-technical assemblage they are part of and, subsequently, designing for. Instead of passively influencing the whole assemblage by exercising their individual agency, they are empowered to take an active stance and directly and deliberately contribute to the shaping of not just the social, but also the technical components of the assemblage. As experts on the most problematic aspects of the system, their unique perspective led to specific and material solutions for the problems of *transparency* and *ex-post explainability*.

The [AMAS](#) case study, by comparison, stands in stark contrast to these observations and should, in many ways, be seen as a cautionary tale of design-by-policy instead of design-by-participation. Neither the initial goals of the system—*efficiency*, *effectiveness* and *bias reduction*—nor the concrete implementation and operationalization of the system reflect the needs of the individually affected stakeholders. The most vulnerable and, at the same time, most directly affected group of them—the *jobseekers*—were not even consulted or included in the process at all. Viewed as a socio-technical assemblage, the [AMS](#) as a whole exercised its distributed agency to create the systems [AMAS](#) as a matter of policy. Its parameters were set by the *decision makers* to address a specific immaterial component of the assemblage, namely the scarcity of resources available for the jobseekers. Following the public controversy and critical analyses, not least in the form of our own research on the case, the system was never deployed nation-wide, thus making a definitive

assessment over whether or not it could have reached its stated policy goals difficult. What became clear, however, are the serious and concerning consequences the use of [AMAS](#) may have brought to those directly affected by it.

Thus, [AMAS](#) represents a missed opportunity to involve the stakeholders in a meaningful participatory design process and elevating them to a position of actively shaping the socio-technical assemblage they are a part of. Instead, the role of the jobseeker was relegated to a passive *patient* of the system's agency, and the caseworkers' agency was, by and large, curtailed and forced to conform to yet another set of procedural bureaucratic rules enacted by the technical components of [AMAS](#). The wholly insufficient implementation of the only feature of the system intended to support a better *ex-post explainability*—the explanatory text fragments—just underscores once more how important participation of stakeholders can be: any application of human-centric design methodologies would have clearly identified these deficits at an early stage of the development process. This observation rests on the assumption that this explanatory feature was designed with the best intentions for the *jobseekers* in mind, given that the hypothetical counterfactual—a wilful and cynical attempt at providing a *fig leaf* to pre-emptively silence the critical voices addressing a lack of transparency—would raise entirely different and arguably much darker questions about the *talismanic* [89] function of certain algorithmic features.

Considering the efforts necessary to implement participatory design methodologies, the many financial, bureaucratic and political hurdles that must be overcome to utilize such approaches fully may be prohibitive for many algorithmic systems, regardless of how evident their benefits may be. Arguably, involving a representative group of both caseworkers and jobseekers in the design process for [AMAS](#) would have required committing resources to the process that may have been difficult to allocate. Beyond the specific implementation of fully committed participatory design processes, however, the auto-ethnographic approach taken for the EnerCoach case study also showed the benefits of merging research and practical perspectives in the study of algorithmic systems.

Thus, a possible path forward emerges between full-scale participatory design and completely ignoring stakeholder input: a gradual transformation of traditional software development processes into more participatory and value-sensitive ones. As the results of my own auto-ethnographic engagement show, elevating *system developers* to a *critical audience* first and foremost would help promote professional accountability between peers, and lead them towards more inclusive and participatory software development and design patterns as well. By fostering the epistemological curiosity of those designing and developing algorithmic systems, and by providing them the tools necessary to evaluate their own creations in an accessible and constructive way, we can help turn engineers into *critical scholars of their own practice*.

Tools like the [A³ framework](#) that can readily be applied outside the realm of academia implicitly nudge practitioners to explore the concrete operationalization of the systems they create, and also helps foreground the implicit assumptions many software development processes rest on. Ideally, by employing qualitative, reflective evaluation and assessment tools like the [A³ framework](#), the system developers will also discover their own

gaps in understanding the needs of other stakeholder groups, and thus develop further arguments for more participatory approaches.

The Stake in Stakeholder

As discussed before in Sections 3.1.1 and 6.1.3, the choice of case studies for an in-depth analysis of algorithmic accountability seems deserving of justification, as the differences between EnerCoach and AMAS are both evident and significant.

On the one hand, the AMAS case study revealed the potential for serious negative consequences its use may have had in terms of bias and discrimination of jobseekers. Concluding that AMAS is worthy of analysis due to what is *at stake* for its stakeholders requires little effort to justify: clearly, an algorithmic system whose predictions directly impact the lives of already vulnerable groups of stakeholders should be accountable for its conduct. On the other hand, EnerCoach’s context of application, the moral quality of its intended use as a (seemingly scientific and fact-based) tool in the fight against climate change, and the fact that its use may be recommended, but is still essentially voluntary does not immediately suggest a particularly pressing need for accountability. These seeming contradictions notwithstanding, the situated ethnography of the EnerCoach case study revealed the fundamental problems of a lack of transparency and ex-post explainability in the system, and the specific impacts these issues had the system’s assemblage. For the energy consultants, the members of the EnerCoach working group and hotline, the many hundreds of Swiss communities using the system, and the larger, arguably more abstract goal of combatting climate change, being able to better hold the system and its outputs to account is definitely an important requirement and a worthy goal to strive towards.

Thus, the choice of case studies helps illustrate the importance to look beyond a surface-level analysis of algorithmic systems based solely on their overall goals, but critically question *what is at stake for the stakeholders* before assuming whether or not algorithmic accountability is a pressing issue for a given system. In fact, I would argue, the aforementioned “*banality*” and “*boringness*” of EnerCoach and many other similar systems is one of the most problematic challenges of algorithmic accountability we need to overcome: if not even we as researchers recognize the necessity for accountability in these systems and tend to overlook them in favour of the more problematic, controversial, but ultimately easily identifiable case studies, how are we to expect the practitioners, the software developers and decision makers creating these systems to pay attention to these issues? After all, as much as AMAS as a case study serves as a placative example for the way algorithmic systems can manifest and embody political paradigms and values, and the dangers that come with making these values *durable* [80] as technology, so does EnerCoach exemplify the same principles through its normative impact on the sustainability efforts in Switzerland.

Contrary to the primary focus of this dissertation on algorithmic accountability as a *mechanism*, the (perhaps surprising) conclusion resulting from these arguments may

be the importance of working towards better establishing algorithmic accountability as a *virtue*. We need to reach a common understanding that *any and all* algorithmic systems exercise power in the world, and thus deserve to be held to account for their conduct. Establishing an agreement on this point not just within the academic fields of computer science and [STS](#), or even between the disciplines in Technical and Natural Sciences, Social Sciences and Humanities, but also within the software development industry and, finally, the general public, is of paramount importance. Without it, we will continue struggling to argue for better and more stringent regulation of algorithmic technologies, and we will continue to fall prone to the wilful exploitation of our attention and cognitive biases by predatory algorithmic systems in the attention economy. But if we can establish the consequences of unchecked power—for instance, by demonstrating with tools like the [A³ framework](#) how even non-controversial or “banal” systems can impact its stakeholders—the necessary investments in measures to improve algorithmic accountability will become much easier to accomplish.

Such an understanding will also help level the playing field between the exciting, bleeding-edge algorithmic systems based on the *buzzword-du-jour* technologies and the much more prevalent, ubiquitous algorithmic infrastructure vying for our attention. In the end, the less controversial nature of the systems on the latter end of this spectrum promises easier access and thus more potential for in-depth and profound insights, and, subsequently, a better chance at creating models of best practice applicable to other systems as well. While this dissertation shows how very different methodologies can be applied in the study of algorithmic systems to reveal the insights necessary for an in-depth analysis, that does not mean that some methods are not preferable over others when available. Ethnomethodological approaches certainly represent, in my view, a particularly fruitful approach due to the richness of the resulting analysis. Given the limitations of access that often *a priori* preclude such approaches, however, alternative strategies such as a *critical document analysis* can, demonstrably, be just as promising. In other words, *limited access* should not be taken as an excuse to assume that a given system simply cannot be analysed in a differentiated, critical and rich manner any more than assuming that non-controversial algorithmic systems can not teach us important lessons about algorithmic accountability and transparency.

Framing the Problem, Working on Solutions

Finally, the [A³ framework](#) represents the core, practical contribution of this dissertation to the field of [CAS](#). Its application to the two case studies also provided a concrete example of its usefulness as an analytic lens to investigate specific accountability processes. In the case of EnerCoach, it revealed the complex layers of domain knowledge and algorithmic literacy necessary to achieve a successful accountability process, and pointed to some of the concrete measures implemented as part of the case study’s interventionist phase as potential solutions contributing to a better overall accountability of the system. For [AMAS](#), it helped identify and assess the limitations that the system’s design, implementation and operationalization placed on both the jobseeker’s and caseworker’s *agency* to

participate in a meaningful accountability process, and provided the necessary starting points for potential remedies.

The [A³ framework](#) does not represent a specific set of guidelines or instructions that would guarantee accountability when applied to a given system. As the theoretical contributions of this dissertation on the nature of algorithmic systems show, guidelines that are both universally applicable and, at the same time, specific and concrete, are simply impossible: the intrinsic qualities of the *heterogeneous*, *ontogenetic* and *contextual* nature of algorithmic systems categorically prohibit such an approach. Instead, the [A³ framework](#) offers a structured, theoretically well-founded *process* for evaluation and assessment of algorithmic accountability. Prioritizing *principles of simplicity and versatility* over complexity and specificity, the framework can be applied by researchers and practitioners alike to a variety of algorithmic systems, and may even be used in educational contexts. This strategic decision follows the underlying conviction discussed before that we must not just pick and choose which systems we consider important enough to hold to account, but rather that we need to strive to apply algorithmic accountability *at scale*. To work towards this goal, we must democratize the process of evaluation and assessment and provide the requisite tools to do so to a wider audience than just domain experts in computer science, [STS](#) or accountability studies.

To this end, the [A³ framework](#) is designed to be used at all stages of algorithm development, both as a prospective and prescriptive, as well as a retrospective tool. Applying the framework with the use of *scenarios* and *personas*, as done in this dissertation, demonstrated its usefulness for the analysis and assessment of algorithmic systems where a *lack of access* or *deficits in system-level transparency* would prevent the application of other accountability frameworks, particularly those with a more narrow scope and a stronger emphasis on quantitative measurements. In the same vein, the use of *scenarios* and *personas* is a common approach for human-centred design as well, which means the [A³ framework](#) could easily be applied during the design process to evaluate and suggest prospective accountability processes. On the other hand, the [A³ framework](#) may also be used as a research tool of qualitative inquiry, by posing its guiding questions directly to affected stakeholders and evaluating their answers. Such an application directly based on empiric data collection may be used in conjunction with other frameworks for accountability evaluation to provide richer and more nuanced insights overall.

In either approach, the use and usefulness of the [A³ framework](#) can be improved by the previous understanding of the algorithmic system it is being applied to. A holistic, qualitative and “*thick*” description of the algorithmic system represents the best foundation for the evaluation and assessment the [A³ framework](#) can provide: the more in-depth and inclusively we study these systems, including not just their technical figurations, but also their context of use, their stakeholders, and the immaterial components like norms and values shaping their their socio-technical assemblages, the more we can learn from the use of the [A³ framework](#).

7.2 Limitations

Any piece of academic research inevitable comes with its list of limitations and caveats. This is particularly true when engaging with complex, even *wicked* challenges such as algorithmic accountability or transparency. None of these limitations invalidate the conclusions presented before, but a thorough and critical reflection on these limitations as a core tenet of accountable research further underscores the relevance and validity of these results. In the following, these limitations and caveats are discussed.

Limitations of Taxonomies

First, the use of *taxonomies*, *conceptualizations* and *classifications* make up a considerable portion of this dissertation's contribution. It should be noted, however, that no claim can be made for the universal applicability or accuracy of these taxonomies: there are, and always will be, exceptions that—aligning with the common saying—“*prove the rule*”. Taxonomies should, particularly for a fluid and fast-developing field such as CAS, never be considered complete or final, and must be interrogated with the same critical perspective as the algorithmic systems they are applied to. Taxonomies, by design, also *simplify* and *abstract reality* to make it possible for humans to grasp, and thus carry the danger of encouraging a reductionist view of the entities they order and classify. Consequentially, any taxonomy, classification or conceptualizations presented in this dissertation, be it the taxonomy of various transparency challenges, the micro- vs. macro-accountability perspective, or even the conceptualization of the accountability process itself, must be considered as a tool characterized by its usefulness, and not as a positivist, descriptive depiction of reality.

Methodological Limitations

Secondly, all methodological choices are the result of inevitable epistemological trade-offs, and the methodologies applied in this dissertation and its constituent case studies are no exception to this rule. To expand on the methodological considerations provided in Chapter 3, the following discloses the limitations introduced by the mode of qualitative inquiry and auto-ethnography, participatory design, document analysis, and the comparative case study approach.

Starting with *qualitative inquiry* as utilized for both case studies, a primary limitation always involves the completeness of the picture painted. Naturally, there can be no proof that expanding the interviews to additional stakeholders might not yield further insights that are not yet represented in the case studies. For EnerCoach, that would potentially include stakeholders with more limited domain knowledge (e.g., facility managers or community administrators). This limitation notwithstanding, it stands to reason that the stakeholders that *were* interviewed have a deep understanding of those stakeholders that were not: For instance, the interview partners from the EnerCoach hotline were

specifically chosen because they can provide overarching insights into the needs and wants of the EnerCoach users they work with on a daily basis.

For [AMAS](#), on the other hand, the limitations given by the lack of access to caseworkers or (potentially) affected jobseekers poses a more substantial challenge. Here, the unwillingness of the AMS to cooperate further with our research study by allowing us access to caseworkers necessitated a more speculative approach of analysis. However, as a remediating factor, the wealth and variety of documents analysed provided some plausible insights into the process and reality of casework as part of the [AMS](#) consultation process, and great care was taken in the creation of scenarios and personas (as used in Chapter [6](#), but also in our prior publications [4](#), [3](#) on the system) to ensure their congruence with the documents provided. Nonetheless, further studies involving both caseworkers and, indeed, jobseekers would surely be a worthwhile endeavour. The *document analysis* performed for this case study is necessarily founded in the assumption that the documents represent a particular manifestation of social reality, but not necessarily a truthful, nuanced or, by any measure, objective one. While discrepancies and contradictions between documents often reveal the extent of subjectivity and tensions within the issuing organisation and document authors, no assumption can be made that all such cases were, or even could be, identified in our analysis.

Beyond these specific limitations for the case studies, methods of *qualitative inquiry* pose general epistemological limitations; first and foremost, they offer little quantifiable evidence or metrics for evaluation, and should not be considered valid methods to achieve such results. For instance, attempts to quantify the *transparency* of the EnerCoach system or quantitatively evaluate the measures devised through the participatory design workshops would be inappropriate, and these results were not pursued as a consequence. That is not to say that quantitative approaches could not yield actionable results or provide valuable insights, as the long history of quantitative methods of evaluation in the Social Sciences and Humanities undoubtedly prove. The potential for these methods notwithstanding, the arguments presented in the requisite sections of Chapter [6](#) and these conclusions support the choice of qualitative methods over quantitative ones insofar as they allow a deeper, more causally oriented analysis of the algorithmic systems and processes of accountability embedded therein. In fact, in these parts of this dissertation I offer a nuanced critique of the over-reliance on quantitative methods in many studies originating from the various subfields of Computer Science related to algorithms and algorithmic systems. Consequentially, I would argue that this methodological choice, as much as it may impose limitations, also represents the *opportunity* to showcase the value of qualitative, in-depth studies as a mode of inquiry into algorithmic systems.

Considering this, a primary limitation of these methods does not necessarily lie in their potential for actionable results, but rather in their broader applicability for future research and assessments. Qualitative methods pose significant challenges in terms of *scalability*, as their application often requires significantly more time and resources than quantitative methods may require. While the qualitative methods could be applied very well to the case studies in this dissertation, analyses of larger-scale systems, perhaps with a more

(geographically or otherwise) distributed set of stakeholders, may simply be infeasible. This dilemma between in-depth qualitative assessments and broader quantitative ones extends beyond the analysis of single systems as well. AIAs and algorithm audits, as well as policies and regulation such as the EU’s GDPR [170], the EU’s proposed *Artificial Intelligence Act* [384, 232] or the Canadian Directive on Automated Decision-Making [172] mandating such assessments, are facing the challenge of finding a middle ground between overly broad recommendations and practically infeasible assessment requirements. The current proposals for such assessments, for instance the CapAI tool presented by Floridi et al. [385] for the AI Act, often try to resolve this through a checklist approach that requires qualitative or quantitative supporting evidence for each check mark. In doing so, they argue, the depth or shallowness of the assessment can be adapted to the context of the algorithmic system. The dangers in such approaches, as they concede in their own disclosure of limitations [385, pp.71-72], lies in the potential for simply skipping over complex issues or difficult to assess aspects of the system. After all, having to conform to a binary assessment for each “check mark” also means the temptation to lean toward quantitative over qualitative methods for the supportive evidence looms large in light of considerations of cost effectiveness. In summary, regulatory mandated algorithm audits or assessments should also clarify where qualitative methods are the preferred methods of inquiry to avoid surface-level assessments that lack depth.

Limitations of (Auto-)Ethnographic Methods

Similar to the limitations of qualitative methods in general as outlined above, the specific use of (auto-)ethnographic approaches for analysing algorithmic systems also presents certain trade-offs worthy of careful consideration. While these approaches offer a number of benefits, as described in Section 3.2.1, they require a high level of integrity and personal responsibility of the researchers employing these methods. Floridi et al. present an illustrative analogy in their analysis of limitations of Ethics-based Audits (EBAs) that is equally applicable for (auto-)ethnographic methodologies:

“[T]he best pipes may improve the flow but do not improve the quality of the water, yet water of the highest quality is wasted if the pipes are rusty or leaky. As the pipes in the analogy, no EBA procedure is morally good in itself. However, they can realise moral goodness if adequately designed and combined with the right values.”

[385, p.71]

Employing well-established practices for (auto-)ethnographic inquiry and a clear disclosure of the auto-ethnographic aspects of an analysis thus is the necessary prerequisite for “realising moral goodness” [385, p.17].

Additionally, (auto-)ethnographic approaches also require comparably high levels of access to the subject matter, i.e., the algorithmic system and its socio-technical assemblage in

question. As shown in the discrepancy in methods between the case studies, EnerCoach fulfilled these requirements of access, while [AMAS](#) clearly did not. Thus, this dissertation and its results indicate a recommendation for these methodologies, albeit with the caveat that they may only be feasible in select cases.

Limitations of the A³ Framework

Finally, the [A³ framework](#) itself has already been discussed and evaluated in context with other frameworks and assessment tools for (algorithmic) accountability, as well as in context with [AIAs](#) and algorithm audits in sections [6.2.4](#) and [6.2.4.2](#). To avoid repeating the arguments made there, I offer only a few additional considerations of the framework's applicability and limitations here. Most pressingly, and paralleling the concerns explicated above, the use of the framework itself cannot guarantee a truthful and ethically valid result. Naturally, a hypothetical *bad-faith* assessment of algorithmic accountability with the help of the [A³ framework](#) can be performed that would show excellent results and intentionally obscure shortcomings. Given the recent proposals for ethical evaluations of algorithmic systems, particular for [AI/ML](#) applications, in the form of *Ethics as a Service* [\[386, 387\]](#), it stands to reason that commercial assessments, performed with the help of tools such as the [A³ framework](#), and subsequently nudged towards a desired outcome by those who pay for it could, at least occasionally, occur. While there is no way to avoid such an intentional misuse, for instance as an attempt of *ethics-washing*, the qualitative nature of analysis inherent to the [A³ framework](#) may offer an outside observer at least a chance at detecting the malicious intent behind such an assessment. Even if the qualitative descriptions given as part of the assessment for the different guiding questions are limited and lack detail, that fact alone may serve as an indication of a lack of rigour and thus raise concerns about the validity of the resulting assessment. Purely quantitative assessment tools, by comparison, may not offer the same chances at detection due to the limiting nature of the (numeric or categorical) results.

In terms of evaluating the [A³ framework](#)'s applicability for a wider range of algorithmic systems, I have presented supporting arguments suggesting it can be used for a wide variety of algorithmic systems and accountability processes. These arguments notwithstanding, the framework has, to this day, only been applied to the two case studies presented in this dissertation. Considering that the initial, iterative and emergent process of designing the framework happened through the comparison of these same case studies, one might have valid concerns over *self-referentiality*. Were the [A³ framework](#) the result of the use of machine learning methods, its validity would need to be seriously questioned as a text-book example of *overfitting*. However, qualitative research is not subject to the same limitations of self-referentiality through inference, and the application of a framework to the case studies prompting its design is not an uncommon practice in the field. The explication of significant differences between the case studies as presented in Section [6.1.1](#) further supports this assumption.

Lastly, a final note on the foundational conceptualization of (algorithmic) accountability informing the [A³ framework](#) is required. While the procedural definition of accountability

as presented by Bovens [22] is most widely used in CAS and related disciplines as shown by Wieringa [23], other competing definitions and concepts for (algorithmic) accountability exist. Kacianka and Pretschner, for instance, provide an overview of such competing definitions, including those rooted in Social Science, Psychology and Organizational Sciences [388, p.3] in addition to the one given by Bovens. Their study also provides an example for an approach to assessing accountability that aims to unify these definitions and perspectives through a different analytic lens, namely *causality*. In contrast, the A³ framework and its focus on *human agency* as its analytic lens only has limited applicability for some of these other definitions of accountability. Thus, for algorithmic systems where Bovens relational and procedural approach may prove less fitting, the A³ framework may not be as useful or applicable either. The number of such instances, however, can be expected to be small, given the *metaphorical* [163] nature of Bovens definition and the resulting *malleability* of this approach. As one primary goal of the A³ framework was not to integrate the *most diverse* definitions of accountability, but focus on the *most widely applied and applicable* one, this limitation constitutes an acceptable trade-off.

7.3 Future Work

Characteristically, good academic research opens up further lines of inquiry as much as it presents answers to previously posed questions. While it is important not to exceed the scope of given research projects such as the ones underlying this dissertation, presenting potential future research endeavours and topics is the foundation upon which the scientific exchange of knowledge rests. To this end, I present to following promising topics that warrant further attention.

In Section 2.4.3, I introduce the notion of *micro-accountability* as a counterpart to the prevalent notion of macro-accountability processes. As the permeation of society with algorithmic systems continues to increase, I posit, investigating the requirements for micro-accountability, studying those micro-accountability processes emerging from the use of these systems, and designing and evaluating measures to support such processes will become more and more relevant. Multiple possible lines of inquiry are worth considering. First, in order to gain a deeper understanding of user expectations and needs in terms of micro-accountability, research in HCI and CSCW, leaning on a critical understanding of the impact and power of different algorithmic systems, should focus on the study of these types of *human-to-computer* and *human-to-human* interactions. Based on the research presented in this dissertation, it is highly likely that these needs and requirements differ significantly between various types of algorithmic systems, user interfaces, stakeholders and their requisite levels of algorithmic literacy and domain knowledge. Further research outlining methods to determine the overall *need* for micro-accountability processes, as well as the specific forms these processes can and should take, is required. Considering inter-disciplinary inquiry by merging methodologies and prior knowledge from disciplines outside of computer science, including Psychology, STS, Sociology, Law and the domain-specific disciplines applicable to the algorithmic system in question is highly recommended.

Furthermore, a better understanding of the relationship between *micro-* and *macro-accountability* could provide valuable insights into the value of better micro-accountability capabilities of algorithmic systems to pave the way for faster, more streamlined, and overall more impactful macro-accountability processes. We need to investigate more closely how a successful—or failed!—micro-accountability process can be escalated into macro-accountability processes to achieve sustainable, positive change for a given algorithmic system. The analytic lens of *emergent interactive agency* can continue to be a helpful tool in conceptualizing and realizing such a transition as well, and guide more specific research foci into these processes as well.

To avoid a purely *ex-post* perspective at micro- and macro-accountability processes that only serves to analyse and highlight the shortcomings of already established algorithmic systems, further research into guidelines and tools that can support these processes during the entire algorithmic lifecycle is sorely needed. Establishing a set of best practices to design algorithmic accountability measures with a human-centric perspective could help practitioners choose a path forward not just for a single algorithmic system, but perhaps entire classes of algorithmic systems. It stands to reason that, for instance, the micro- and macro-accountability requirements for the variety of credit scoring or risk-assessment systems share similar characteristics, and that commonalities can be formulated into such modular best-practice measures applicable to most future systems of a certain class or in a given domain. These *best practice examples* can also be considered a valuable contribution to accountability policy and regulation: certifying certain practices as applicable and sufficient could lead to an expedited assessment process for AIAs or other audit procedures.

This dissertation also adds to the growing chorus of scholarly voices calling for better, more applicable algorithmic regulation mandating such assessments, particularly for those application domains where the use of algorithmic systems has a tangible, and potentially negative, impact on humans, or those algorithmic technologies particularly prone to such impacts (i.e., AI/ML technologies). Such regulation could help curtail the purely profit-driven development of potentially dangerous technologies private enterprises and large technology corporations engage in. How best to design such regulation—both from a *legislative* and *executive* standpoint—will undoubtedly continue to be a significant challenge and the focus of further research of scholars in Law, Political Science and STS.

Beyond the private sector, academia itself must consider integrating ethical considerations and the study of potential social implications more fundamentally into technology development and innovation as well. While regulation of academic research conflicts with the fundamental freedom of Science and thus may not be an appropriate way to ensure this, we all, as members of the global scientific community, simply must do better in regards to recognizing and accepting moral responsibility for the potential consequences of our research. To do so, new approaches to how we teach these concepts in higher education and how we can impress upon our students this moral responsibility will be needed, along with a stronger commitment to interdisciplinarity as well. For technology research, we must develop more and better suitable tools and guidelines

that help us evaluate technological innovations and new techniques in regards to their potential for supporting or hindering these micro- and macro-accountability processes. Ideally, these tools—including guiding frameworks such as the [A³ framework](#)—can help avoid developing technologies that may fulfil specific technological requirements at the cost of human agency and accountability, and put the focus on human-centric innovation.

At the same time, we must accept the practical limitations for micro-accountability processes posed by the large scale and distribution of algorithmic systems, and develop alternative methods for such processes that do not rely on purely human-to-human accountability processes. In Section [2.4.5](#), I introduced the notion of *artificial accountability* as an accountability relationship between human and non-human agents, not as a replacement, but as an addition to the taxonomy of human-to-human accountability processes. Further research—both in terms of basic and applied research—is required to explore the potential for such artificial agents to participate in accountability processes as the *actor* giving an account to various human *fora*. As we look into existing automated customer relations systems—from automated telephone support systems to chat bots—and conceptualize some of the processes as a form of *artificial accountability*, we can learn about the existing approaches, their shortcomings and strengths, and suggest improvements that can lead to empowering the stakeholders of these systems and elevate its users to a *critical audience*. Here, like in almost all of the previously suggested lines of inquiry, an inter-disciplinary approach should be considered most promising to integrate the various methodologies, insights and perspectives offered by both Natural or Technical Sciences and Social Science and the Humanities.

Finally, the work presented in this dissertation culminated in the proposal of the [A³ framework](#). To address some of the limitations outlined in the previous section, further research into the applicability and usefulness of the framework itself and the larger approach of human agency as an analytic lens is required. This may lead to a refinement of the [A³ framework](#) and the development of more granular, domain- or application-specific versions that maximise its impact. Applying the framework to a variety of algorithmic systems will also help highlight its deficits and suggest avenues for such refinements. Finally, the simplicity and versatility of the analytic lens of human agency for accountability processes shows significant potential not just in research, but in teaching as well, and should be evaluated in the context of higher education courses on [CAS](#) and its related disciplines.

Lastly, evaluating the framework in the context of *algorithm audits* and [AIAs](#) as discussed in Section [6.2.4.2](#) provided directions for further research to adapt the framework for, and integrate it into, these processes as a complementary methodology to other qualitative of quantitative research methodologies. Adapting the framework to be better suit the requirements of *regulatory inspections* and *algorithmic impact assessments* will certainly be a worthwhile endeavour.

7.4 Final Remarks

At the beginning of this dissertation, I posed the following primary research question:

“How can algorithmic systems, in all their heterogeneity, complexity and various application domains, be analysed, designed and improved to satisfy higher standards of accountability towards its stakeholders, affected humans and society at large?”

Attempting to answer this question directly might either suggest a fundamental lack of understanding of the problem, or signify an almost comical case of *hubris*. To avoid the impression of either, I offer the following learnings one might take from this dissertation not as an answer, but—in line with the Buddhist’ Lojong [389, p.145] slogan to “*abandon any hope of fruition*”—as a motivation for a further pursuit of this question despite its apparent lack of answerability.

Algorithmic systems are, undoubtedly, “*here to stay*”, and to suggest otherwise would be a very difficult argument to make indeed. Similarly, their real and tangible impacts on our world are undeniable. Whether or not we think these impacts are, by and large or even just for specific instances, *positive* or *negative*, should make no difference to the implicit normative assumption made in the research question above, namely that impactful technologies must “*satisfy higher standards of accountability*”. If we can agree on that point alone, I posit that we have a shared moral obligation as scientists and academics, as engineers and sociologists and scholars of Science and Technology Studies, as algorithm developers and as algorithm users, to pursue further insights that can contribute both to our understanding of the question and, bit by bit, to its theoretical and practical answers. Complex questions demand complex answers, and therein may lie the simplest insight to help address this primary question: no single discipline, or individual researcher, can provide the tools and knowledge required for as *wicked* a problem as *algorithmic accountability*.

To further our pursuit of answers to this problem, *embracing inter-disciplinary collaboration* between the social, technical and natural sciences, and boldly challenging disciplinary and methodological divides truly is the only plausible path forward towards better, more trustworthy, useful, safer and—ultimately—more *accountable algorithmic systems*.

Appendix

A.1 EnerCoach Stakeholder Interview Guideline

The following interview guidelines (in their original German version) were utilized as part of the EnerCoach case study.

1. Stakeholders

- (a) Welche Rolle nimmst Du im Rahmen des EnerCoach-Projekts ein?
 - (i) Was sind Deine Aufgaben und Verantwortlichkeiten?
 - (ii) Welche Expertisen bringst Du in das Projekt mit ein?
 - (iii) Welchen Prozentsatz Deiner beruflichen Tätigkeit nimmt das EnerCoach-Projekt in etwa ein?
- (b) Welche anderen Stakeholder kannst Du identifizieren?
 - (i) Welche Rollen nehmen sie ein, was sind ihre Aufgaben und Verantwortlichkeiten?
- (c) Welche Organisationen sind an dem Projekt (direkt oder indirekt) beteiligt?
 - (i) Welchen Einfluss nehmen diese Organisationen?

2. Projekt-Struktur und Eigenschaften

- (a) Wie lassen sich die Zielsetzungen des EnerCoach Projektes beschreiben?
- (b) Welche Funktionen erfüllt das Projekt für die BenutzerInnen, die BetreiberInnen, und andere Stakeholder?
 - (i) Wie verwenden die BenutzerInnen die Ergebnisse der Algorithmen?
 - (ii) Welche Konsequenzen haben die Ergebnisse der Algorithmen für die Stakeholder?

- (iii) Welche worst-case Szenarios ergeben sich aus ‘falschen’ Eingaben oder Berechnungen?
- (c) Wie ist das Projekt entstanden, was war die Entwicklungsgeschichte des EnerCoach Tools?
- (d) Wie ist das Projekt finanziert?
- (e) Wer entscheidet über Implementationsdetails der Algorithmen (Reports, Datenstrukturen, Grunddaten wie Energieträger, etc.)
- (f) Wie und von welchen Entscheidungsträgern werden Entscheidungen über potentielle Weiterentwicklungen & Anpassungen getroffen?

3. Algorithmic Transparency, Accountability & UserInnen-Wissen

- (a) Wie würdest Du den aktuellen Zustand des Tools in Bezug auf die Transparenz der eingesetzten Algorithmen einschätzen?
 - (i) Welche Aspekte von Transparenz des Systems sind gut umgesetzt, wo siehst Du noch Aufholbedarf?
 - (ii) Woher beziehst Du Deine Einschätzung?
- (b) Wie würdest Du den Wissens- und Verständnisstand der BenutzerInnen in Bezug auf das Tool sowie die zugrundeliegenden Algorithmen charakterisieren?
 - (i) Kannst Du eine Schätzung abgeben, welcher Prozentsatz der BenutzerInnen mit hohem, mittlerem und geringem Wissensstand ausgestattet sind?
- (c) Welche Weiterbildungsmaßnahmen in Bezug auf EnerCoach gibt es für die BenutzerInnen?
 - (i) Sind diese freiwillig oder verpflichtend?
- (d) Welche anderen Informationsquellen oder Ressourcen stehen den BenutzerInnen zur Verfügung, um sich über das System und die verwendeten Algorithmen zu informieren?
- (e) Welche Einschränkungen bezüglich der Information über Implementations-Details der Algorithmen gibt es?
 - (i) Gibt es Dinge, die die UserInnen nicht über das System wissen sollen oder dürfen, und wenn ja, warum?

A.2 EnerCoach Reporting Sample Screenshots

The following figures showcase the EnerCoach reporting capabilities through a series of screenshots.

Although the data depicted are sample data from a demo community, the community in question is an amalgamated, plausible example modelled after real EnerCoach communities. As such, this community was created by the EnerCoach Working Group and is being used in training sessions to illustrate the functionalities and capabilities of the EnerCoach system with a realistic data set.

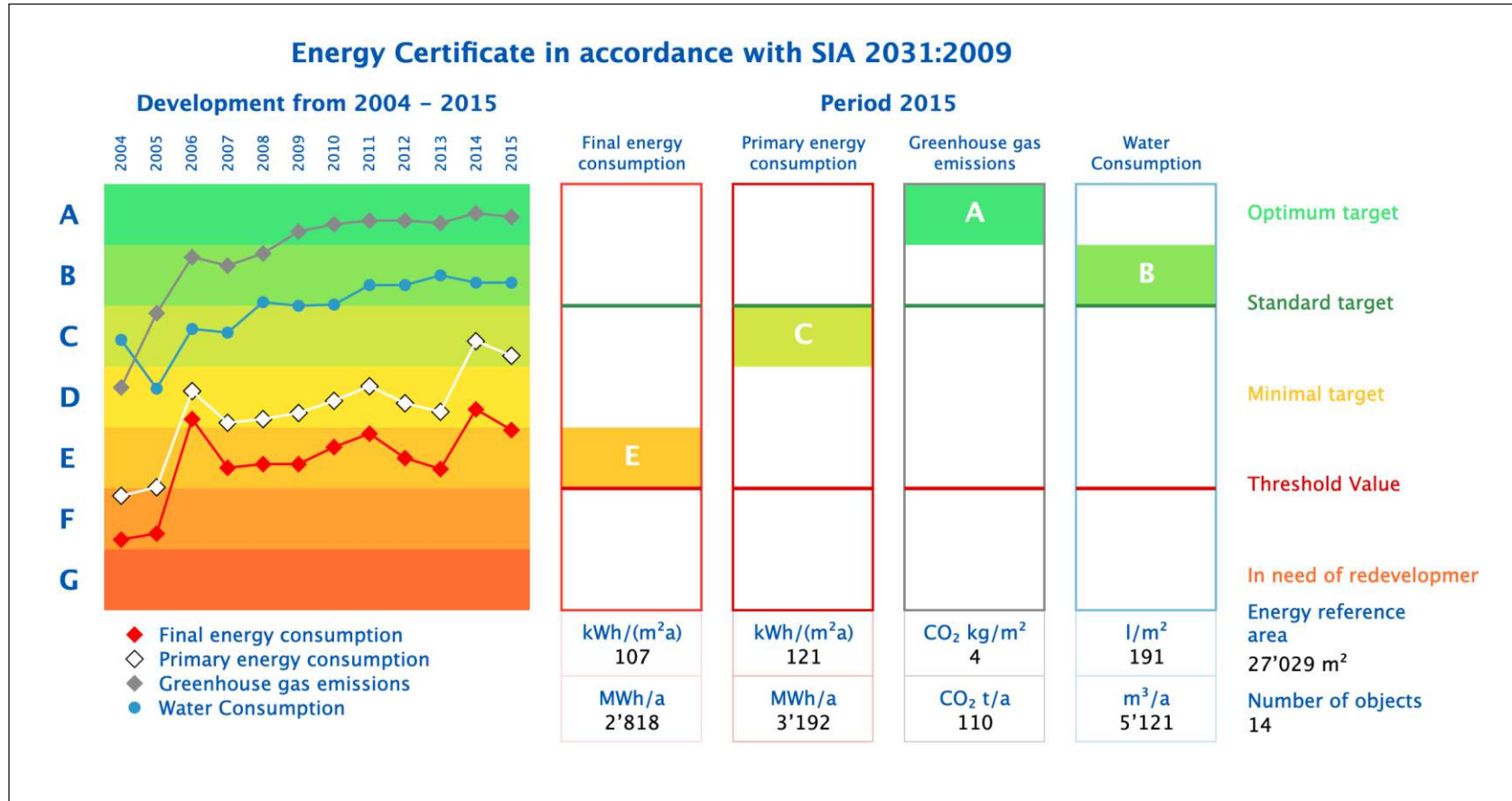


Figure A.1: Energy certificate report showing the sustainability performance of a sample community.

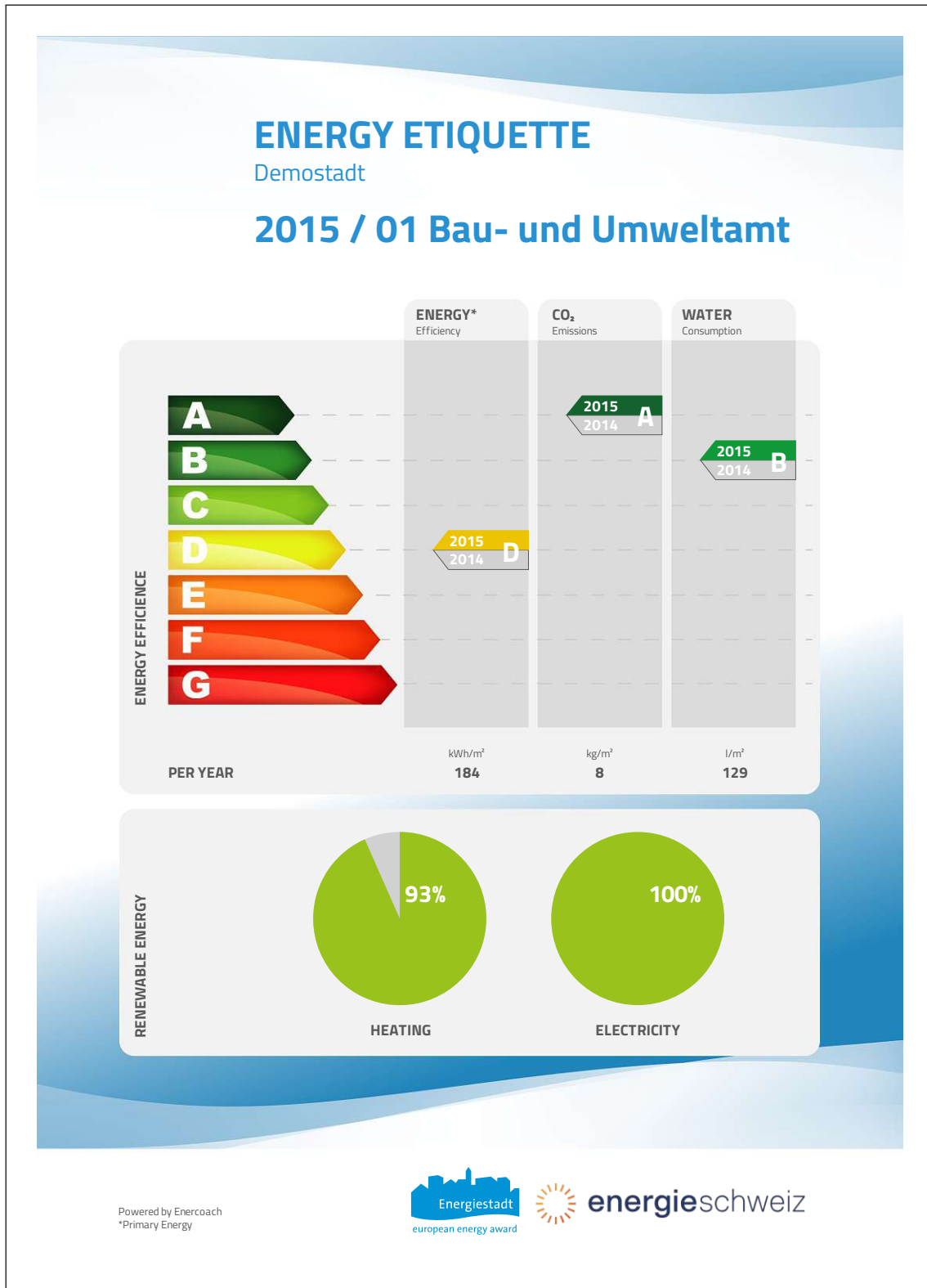


Figure A.2: Poster report showing the report plaque of a sample building.



Figure A.3: Key Figure Reports (Electricity, Heat and Water) of a sample community.

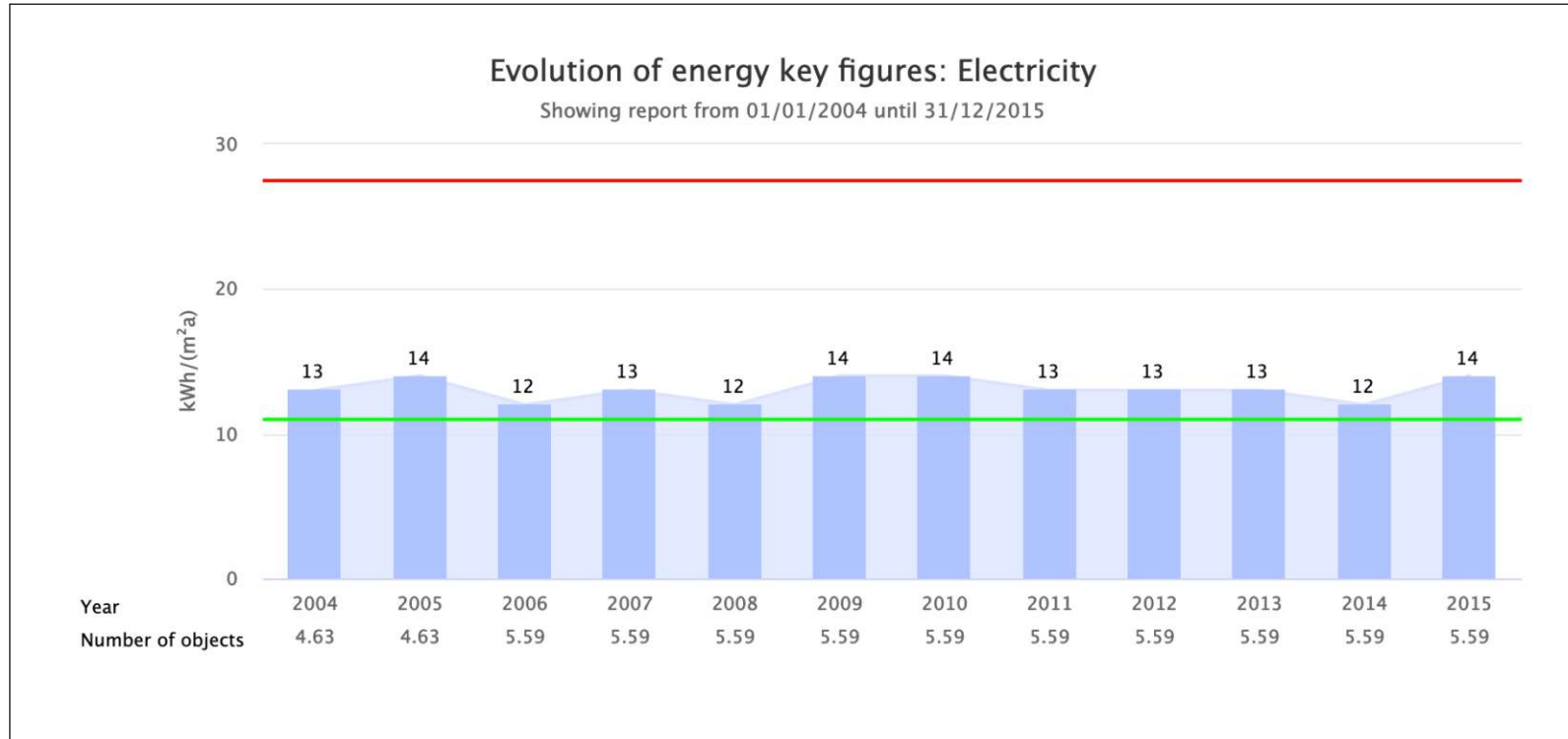


Figure A.4: Key Figure Report (Electricity), filtered by building category 'schools' and showing target and threshold values for key figures of a sample community.

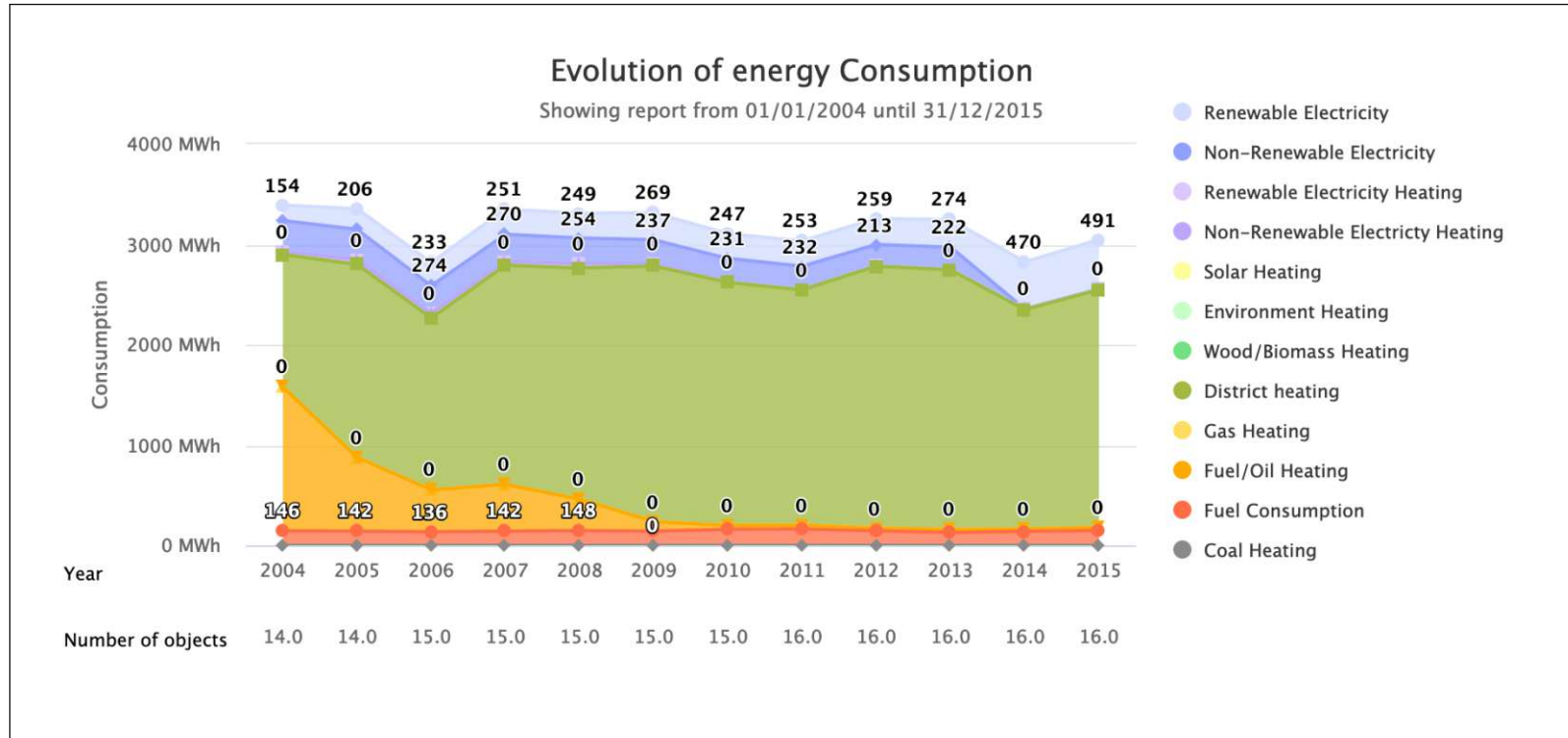


Figure A.5: Evolution of energy consumption report of a sample community.

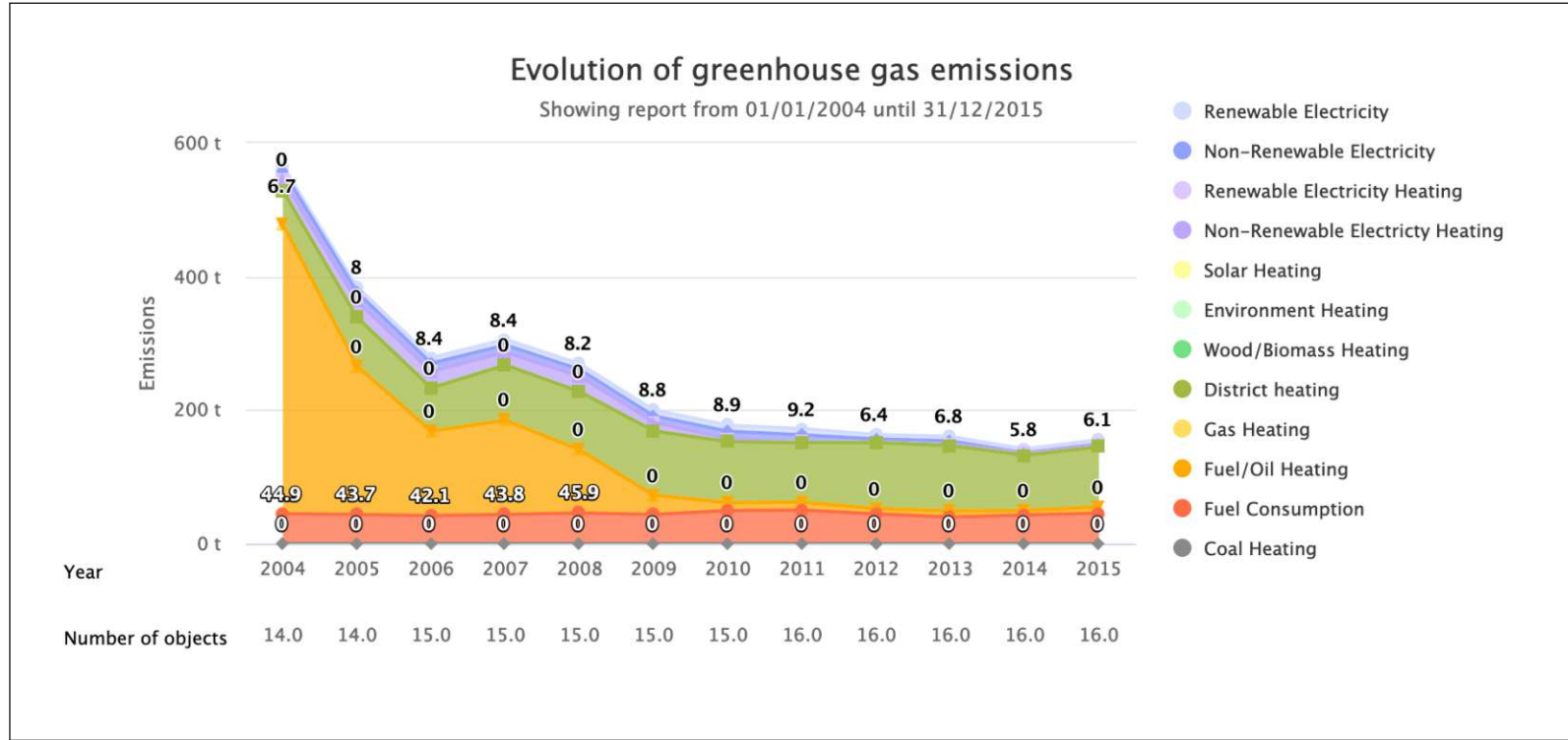


Figure A.6: Evolution of greenhouse gas emissions report of a sample community.

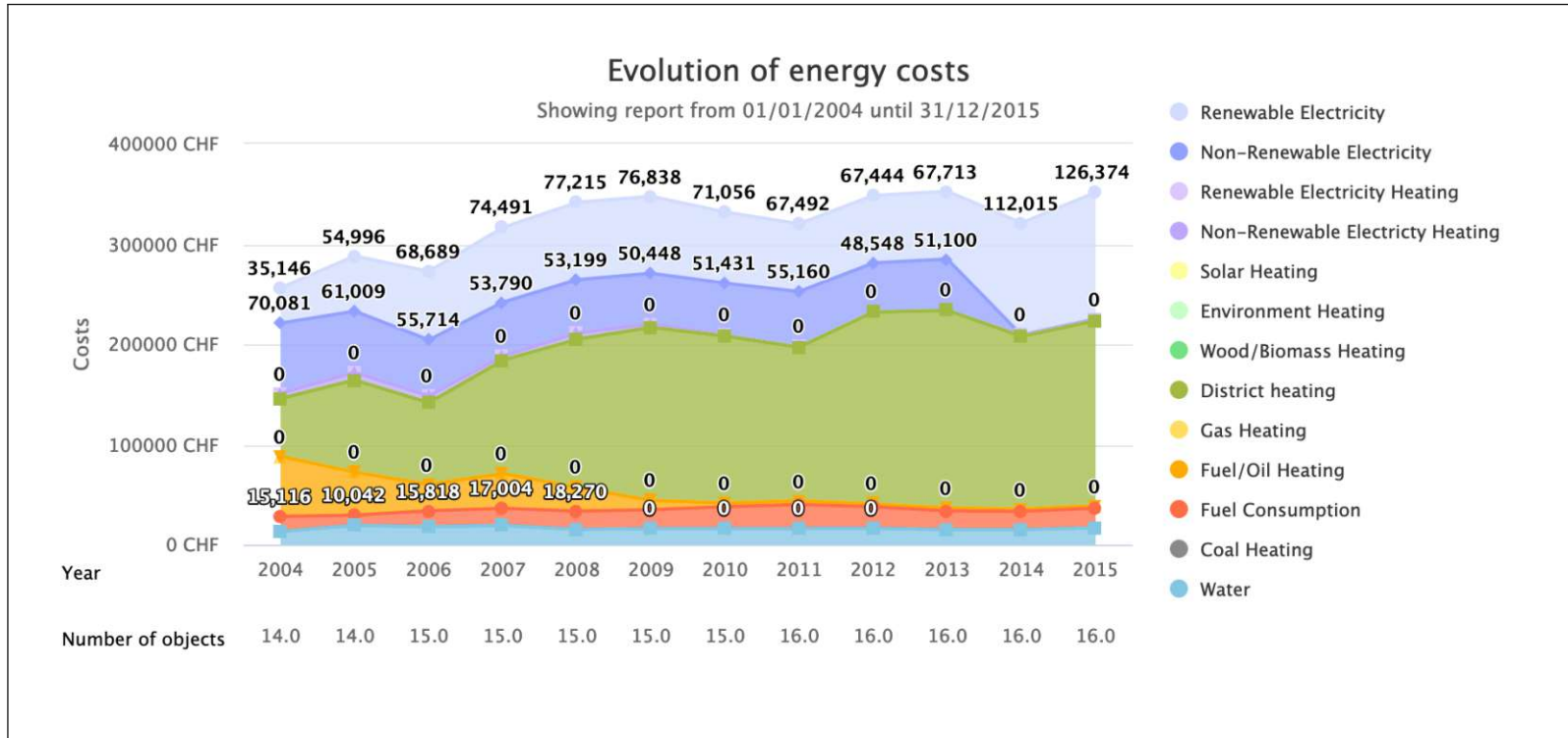


Figure A.7: Evolution of energy costs report of a sample community.

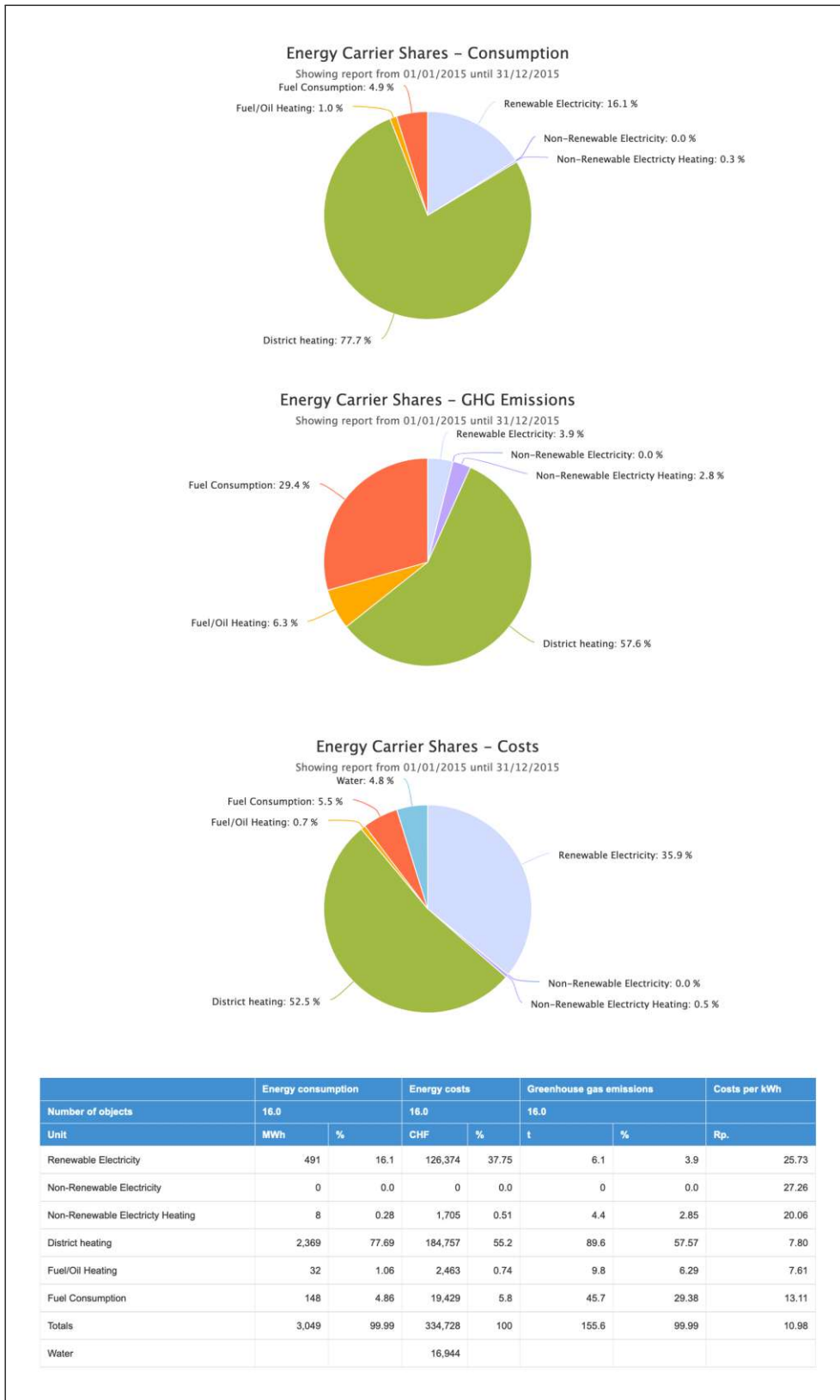


Figure A.8: Energy carrier shares report of a sample community.

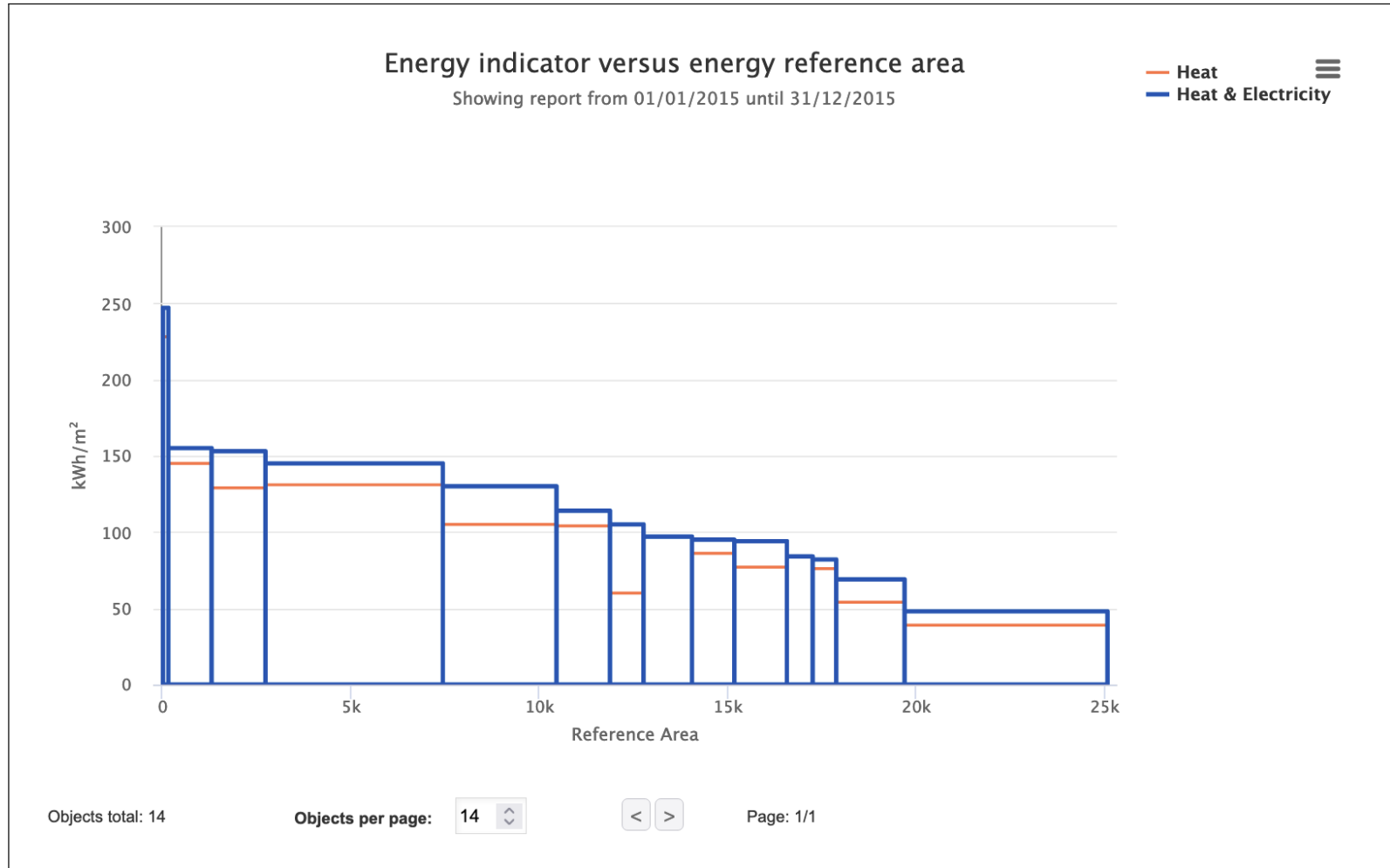


Figure A.9: Energy indicator vs. reference area report of a sample community.

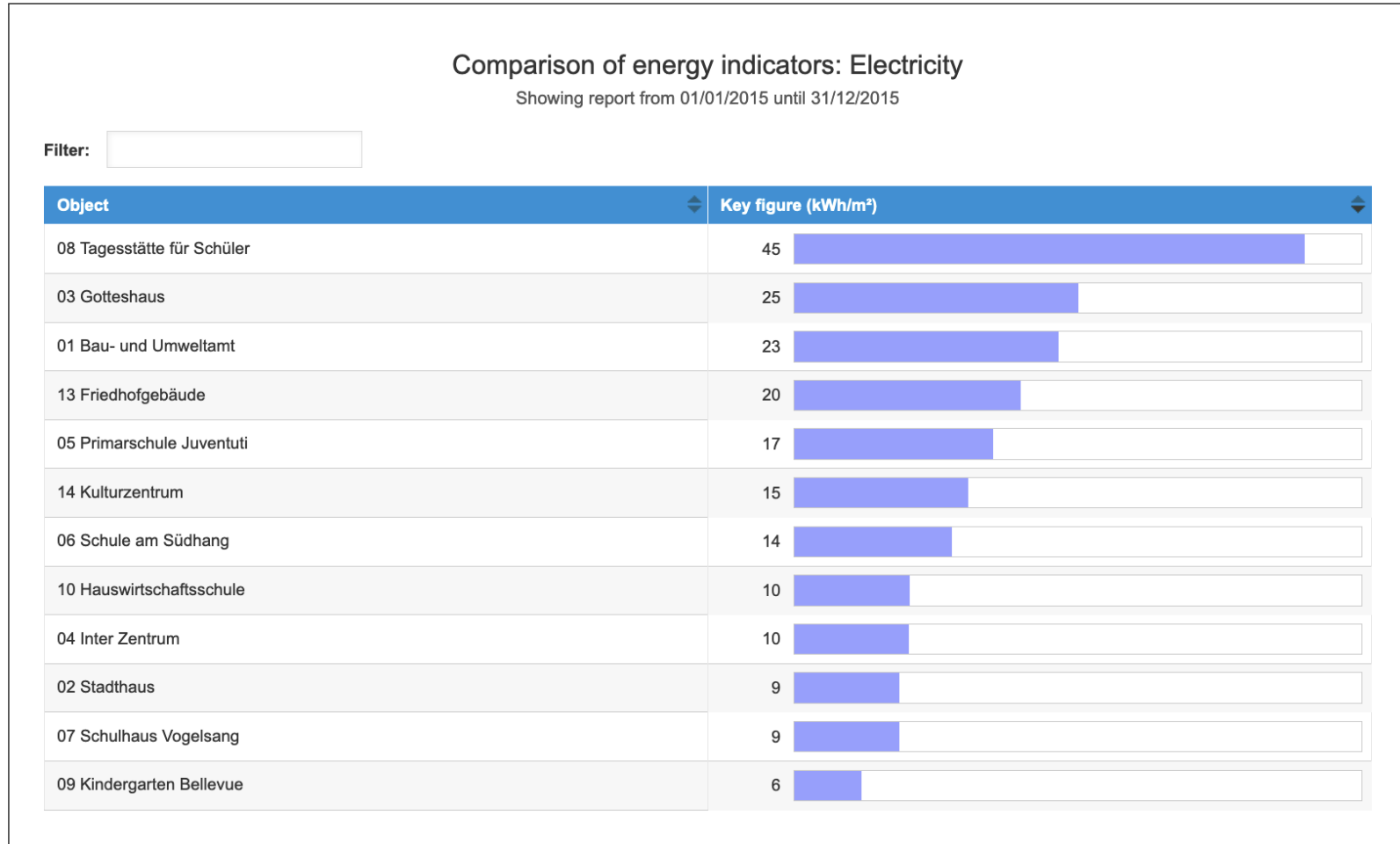


Figure A.10: Key figure comparison (electricity) report of a sample community.

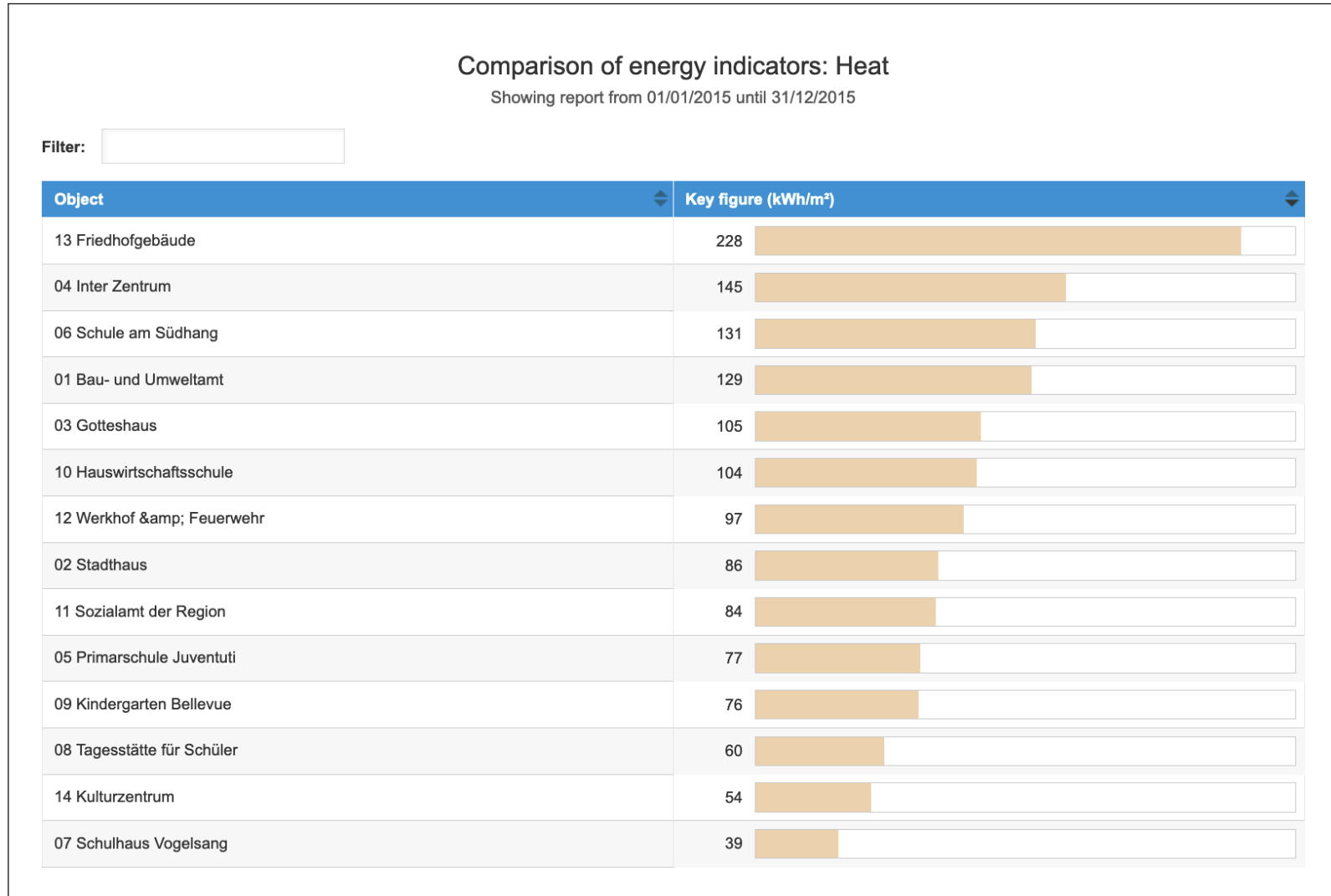


Figure A.11: Key figure comparison (heat) report of a sample community.

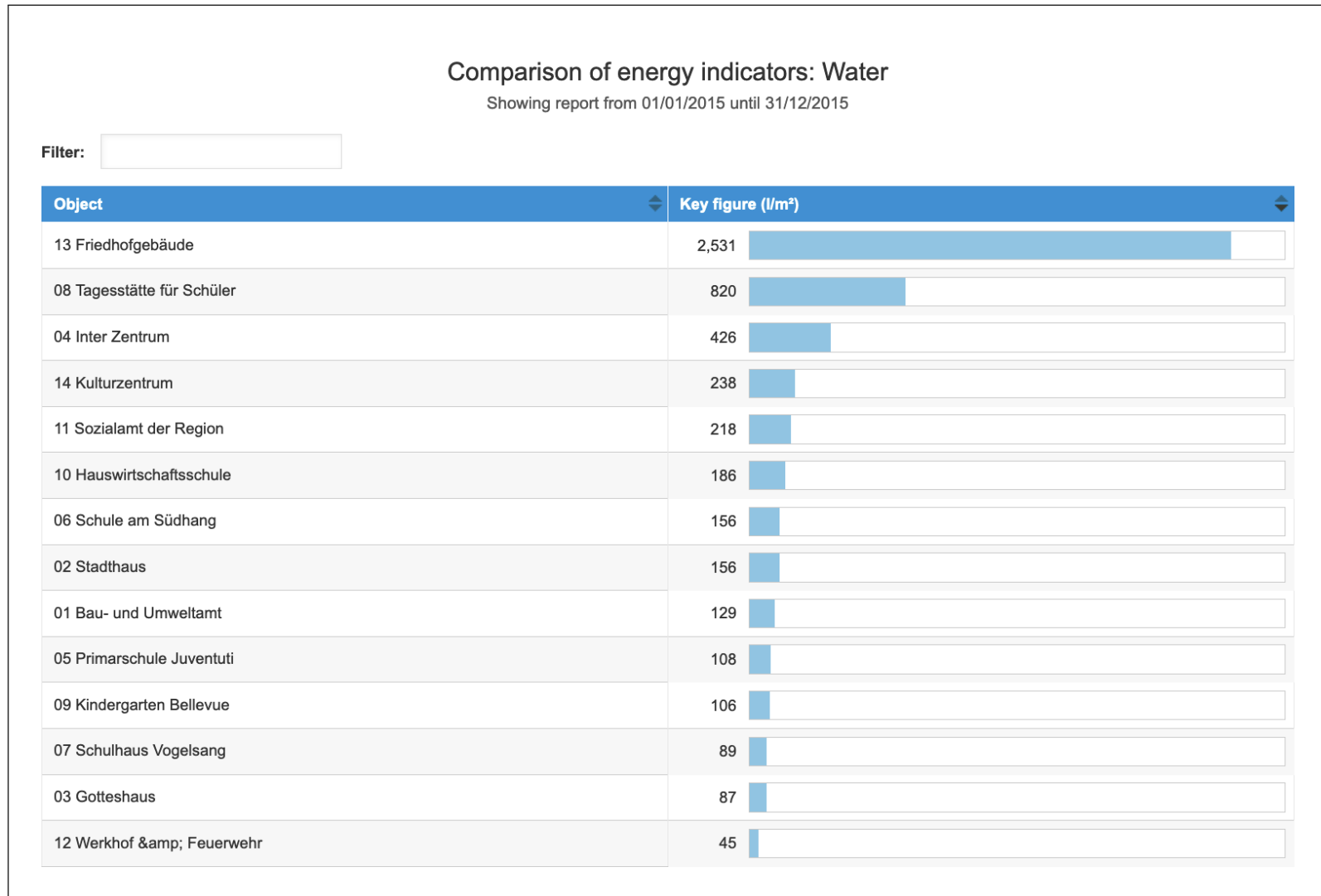


Figure A.12: Key figure comparison (water) report of a sample community.

| Energy City - Renewable Energy: Electricity | | |
|--|------------------------------|--------------------------|
| Showing report from 01/01/2015 until 31/12/2015 | | |
| Electricity Consumption/Production for community buildings | MWh | MWh |
| Total Electricity Consumption of public buildings | 2,289 | |
| Renewable Electricity Consumption as designated | Renewable electricity | Thereof Certified |
| Hydropower | 0 | --- |
| Other Renewable | 0 | --- |
| Promoted Power | 0 | --- |
| Own Facilities / Purchased Certified Electricity | | |
| Hydropower | 2,280 | 461 |
| Solar Power (Photovoltaics) | 9 | 9 |
| Wind Power | 0 | 0 |
| CHP Wastewater Treatment Plant (Biogas) | 0 | 0 |
| CHP Waste Incineration Plant (50%) | 0 | 0 |
| CHP Biomass (Biowaste, Wood, Biogas, etc.) | 0 | 0 |
| Other Facilities (new, renewable) | 0 | 0 |
| Total Electricity (renewable) | 2,289 | 470 |
| Share of Total Electricity Consumption | 100.0 % | 21.0 % (21.0 %) |

Potential 8.0 Points; Rating 52.0%

Figure A.13: EnergyCity: Renewable energy (electricity) report of a sample community.

| Energy City - Renewable Energy: Heat | | | |
|---|---------------------|-----------------|------------------|
| Showing report from 01/01/2015 until 31/12/2015 | | | |
| Energy Carriers | Thermal Consumption | Renewable Share | Renewable energy |
| | MWh | % | MWh |
| Renewable energy | | | |
| Solar thermal | 0 | 100.0 % | 0 |
| Environmental heat | 0 | 100.0 % | 0 |
| Renewable combustibles | | | |
| Wood energy | 0 | 100.0 % | 0 |
| Biogas | 0 | 100.0 % | 0 |
| Fossil combustibles | | | |
| Natural Gas | 0 | 0.0 % | 0 |
| Propane, butane | 0 | 0.0 % | 0 |
| Fuel EL | 32 | 0.0 % | 0 |
| Other | 0 | 0.0 % | 0 |
| District heating | | | |
| Defined district heating | 2,369 | 98.0 % | 2,322 |
| Electricity - heat | | | |
| Heatpump (defined electricity mix) | 8 | 0.0 % | 0 |
| Electricity (direct heating) | 0 | 0.0 % | 0 |
| Totals | 2,410 | 96.3 % | 2,322 |

Potential 8.0 Points; Rating 100.0%

Figure A.14: EnergyCity: Renewable energy (heat) report of a sample community.

Energy City - Greenhouse Gas Intensity

Showing report from 01/01/2015 until 31/12/2015

| Heat | | | | | | | | |
|---------------------------|--------------|-----------------|--------------|-----------------------|-----------------------|-----------------------|-----------------------------------|----------------|
| Category | # of Objects | Reference Value | Emissions | Key figure | Target Value | Threshold Value | Weights | Target Reached |
| | | m ² | t | kg/(m ² a) | kg/(m ² a) | kg/(m ² a) | % | % |
| I Apartment building | 1.7 | 1,969 | 18.5 | 9.42 | 15.1 | 37.75 | 10.3 % | 100.0 % |
| II Family house | | | | | | | | |
| III Administration | 3.4 | 4,315 | 22.5 | 5.22 | 10.9 | 27.25 | 16.3 % | 100.0 % |
| IV Schools | 5.5 | 13,228 | 38.6 | 2.92 | 11.2 | 28 | 51.4 % | 100.0 % |
| V Commercial | | | | | | | | |
| VI Restaurants | | | | | | | | |
| VII Meeting places | 1.5 | 3,376 | 14.9 | 4.41 | 13.2 | 33 | 15.5 % | 100.0 % |
| VIII Hospitals | | | | | | | | |
| IX Industry | | | | | | | | |
| X Warehouses | 0.7 | 1,300 | 4 | 3.04 | 8.1 | 20.25 | 3.7 % | 100.0 % |
| XI Sports buildings | 0.1 | 427 | 2.6 | 6.19 | 14.6 | 36.5 | 2.2 % | 100.0 % |
| XII Indoor swimming pools | | | | | | | | |
| Total | 12 | 24,616 | 101.2 | 4.11 | | | Weighed Target Fulfillment | 99.3 % |

Grasped percentage of the energy reference area of all municipally owned buildings

Potential 4.0 Points; Rating 99.30 %

| Electricity | | | | | | | | |
|---------------------------|--------------|-----------------|------------|----------------------|-----------------------|-----------------------|-----------------------------------|----------------|
| Category | # of Objects | Reference Value | Emissions | Key figure | Target Value | Threshold Value | Weights | Target Reached |
| | | m ² | t | g/(m ² a) | kg/(m ² a) | kg/(m ² a) | % | % |
| I Apartment building | 1.0 | 426 | 0.1 | 291.87 | 8.9 | 22.25 | 2.7 % | 100.0 % |
| II Family house | | | | | | | | |
| III Administration | 2.0 | 3,156 | 0.8 | 263.12 | 11 | 27.5 | 24.6 % | 100.0 % |
| IV Schools | 5.6 | 13,228 | 1.9 | 141.57 | 5.1 | 12.75 | 47.8 % | 100.0 % |
| V Commercial | | | | | | | | |
| VI Restaurants | | | | | | | | |
| VII Meeting places | 2.0 | 3,376 | 0.5 | 143.06 | 8 | 20 | 19.1 % | 100.0 % |
| VIII Hospitals | | | | | | | | |
| IX Industry | | | | | | | | |
| X Warehouses | 0.2 | 616 | 0.1 | 176 | 5.9 | 14.75 | 2.6 % | 100.0 % |
| XI Sports buildings | 0.1 | 427 | 0.1 | 191.77 | 7.2 | 18 | 2.2 % | 100.0 % |
| XII Indoor swimming pools | | | | | | | | |
| Total | 10 | 21,230 | 3.5 | 164.9 | | | Weighed Target Fulfillment | 98.9 % |

Grasped percentage of the energy reference area of all municipally owned buildings

Potential 4.0 Points; Rating 98.90 %

Potential 8 Points; Rating 99.1%

Figure A.15: EnergyCity: Greenhouse gas intensity report of a sample community.

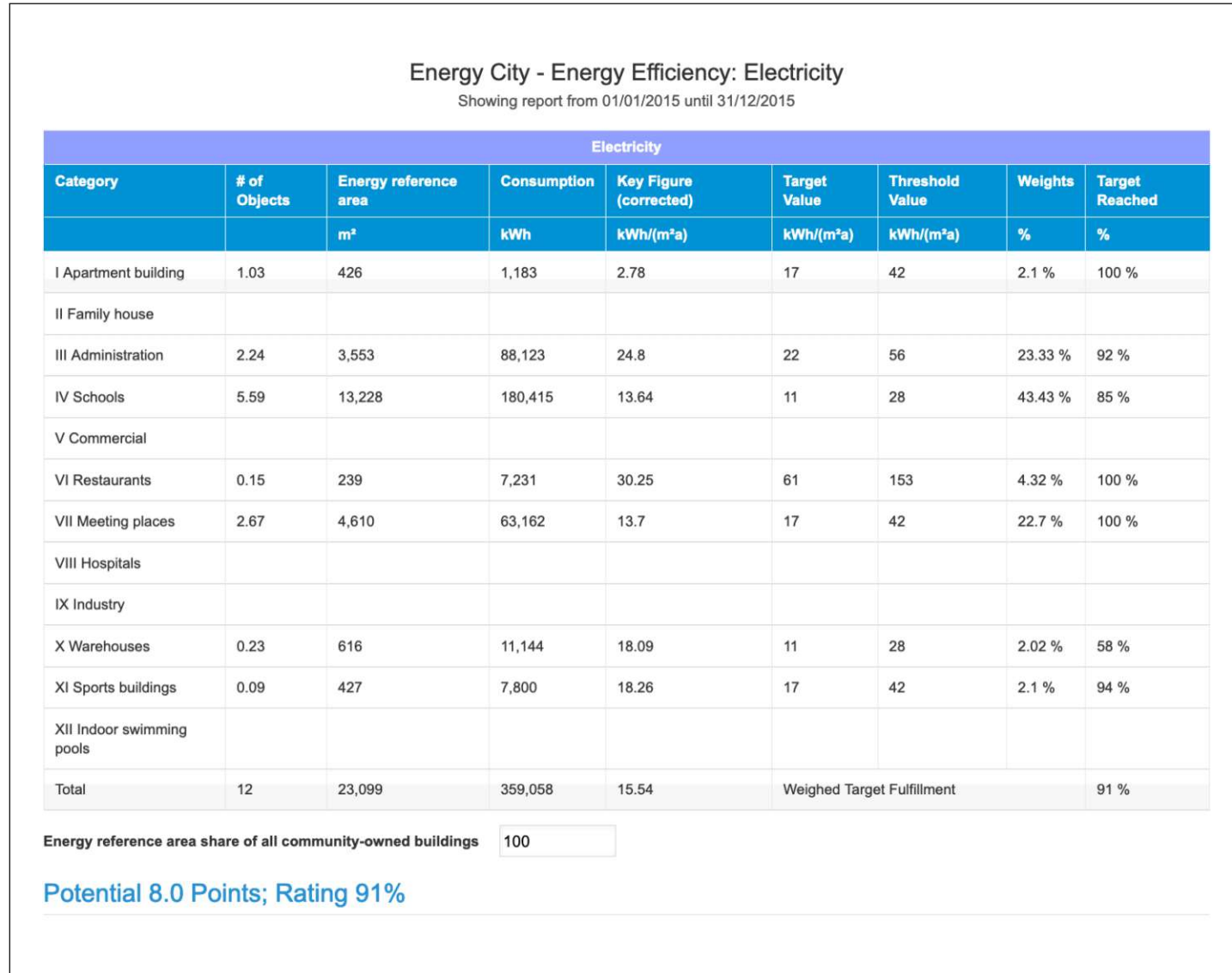


Figure A.16: EnergyCity: Energy efficiency report (electricity) of a sample community.

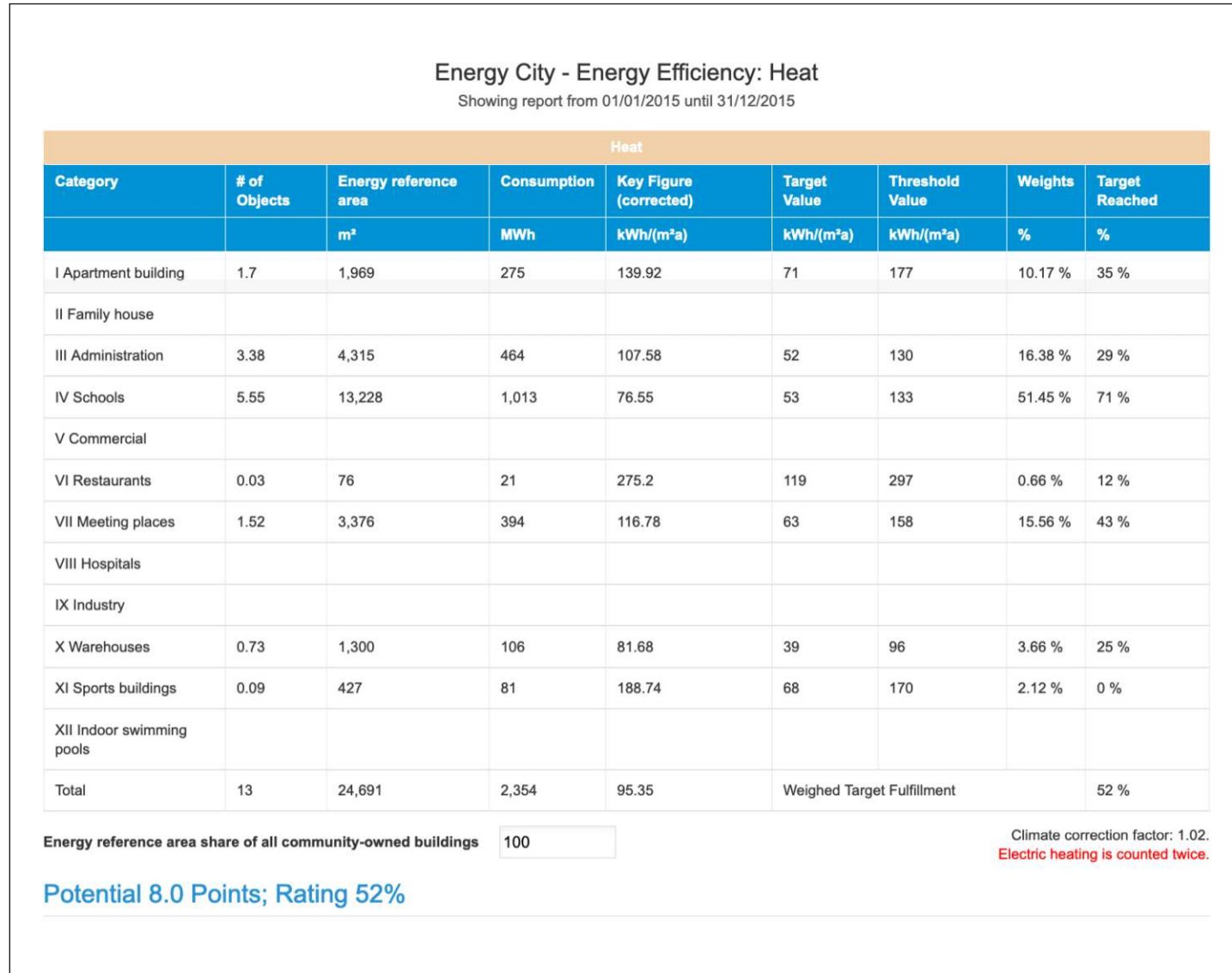


Figure A.17: EnergyCity: Energy efficiency report (heat) of a sample community.

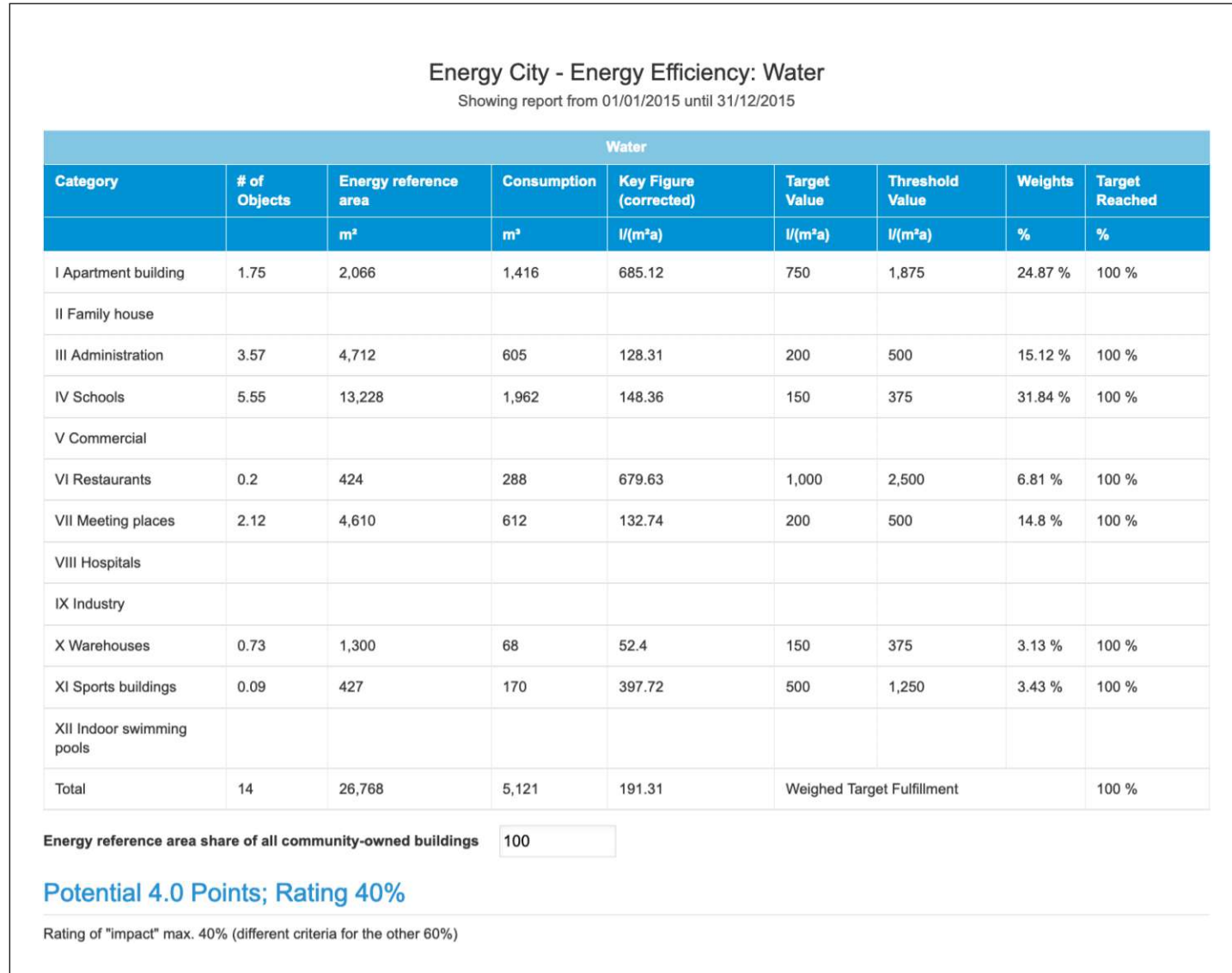


Figure A.18: EnergyCity: Energy efficiency report (water) of a sample community.

A.3 AMAS Case Study Document Index

The following index lists all documents analysed as part of the AMAS case study.

For brevity, document types have been shortened according to this schema:

| | |
|----------------------------------|---------------------------------------|
| AG Meeting Agenda | PP Position Paper |
| CALC Calculation Tool | PRES Presentation / Slides |
| CHECK Checklist | PROT (Meeting) Protocol |
| CONS Consent Form | QA Questionnaire (annotated) |
| DEX Data Excerpt | Q Questionnaire |
| DI Discussion Input | RA Rant |
| DOC Documentation | REQ Requirements Specification |
| GDPR GDPR-Information | RU Ruling |
| GD Guidelines | R Report |
| HB Handbook | SC Screenshots |
| IG Interview Guidelines | SPEC Specification |
| INFO Handout / Info Sheet | SUP Supplement |
| IQR Response to Inquiry | TC Terms and Conditions |
| IQ Inquiry | TD Target Definitions |
| L Letter | T Tender |
| NO Notes | WP Working Paper |
| POL Policy Brief | |

Wherever applicable, multiple document type classifications can be listed.

Document titles were extracted from the documents themselves and are presented *as is*, including somewhat idiosyncratic formatting.

Where authorship could not be determined for individual persons, the issuing organization is listed instead. For documents with multiple authors and differing affiliations, the organization is listed in brackets after the names of the affiliated authors.

Finally, publishing dates refer to either dates mentioned explicitly within the documents, with dates extracted from the file metadata in case no publication date could be found within the document itself.

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|---|---|-------|------------|
| AGB | ALLGEMEINE VERTRAGSBEDINGUNGEN ARBEITSMARKTSERVICE | AMS | TC | 2015-09-01 |
| AUSSCHR_1 | AUSSCHREIBUNGSUNTERLAGE VERGABEVERFAHREN „Entwicklung eines Integrationschance-Prognosemodell zur KundInnensegmentierung im AMS“ | AMS | T | 2015-11-09 |
| AUSSCHR_2 | FORMVORLAGEN UND FORMBLÄTTER | AMS | T | 2015-11-09 |
| BEGL_1 | Beschäftigungsintegration von AMS-Kundinnen/-Kunden Beobachtung, Analyse, Prognose | Jürgen Holl, Günter Kernbeiß, Michael Wagner-Pinter (Synthesis) | SUP | 2019-04-17 |
| BER_1 | Evaluierung des Betreuungsformates für Personen mit multiplen Vermittlungshindernissen (BBEN) - Management Summary | Eva Auer, Petra Tamler (AMS), Friederike Weber (Prospect Research and Solution) | R | 2019-01-01 |
| BER_10 | Arbeitsmarktchancen als Merkmal zur Bildung von KundInnengruppen im AMS | Ernst Haider (AMS) | R | 2018-04-20 |
| BER_11 | Profiling 2008/09 Schlussfolgerungen aus den internationalen Erfahrungen mit Profiling und Targeting und der Pilotversuch 2008/09 | Synthesis | R | 2009-05-01 |
| BER_12 | Bericht des Rechnungshofes: Arbeitsmarktservice (AMS) | Rechnungshof der Republik Österreich | R | 2017-12-22 |
| BER_13 | Arbeitsmarktpolitische Zielvorgaben | BM Beate Hartinger-Klein | R, TD | 2019-02-01 |
| BER_14 | Bericht der Volksanwaltschaft an den Nationalrat und an den Bundesrat: Kontroll der öffentlichen Verwaltung | Volksanwaltschaft | R | 2019-03-01 |
| BER_15 | Vom Arbeitsmarkt vorgegeben: Wie die Verteilung von individuellen Integrationschance das Handlungsfeld des AMS strukturiert | Petra Gegeritsch, Stefanie Gude, Paul Timar, Michael Wagner-Pinter (Synthesis) | R | 2011-11-01 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|--|--|------|------------|
| BER_2 | Profiling-systeme für eine effektive Arbeitsmarktintegration: Neue Ansätze für berufliches Profiling und holistisches Assessment | Jenny Bimrose, Sally Anne Barnes (Warwick Institute of Employment Research) | PP | 2011-05-01 |
| BER_3 | Bericht vom PES - Profiling Workshop: “Profiling for customers at risk from long-term unemployment” | Trude Hausegger (Prospect Research and Solution) | R | 2015-04-01 |
| BER_4 | Evaluierung des Betreuungsformates für Personen mit multiplen Vermittlungshindernissen (BBEN) – Endbericht | Eva Auer, Petra Tamler (AMS), Friederike Weber (Prospect) | R | 2019-01-01 |
| BER_5 | Evaluierung der PPC-Pilotierung: Perspektivencheck Begleitforschung zur Pilotierung im Auftrag des AMS Österreich | Claudia Liebeswar, Mario Taschwer, Andrea Egger-Subotitsch (abif, AMS) | R | 2019-02-01 |
| BER_6 | Arbeitsmarktchancen-Assistenzsystem – AMS-Interne Befragung | Michael Auer, Tobias Krüse (AMS) | R | 2019-11-18 |
| BER_7 | AMS-Chance 2020: Das AMS-Arbeitsmarktchancen-Modell | Ernst Haider, Judit Marte-Huainigg, Marius Wilk (AMS), Jutta Gamper, Günter Kernbeiß, Michael Wagner-Pinter (Synthesis) | R | 2019-12-01 |
| BER_8 | Projektkonzept Neue Segmentierungsstrategie – “Integrationschancendescriptor” (ICD) | Ernst Haider (AMS) | R | 2015-06-22 |
| BER_9 | Interne Kommunikation, strategische Diskussion und Statusberichte zur Umsetzung einer neuen Strategie des AMS in der Arbeitsmarktpolitik | AMS | R | 2015-06-22 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|--|---|------|------------|
| BRIEF_1 | AMS-Arbeitsmarktchancenmodell: Algorithmus zur Chancenbewertung | Herbert Buchinger (AMS) | L | 2015-03-15 |
| BRIEF_2 | Arbeitsmarktchancendescriptor: Algorithmus zur Chancenbewertung am Arbeitsmarkt | Herbert Buchinger (AMS) | L | 2018-12-10 |
| BRIEF_3 | Beantwortung der Datenanfrage des Institut für Technikfolgen-Abschätzung (ITA) der Österreichischen Akademie der Wissenschaften (ÖAW) zum Projekt »Soziotechnische Analyse des AMS Algorithmus« | Ernst Haider (AMS) | IQR | 2020-02-04 |
| DAT_1 | Schnittstellenbeschreibung | Synthesis | DEX | 2019-11-11 |
| DAT_2 | RGS Typen / Modell inkl. Einteilung | Judit Marte-Huainigg (AMS) | DEX | 2019-12-17 |
| DAT_3 | Trefferquoten: Datentabelle | AMS | DEX | 2020-01-16 |
| DOK_1 | Das AMS-Arbeitsmarktchancen-Modell: Dokumentation zur Methode | Jürgen Holl, Günter Kernbeiß, Michael Wagner-Pinter (Synthesis) | DOC | 2018-10-01 |
| DOK_2 | Das Assistenzsystem AMAS: Zweck, Grundlagen, Anwendung | Jutta Gamper, Günter Kernbeiß, Michael Wagner-Pinter (Synthesis) | DOC | 2020-02-01 |
| DOK_3 | Die Informationsbasis des Assistenzsystems AMAS. Statistisch methodische Dokumentation zur Version »Logistische Regression« | Synthesis | DOC | 2020-02-01 |
| DSB | Bescheid der Datenschutzbehörde | Austrian Data Protection Agency (DSB) | RU | 2020-08-19 |
| HAND_1 | Handbuch AMS-REQ-000930 AMAS Arbeitsmarktchancen – Assistenz- System | AMS, IBM | HB | 2019-11-20 |
| INFO_1 | Kundinnen-Information zur Datenschutz-Grundverordnung – DSGVO | AMS | GDPR | N/A |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|---|--|---------|------------|
| NOTES_1 | Die Veränderung der Integrationschancen im Laufe eines Geschäftsfalles: Anomalie oder Informationsquelle? | Jürgen Holl, Günter Kernbeiß, Michael Wagner-Pinter (Synthesis) | DI | 2017-01-01 |
| NOTES_10 | Anfragebeantwortung AMS / GBA | Herbert Buchinger (AMS) | IQ, IQR | 2019-04-24 |
| NOTES_2 | Klausur des Vorstandes mit den LandesgeschäftsführerInnen 25. – 27.2.2015: Zusammenfassund und Vorschläge | AMS | NO | 2015-01-19 |
| NOTES_3 | Unterlage zu den Fragestellungen vom 16.3.2020 | Synthesis | IQR | 2020-03-30 |
| NOTES_4 | »AMS-Algorithmus« am Prüfstand | Michael Wagner-Pinter (Synthesis) | RA | 2020-03-30 |
| NOTES_5 | ITA_unklarheiten_Feedback | Günter Kernbeiß (Synthesis) | IQR | 2020-09-04 |
| NOTES_6 | Personenbezogene Wahrscheinlichkeitsaussagen (»Algorithmen«): Stichworte zur Sozialverträglichkeit | Jürgen Holl, Günter Kernbeiß, Michael Wagner-Pinter (Synthesis) | DOC | 2019-05-09 |
| NOTES_7 | Anfragebeantwortung AMS / Epicenter.works | Marius Wilk (AMS) | IQR | 2019-08-16 |
| NOTES_9 | Verifizierung letzter Unklarheiten | Fabian Fischer (TU Wien) | IQ | 2020-08-31 |
| PARL_1 | Parlamentarische Anfragen (Heinisch-Hosek): "Personalisierte Arbeitsmarktbetreuung" durch das AMS und Algorithmus zur Segmentierung von beim AMS vorgemerkten Arbeitssuchenden | BM Gabriele Heinisch-Hosek | IQ | 2018-11-22 |
| PARL_2 | Parlamentarische Anfrage: Beantwortung | BM Beate Hartinger-Klein | IQR | 2019-01-17 |
| PFLI_1 | Pflichtenheft KundInnensegmentierung | Ernst Haider, Marius Wilk (AMS) | REQ | 2016-06-15 |
| PFLI_2 | Pflichtenheft KundInnensegmentierung | Ernst Haider, Marius Wilk (AMS) | REQ | 2018-07-30 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|--|--|-------|------------|
| POLICY_1 | Profiling tools for early identification of jobseekers who need extra support | Kristine Langenbacher, Theodora Xenogiani (OECD) | POL | 2018-12-01 |
| PPC_CHECK_1 | Checkliste Berufswahl | AMS | CHECK | N/A |
| PPC_FORM_1 | Information zu Zielsetzung und diagnostische Verfahren | AMS | CONS | N/A |
| PPC_FRAGE_1 | Fragebogen für den Einstieg in den „Perspektivencheck“ | AMS | Q | N/A |
| PPC_FRAGE_2 | Fragebogen Arbeitsbewältigung | AMS | Q | N/A |
| PPC_FRAGE_3 | Arbeitsbewältigungsindex Verrechnungsbogen (Anleitung f. SB zum Ausfüllen des Berechnungstools) | AMS | QA | N/A |
| PPC_FRAGE_4 | Ergebnisbericht: Mit neuen Perspektiven zu besseren Arbeitsmarktchancen | AMS | Q | N/A |
| PPC_FRAGE_5 | Ergebnisbericht: Mit neuen Perspektiven zu besseren Arbeitsmarktchancen: Ausfüllhilfe | AMS | QA | N/A |
| PPC_FRAGE_6 | Dokumentationsblatt Handlungsempfehlungen | AMS | Q | N/A |
| PPC_HAND_1 | Handanweisung Perspektivencheck | AMS | HB | N/A |
| PPC_LEIT_1 | Leitfadengestütztes Interview: Termin 1 | AMS | IG | N/A |
| PPC_LEIT_2 | Leitfadengestütztes Interview: Termin 2 | AMS | IG | N/A |
| PPC_RICHT_2 | Unterlage zur Begehrensstellung zur Förderung des Projektes Perspektivencheck zur Abklärung der Arbeitsmarktchancen | AMS | SUP | 2019-10-01 |
| PPC_RICHT_3 | „Perspektivencheck“ (BBEP) Umsetzungsinfo für RGS und LGS | AMS | GD | 2019-10-01 |
| PPC_TOOL_1 | Berechnungstool Arbeitsbewältigung | AMS | CALC | N/A |
| PROT_10_TO | Tagesordnung 5. Workshops zum Datentransfer und zur Modellimplementierung bei IBM | Synthesis | PROT | 2016-04-08 |
| PROT_10_WS | Jour Fixe/Workshop Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-04-08 |
| PROT_11_WS | AMS-REQ-619 KundInnenSegmentierung Projekt Kick Off | AMS | PRES | 2016-04-11 |
| PROT_11_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-04-11 |
| PROT_11_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-04-11 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|--------------------|---|--------------------------------|-------------|------------------|
| PROT_12_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-05-17 |
| PROT_12_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-05-17 |
| PROT_13_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-06-06 |
| PROT_13_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-06-06 |
| PROT_14_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-06-20 |
| PROT_14_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-06-20 |
| PROT_15_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-06-27 |
| PROT_15_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-06-27 |
| PROT_16_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-08-17 |
| PROT_16_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-08-17 |
| PROT_17_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-08-24 |
| PROT_17_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-08-24 |
| PROT_18_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-09-05 |
| PROT_19_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2017-01-25 |
| PROT_1_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-01-11 |
| PROT_1_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-01-11 |
| PROT_20_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2017-02-20 |
| PROT_21_JF | Jour Fixe Protokoll AMS IT: KundInnenSegmentierungsUpdate 2018-Besprechung | AMS | PROT | 2017-06-14 |
| PROT_21_TO | Tagesordnung Jour Fixe - Synthesis "AMS Chance Update" | Synthesis | PROT | 2017-06-14 |
| PROT_22_JF | Jour Fixe Protokoll AMS IT: KundInnenSegmentierungsUpdate 2018-Besprechung | AMS | PROT | 2017-09-07 |
| PROT_22_TO | Tagesordnung Jour Fixe - Synthesis "AMS Chance Update" | Synthesis | PROT | 2017-09-07 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|---|-------------------------|------|------------|
| PROT_2_TO | Tagesordnung 1. Workshops zum Datentransfer und zur Modellimplementierung bei IBM | Synthesis | PROT | 2016-01-22 |
| PROT_2_WS | Protokoll AMS IT AMS-REQ-536 KundInnensegmentierung Workshop IBM-Synthesis | AMS | PROT | 2016-01-22 |
| PROT_3_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-01-25 |
| PROT_3_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-01-25 |
| PROT_4_TO | Tagesordnung 2. Workshops zum Datentransfer und zur Modellimplementierung bei IBM | Synthesis | PROT | 2016-02-12 |
| PROT_4_WS | Notizen Workshop | Synthesis | PROT | 2016-02-12 |
| PROT_5_TO | Tagesordnung Jour Fixe "Integrationschancen" - Synthesis | AMS | PROT | 2016-02-15 |
| PROT_6_TO | Tagesordnung 3. Workshops zum Datentransfer und zur Modellimplementierung bei IBM | Synthesis | PROT | 2016-02-26 |
| PROT_6_WS | Jour Fixe/Workshop Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-02-26 |
| PROT_7_JF | Jour Fixe Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-02-29 |
| PROT_7_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-02-29 |
| PROT_8_TO | Tagesordnung 4. Workshops zum Datentransfer und zur Modellimplementierung bei IBM | AMS | PROT | 2016-03-11 |
| PROT_8_WS | Jour Fixe/Workshop Protokoll AMS IT: KundInnensegmentierung Projektstatus Meeting | AMS | PROT | 2016-03-11 |
| PROT_9_TO | Tagesordnung Jour Fixe - Synthesis | Synthesis | PROT | 2016-03-14 |
| PROT_ORG_1 | KundInnensegmentierung und neue Betreuungsstrategie im AMS: Stand der Arbeitsgruppen | Ernst Haider (AMS) | PRES | 2015-11-03 |
| PROT_ORG_10 | Protokoll Verwaltungsratsitzung: TOP 10 Personalisierte Betreuung von Arbeitslosen aufgrund von Arbeitsmarktchancen als Merkmal zur Bildung von KundInnengruppen im AMS | AMS | PROT | 2018-05-08 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|--|-------------------------|------|------------|
| PROT_ORG_11 | Strategieausschuss: 5. Funktionsperiode, 21. Sitzung, Protokoll, Teil 1 | AMS | PROT | 2019-11-06 |
| PROT_ORG_12 | Tagesordnung Präsidium 3.1..2015 - Screenshot | AMS | AG | 2015-11-03 |
| PROT_ORG_13 | Präsidium des Verwaltungsrates - 4. Funktionsperiode - Protokoll 14. Sitzung | AMS | PROT | 2015-11-03 |
| PROT_ORG_2 | Strategieausschuss: AM-Chancen als Merkmal zur Bildung von KundInnengruppen | AMS | PROT | 2018-07-03 |
| PROT_ORG_3 | Protokoll Verwaltungsratsitzung: TOP 15.1. Personalisierte Arbeitsmarktbetreuung Assistenz-System | AMS | PROT | 2018-12-04 |
| PROT_ORG_4 | Protokoll Verwaltungsratsitzung: TOP 9 Grundsatzbeschluss zur Umsetzung der PAMAS-Strategie | AMS | PROT | 2019-06-25 |
| PROT_ORG_5 | Protokoll Verwaltungsratsitzung: TOP 5 Grundsatzbeschluss zur Umsetzung der PAMAS-Strategie | AMS | PROT | 2019-09-17 |
| PROT_ORG_6 | Protokoll Verwaltungsratsitzung: TOP 4 Entwicklung eines Prognosemodells zur KundInnensegmentierung (Syntheseforschung, Univ.-Prof. Dr. Wagner-Pinter) | AMS | PROT | 2016-03-08 |
| PROT_ORG_7 | Protokoll Verwaltungsratsitzung: TOP 14 BBE Neu für Personen mit multiplen Vermittlungshindernissen: Konzeption und Erfolgsmessung | AMS | PROT | 2016-11-08 |
| PROT_ORG_8 | Protokoll Verwaltungsratsitzung: TOP 9 Strategie im AMS - BBENeu - Modellvergleiche aus den Bundesländern | AMS | PROT | 2016-12-14 |
| PROT_ORG_9 | Protokoll Verwaltungsratsitzung: TOP 6 Umsetzung der Pilotprojekte für Personen mit multiplen Vermittlungshindernissen | AMS | PROT | 2017-03-28 |
| PRÄS_1 | AMS: strategische Herausforderungen - Überleben für weitere 20 Jahre | Herbert Buchinger (AMS) | PRES | 2015-02-01 |
| PRÄS_2 | Personalisierte Arbeitsmarktbetreuung - KundInnengruppen im AMS | Herbert Buchinger (AMS) | PRES | 2018-06-01 |
| PRÄS_3 | Arbeitsmarktchancen als Merkmal zur Bildung von KundInnengruppen im Arbeitsmarktservice | Ernst Haider (AMS) | PRES | 2018-06-01 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|-------------|---|---|------|------------|
| PRÄS_4 | KundInnengruppen nach Arbeitsmarktchancen: kumulative Erweiterung/Einschränkung | AMS | PRES | 2019-12-17 |
| PRÄS_5 | AMS-Profilung im Pilotbetrieb | Elisabeth Oehry (BGS/SfA), Sabine Putz (BGS/ABI), Daniel Kamleitner (Synthesis) | PRES | 2008-01-01 |
| PRÄS_6 | Profilungtool: Umsetzung im AMS Wien | Thomas Haider / LGS Wien (AMS) | PRES | 2012-08-27 |
| PRÄS_7 | KundInnensegmentierung OÖ | AMS/Synthesis | PRES | 2015-07-09 |
| RICHT_1 | Bundesrichtlinie Kernprozess Arbeitskräfte unterstützen | Herbert Buchinger, Johannes Kopf (AMS) | GD | 2019-12-20 |
| RICHT_2 | Arbeitsmarktchancen – Assistenz-System (AMAS): KundInnenfragen beantworten | AMS | GD | 2019-12-16 |
| SCHU_1 | Programm AMAS TTT Veranstaltung (Train the Trainers) | AMS | AG | 2020-03-16 |
| SCHU_2 | AMAS ganz kurz | Bettina Huber (AMS) | INFO | 2019-12-09 |
| SCHU_3 | HANDOUT - SfA - AMAS – Arbeitsmarktchancen – Assistenz-System | Alexander Peinhaupt (AMS) | INFO | 2020-01-27 |
| SCHU_5 | Eine kompakte Information über das Arbeitsmarktchancen Assistenz-System (AMAS) für externe Trainer und Trainerinnen des AMS | Bettina Huber, Judit Marte-Huainigg (AMS) | INFO | 2020-03-09 |
| SCHU_6 | Übersicht über die Kontaktintervalle | Gerlinde Wieser-Böhm (AMS) | GD | 2020-01-01 |
| SCHU_7 | Arbeitsmarktchancen Assistenz-System - Informationsworkshop | Karin Ostermann, Bettinga Huber, Ernst Haider, Karin König, Judit Marte-Huainigg (AMS) | PRES | 2020-02-14 |
| SCR_1 | Screenshots AMS interne Community - Info f. PAMAS | AMS | SC | N/A |
| SCR_2 | Screenshots AMS eLearning Plattform - PAMAS | AMS | SC | N/A |
| SPEZ_1 | Variablen für das Basismodell | Günter Kernbeiß (Synthesis) | SPEC | 2019-06-03 |

Continued on next page

Continued from previous page

| Document ID | Title | Author(s) / Affiliation | Type | Published |
|--------------------|--|---|-------------|------------------|
| SPEZ_2 | Textbausteine zur Erläuterung der Arbeitsmarktchancen Modell 2020 | AMS | SPEC | 2019-12-17 |
| SYN_1 | Assessment des informationstechnischen Systems PAMAS des Arbeitsmarktservice Österreich - Worum geht es? | Synthesis | DOC | N/A |
| TO_1 | Tagesordnung Wissensvermittlung Kommunikation von AMAS Kommunikation – Grundlage 6 Phasen Modell | AMS | AG | N/A |
| WP_1 | Statistical Profiling in Public Employment Services: An international comparison | Sam Desiere, Kristine Langenbucher, Ludo Struyven (OECD) | WP | 2019-02-13 |

A.4 AMAS Case Study Explanation Texts

The following table lists the text fragments and conditions for the additional segment information of the **AMAS** system for either the *high* or *low* chances segment. The table is based on multiple tables in **BER_7** and **SPEZ_2**. The German texts have been translated by the author.

| Fragment ID | Text shown | Condition I | Condition II |
|---|--|--|--|
| "Encouraging" (applied for high chances) | | | |
| High continuity in employment | <i>"You were employed over a long period of time in recent years."</i> | Days of gainful employment within 4 years \geq 75 % | Duration of current unemployment: max. milestone 3 |
| Experience with supportive labour market policy measures | <i>"You showed willingness to take part in supportive measures provided by the AMS."</i> | Education: None for age <50 or grade school+ for age >50 | Measures claimed: 2+ |
| Prior experience with job applications | <i>"You started at a new job repeatedly, and subsequently have experience in employment in the labour market."</i> | Cases within 4 one year intervals: 2+ | Cases with duration longer than 180 days: 0 |
| Established job-specific vocational training | <i>"You have vocational training with a formal graduation of at least the level of a skilled worker."</i> | Education: Apprenticeship, vocational school or high-, secondary school or university degree | |

Continued on next page

Continued from previous page

| Fragment ID | Text shown | Condition I | Condition II |
|---|--|--|--|
| "Particularly challenging" (applied for low chances) | | | |
| Deficits in competency | <i>"You only completed your mandatory education and/or have limited German language skills, which makes your job search harder."</i> | Education: Mandatory school | Citizenship: Non-EU |
| No vocational training | <i>"You have no additional vocational training beyond mandatory school."</i> | Education: Mandatory school | Citizenship: Austria or EU |
| Significant gaps in employment | <i>"You were only employed over limited periods of time during the last few years."</i> | Days of gainful employment within 4 years: <75 % | Duration of current unemployment: min. milestone 12 |
| Low utilization of supportive measures | <i>"You have not taken part in any supportive measures offered by the AMS in recent years."</i> | Measures claimed: 0 | Cases with duration longer than 180 days \geq 1+ |
| Obligations of care | <i>"You have obligations of care for members of your immediate family."</i> | Obligations of care: Yes | Age <50 |
| Health impairments | <i>"You have health impairments that make your job search harder."</i> | Health Impairments: Yes | Education: Apprenticeship, vocational school or higher if age \leq 50; None if age >50 |
| Advanced age | <i>"For people of an advanced age like you it is more difficult to find new employment."</i> | Age >50 | Education: Apprenticeship, vocational school or higher |

List of Figures

| | |
|---|-----|
| 1.1 CAS related disciplines | 10 |
| 1.2 Number of CAS-related publications | 11 |
| 1.3 Dissertation Structural Flow | 19 |
| 2.1 Big Data | 34 |
| 2.2 Transparency Issues | 41 |
| 2.3 XAI approaches | 45 |
| 2.4 Computer Says No | 50 |
| 2.5 Public Accountability Taxonomy | 51 |
| 2.6 Accountability Cube | 66 |
| 2.7 Bandura’s Triadic Reciprocal Causation Model | 78 |
| 3.1 EnerCoach Research Project Timeline | 101 |
| 3.2 AMAS Research Project Timeline | 116 |
| 4.1 “Homepage” of the original EnerCoach Tool | 131 |
| 4.2 EnerCoach Code Snippet | 132 |
| 4.3 EnerCoach stakeholder groups and their interactions. | 139 |
| 4.4 EnerCoach simplified data model | 146 |
| 4.5 EnerCoach Mix Period Example Form | 149 |
| 4.6 EnerCoach user counts vs. communities. | 160 |
| 4.7 Screen capture of early EnerCoach support request | 164 |
| 4.8 EnerCoach report missing data table | 166 |
| 4.9 Software Engineer Temptations | 171 |
| 4.10 EnerCoach Entity Cards | 173 |
| 4.11 Impressions of the PD workshop visualization exercise. | 174 |
| 4.12 EnerCoach Certificate Report Visualization (Initial Version) | 178 |
| 4.13 EC Stakeholder Knowledge Requirements | 179 |
| 4.14 EnerCoach Mock-Up Examples | 182 |
| 4.15 EnerCoach Report Visualizations (Final Version) | 184 |
| 4.16 EnerCoach Key Figure Report Adaptations | 185 |
| 5.1 AMAS Coefficients | 193 |
| 5.2 Profiling Methodologies | 198 |
| 5.3 AMAS Profiling Process Flowchart | 210 |
| | 351 |

| | | |
|------|--|-----|
| 5.4 | AMAS Personal Data and Constellations | 213 |
| 5.5 | AMAS IC Score Sample Form | 215 |
| 5.6 | AMAS IC Score Explanations Form | 219 |
| 5.7 | AMAS Documentation Table - All Populations by Gender and Age | 223 |
| 5.8 | AMAS Documentation Table - Complete Data Populations by Gender and Age | 224 |
| 5.9 | AMAS Documentation Table - Incomplete Data / "Migration Background" Population by Gender and Age | 224 |
| 6.1 | Procedural model of the accountability process for the A ³ framework | 253 |
| 6.2 | XKCD: "Exploits of a Mom" | 257 |
| 6.3 | A ³ framework Guiding Questions | 258 |
| 6.4 | Bødker and Klokmoose's Human-Artifact Model | 281 |
| 6.5 | AIA/Audit Methods Taxonomy | 285 |
| A.1 | Energy certificate report showing the sustainability performance of a sample community | 320 |
| A.2 | Poster report showing the report plaque of a sample building | 321 |
| A.3 | Key Figure Reports (Electricity, Heat and Water) of a sample community | 322 |
| A.4 | Key Figure Report (Electricity), filtered by building category 'schools' and showing target and threshold values for key figures of a sample community | 323 |
| A.5 | Evolution of energy consumption report of a sample community | 324 |
| A.6 | Evolution of greenhouse gas emissions report of a sample community | 325 |
| A.7 | Evolution of energy costs report of a sample community | 326 |
| A.8 | Energy carrier shares report of a sample community | 327 |
| A.9 | Energy indicator vs. reference area report of a sample community | 328 |
| A.10 | Key figure comparison (electricity) report of a sample community | 329 |
| A.11 | Key figure comparison (heat) report of a sample community | 330 |
| A.12 | Key figure comparison (water) report of a sample community | 331 |
| A.13 | EnergyCity: Renewable energy (electricity) report of a sample community | 332 |
| A.14 | EnergyCity: Renewable energy (heat) report of a sample community | 333 |
| A.15 | EnergyCity: Greenhouse gas intensity report of a sample community | 334 |
| A.16 | EnergyCity: Energy efficiency report (electricity) of a sample community | 335 |
| A.17 | EnergyCity: Energy efficiency report (heat) of a sample community | 336 |
| A.18 | EnergyCity: Energy efficiency report (water) of a sample community | 337 |

List of Tables

| | |
|--|-----|
| 3.1 EnerCoach qualitative codes overview | 108 |
| 3.2 EnerCoach codebase number of files and lines | 109 |
| 3.3 AMAS document categories overview | 118 |
| 4.1 EnerCoach community types currently in use in Switzerland. | 147 |
| 4.2 EnerCoach zone categories in use. | 150 |
| 5.1 AMAS variables and values overview | 207 |
| 5.2 AMAS Classifications and Group Labels | 217 |

Glossary

Eidgenössischer Gebäudeidentifikator The EGID is a unique identifier of a building in the Swiss Federal Register of Buildings and Housing, sometimes also referred to as Federal Building Identifier. Primarily used to uniquely assign each registered inhabitant in Switzerland to a building (and possible apartment unit in the building via its sister identifier, the EWID or Eidgenössische Wohnungsidentifikator), the EGID also serves as a unique identifier for a given building in the EnerCoach tool.

148

Kilowatt-hour Kilowatt-hours [390] are a unit of energy over time. One kilowatt-hour is equal to a kilowatt of power sustained over one hour. kWh are a non-standard unit, as they are not directly represented in the International System of Units (ISU); the corresponding IS unit is the Megajoule, with $1kWH = 3.6MJ$. 147

Megajoule Joule [391] is a derived unit of energy in the ISU. It can be derived from the units of *force* and *distance* as a force of one Newton displacing a given mass by one meter, or $J = N * m$ with J (Joule), N (Newton) and m (Meter) (in a vacuum and excluding any other impact factors). 147

Acronyms

- AAA** Algorithmic Accountability Act. [60](#)
- ACA** Austrian Court of Audit. [113](#)
- ACM** Association for Computing Machinery. [62](#), [63](#)
- ADEME** Agence de la transition écologique. [137](#)
- ADM** Automated Decision-Making. [ix](#), [xiii](#), [3](#), [40](#), [56](#), [60](#), [68](#), [84](#), [98](#), [191](#), [195](#)–[197](#)
- ADS** Automated Decision Support. [40](#), [98](#), [191](#), [195](#), [197](#), [199](#), [238](#)
- AEA** Austrian Energy Agency. [137](#)
- AFSI** Austrian Federation of Social Insurances. [206](#), [207](#)
- AI** Artificial Intelligence. [ix](#), [xiii](#), [3](#), [11](#), [84](#), [133](#)
- AI/ML** Artificial Intelligence/Machine Learning. [11](#), [17](#), [75](#), [92](#), [244](#), [247](#)–[250](#), [288](#), [298](#), [311](#), [313](#)
- AIA** Algorithmic Impact Assessment. [21](#), [60](#), [61](#), [68](#), [195](#), [241](#), [284](#), [287](#), [288](#), [310](#), [311](#), [313](#), [314](#)
- AKOÖ** Austrian Chamber of Labour for Upper Austria. [113](#), [115](#), [125](#), [192](#)
- AMAS** Arbeitsmarkt-Assistenz-System. [x](#), [xiii](#)–[xv](#), [6](#), [8](#), [20](#), [32](#), [37](#), [49](#), [59](#), [62](#), [65](#), [97](#), [98](#), [113](#)–[115](#), [119](#), [120](#), [122](#), [123](#), [125](#), [126](#), [187](#), [191](#), [192](#), [194](#), [199](#)–[202](#), [204](#)–[206](#), [209](#), [210](#), [214](#), [216](#)–[220](#), [222](#)–[226](#), [228](#), [231](#), [233](#)–[238](#), [241](#)–[250](#), [254](#), [260](#), [266](#), [268](#)–[276](#), [279](#), [287](#)–[289](#), [294](#), [297](#), [298](#), [300](#), [303](#)–[306](#), [309](#), [311](#), [349](#)
- AMS** Public Employment Service Austria. [xiii](#), [xiv](#), [6](#), [32](#), [52](#), [62](#), [113](#)–[116](#), [119](#), [120](#), [125](#), [191](#), [192](#), [196](#), [199](#)–[203](#), [238](#), [248](#), [274](#), [303](#), [309](#)
- ANT** Actor-Network Theory. [4](#), [7](#), [24](#), [28](#), [29](#), [31](#), [32](#), [76](#), [77](#), [81](#), [85](#), [89](#), [189](#), [293](#), [296](#), [301](#)
- API** Application Programming Interface. [286](#)

- A³ framework** Algorithmic Accountability Agency Framework. [x](#), [xi](#), [xiv-xvi](#), [6](#), [7](#), [18](#), [20](#), [21](#), [76](#), [89](#), [95](#), [121](#), [124](#), [188](#), [191](#), [236](#), [239](#), [241](#), [246](#), [247](#), [250-254](#), [258-260](#), [266](#), [273-277](#), [279](#), [280](#), [282-284](#), [286-292](#), [294](#), [295](#), [301](#), [303](#), [304](#), [306](#), [307](#), [311](#), [312](#), [314](#), [352](#)
- BAM** Labour Market Chance as assessed by Caseworker. [214-216](#), [228](#), [246](#), [271](#)
- BBE** External Counselling and Support Institutions. [217](#), [218](#)
- BBEP** Evaluation of Prospects Measure. [208](#), [217](#), [218](#)
- BEMS** Building Energy Management System. [135](#)
- C!S** Centre for Informatics and Society. [113](#), [115](#), [133](#), [191](#)
- CAM** Computer-Supported Labour Market Chance. [214-216](#), [228](#), [246](#), [271](#)
- CAS** Critical Algorithm Studies. [ix](#), [xi](#), [xiii](#), [4](#), [8-15](#), [17](#), [20](#), [23](#), [40](#), [50](#), [54](#), [55](#), [75](#), [91](#), [95](#), [97](#), [99](#), [100](#), [121](#), [124](#), [126](#), [127](#), [130](#), [132](#), [158](#), [171](#), [187](#), [188](#), [194](#), [195](#), [242](#), [244](#), [247](#), [250](#), [259](#), [260](#), [284](#), [288](#), [292](#), [295](#), [297](#), [306](#), [308](#), [312](#), [314](#)
- CDS** Critical Data Studies. [12](#), [55](#)
- CDSS** Clinical Decision Support Systems. [38](#)
- CEO** Chief Executive Officer. [196](#), [202](#)
- CFAA** Computer Fraud and Abuse Act of 1986. [286](#)
- CNN** Convolutional Neural Network. [44-46](#)
- COMPAS** Correctional Offender Management Profiling for Alternative Sanctions. [26](#), [35](#), [36](#), [40](#), [303](#)
- CRUD** Create, Read, Update, Delete. [153](#)
- CSCW** Computer-Supported Cooperative Work. [ix](#), [xiii](#), [3](#), [8](#), [10](#), [15](#), [27](#), [138](#), [186](#), [238](#), [259](#), [292](#), [312](#)
- DSB** Austrian Data Protection Agency. [32](#), [201](#), [230](#), [266](#), [267](#), [272](#)
- EBA** Ethics-based Audit. [310](#)
- EEA** European Energy Awards. [130](#), [137](#), [138](#), [143](#), [156](#), [157](#), [161](#)
- EMS** Energy Management System. [135](#)
- FAccT** Fairness, Accountability and Transparency in Algorithmic Systems. [10-12](#), [35](#)

FOIA Freedom of Information Act. [65](#)

GBA Ombud for Equal Treatment. [113](#), [229](#), [246](#)

GDPR General Data Protection Regulation. [55](#), [59](#), [72-74](#), [287](#)

HCI Human-Computer Interaction. [ix](#), [xiii](#), [3](#), [8](#), [10](#), [15](#), [21](#), [75](#), [101](#), [134](#), [138](#), [186](#), [187](#), [238](#), [241](#), [259](#), [280](#), [283](#), [292](#), [301](#), [302](#), [312](#)

IC Integration Chance. [211-214](#), [219](#), [224](#), [228](#), [229](#), [231](#), [232](#), [269](#)

IoT Internet of Things. [133](#), [135](#)

IPCC Intergovernmental Panel on Climate Change. [133](#)

ISU International System of Units. [355](#)

ITA Institute of Technology Assessment. [113](#), [115](#), [191](#), [192](#)

KPI Key Performance Indicator. [197](#)

ML Machine Learning. [11](#)

NLP Natural Language Processing. [46](#)

NPM Node Package Manager. [61](#)

OBA Online Behavioural Advertising. [87](#)

OECD Organisation for Economic Co-operation and Development. [119](#), [197](#), [198](#)

PII Personal Identifiable Information. [32](#)

SCT Social Cognitive Theory. [5](#), [77](#), [82](#), [83](#), [90](#), [290](#), [301](#)

SDG Sustainable Development Goals. [133](#)

SIA Swiss Society of Engineers and Architects. [149](#), [152](#), [157](#), [173](#), [176](#)

ST Structuration Theory. [77](#)

STS Science and Technology Studies. [ix](#), [xiii](#), [3](#), [9](#), [10](#), [12](#), [13](#), [15](#), [23](#), [50](#), [75](#), [120](#), [288](#), [291](#), [306](#), [307](#), [312](#), [313](#)

TPS Thermal Production System. [151](#), [155](#), [159](#), [162](#)

WIFO Austrian Institute of Economic Research. [203](#)

XAI Explainable AI. [40](#), [44-46](#), [86](#), [249](#), [250](#), [298](#)

ÖAW Austrian Academy of Sciences. [113](#), [191](#)

Bibliography

- [1] F. Cech, “Tackling Algorithmic Transparency in Communal Energy Accounting through Participatory Design,” in *ACM Communities and Technologies Conference 2021*, ser. C&T ’21: Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech. ACM, 2021, pp. 258–268. [Online]. Available: <https://doi.org/10.1145/3461564.3461577>
- [2] —, “Beyond Transparency,” in *Companion of the 2020 ACM International Conference on Supporting Group Work*, ser. GROUP ’20: The 2020 ACM International Conference on Supporting Group Work, vol. 41. ACM, January 2020, pp. 11 – 14. [Online]. Available: <https://doi.org/10.1145/3323994.3371015>
- [3] D. Allhutter, F. Cech, F. Fischer, G. Grill, and A. Mager, “Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective,” *Frontiers in Big Data*, vol. 3, pp. 1 – 28, February 2020. [Online]. Available: <https://doi.org/10.3389/fdata.2020.00005>
- [4] —, “Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS),” ITA/ÖAW, Tech. Rep. 2020-02, 2020. [Online]. Available: <https://doi.org/10.1553/ita-pb-2020-02>
- [5] B. Wagner, P. Lopez, F. Cech, G. Grill, and M.-T. Sekwenz, “Der AMS-algorithmus.” *Zeitschrift für kritik - recht - gesellschaft*, no. 2, p. 191, 2020. [Online]. Available: <https://doi.org/10.33196/juridikum202002019101>
- [6] F. Cech, “The Agency of the Forum: Mechanisms for Algorithmic Accountability through the Lens of Agency,” *Journal of Responsible Technology*, vol. 7-8, October 2021. [Online]. Available: <https://doi.org/10.1016/j.jrt.2021.100015>
- [7] —, “Exploring emerging topics in social informatics: An online real-time tool for keyword co-occurrence analysis,” in *Social Informatics. SocInfo 2017. Lecture Notes in Computer Science*, vol. 10540 LNCS, January 2017, pp. 527 – 536. [Online]. Available: http://link.springer.com/10.1007/978-3-319-67256-4_42
- [8] F. Cech and M. Wagner, “Erollin’ on green: A case study on Eco-Feedback Tools for eMobility,” in *Proceedings of the 9th International Conference on Communities & Technologies - Transforming Communities*, ser. Communities and Technologies 2019, June 2019, pp. 121 – 125. [Online]. Available: <https://doi.org/10.1145/3328320.3328402>
- [9] S. Human and F. Cech, “A Human-Centric Perspective on Digital Consenting: The Case of GAFAM,” in *Human Centred Intelligent Systems*, A. Zimmermann, R. J. Howlett, and L. C. Jain, Eds. Singapore: Springer Singapore, 2021, pp. 139–159. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-5784-2_12
- [10] H. Tellioglu, M. Habiger, and F. Cech, “Infrastructures for sense making,” in *InfraHealth 2017: Proceedings of the 6th International Workshop on Infrastructures for Healthcare*. European Society for Socially Embedded Technologies (EUSSET), June 2017, pp. 1–4. [Online]. Available: <https://dl.eusset.eu/handle/20.500.12015/2908>
- [11] C. Kelty, “Qualitative research in the age of the algorithm: New challenges in cultural anthropology,” in *Lecture at the Research Libraries Group 2003 Annual Meeting: Rethinking the Humanities in a Global Age*, Boston Public Library, Boston, MA, May, vol. 5, 2003.
- [12] M. Ziewitz, “Governing algorithms,” *Science, Technology, & Human Values*, vol. 41, no. 1, pp. 3–16, September 2015. [Online]. Available: <https://doi.org/10.1177/0162243915608948>

- [13] L. Andrews, “Public administration, public leadership and the construction of public value in the age of the algorithm and ‘big data’,” *Public Administration*, vol. 97, no. 2, pp. 296–310, August 2018. [Online]. Available: <https://doi.org/10.1111/padm.12534>
- [14] E. Finn, “Coda: The algorithmic imagination,” in *What Algorithms Want*, ser. What Algorithms Want. The MIT Press, March 2017. [Online]. Available: <https://doi.org/10.7551/mitpress/9780262035927.003.0007>
- [15] L. Rainie and J. Anderson, “Code-dependent: Pros and cons of the algorithm age,” Pew Research Center, Tech. Rep., 2017. [Online]. Available: <https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/>
- [16] B. Franklin and A. Smyth, *The Writings of Benjamin Franklin*, ser. The Writings of Benjamin Franklin. Macmillan, 1856, no. v. 10. [Online]. Available: <https://books.google.at/books?id=FjdLAAAAYAAJ>
- [17] R. Berman, “This A.I. can predict how long you’ll live—and it’s free,” *Big Think*, April 2022. [Online]. Available: <https://bigthink.com/surprising-science/an-ai-algorithm-predicts-your-expiration-date> (Accessed: 2022-11-06)
- [18] Y. Guo, F. Farooq, D. d. Roux, B. Perez, A. Moreno, M. d. P. Villamil, and C. Figueroa, “Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach,” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 215–222, 2018. [Online]. Available: <https://doi.org/10.1145/3219819.3219878>
- [19] C. Groves, H. v. Lente, C. Selin, and K. Konrad, *Performing and Governing the Future in Science and Technology*. MIT Press, September 2018.
- [20] J. Sadowski and R. Bendor, “Selling smartness: Corporate narratives and the smart city as a sociotechnical imaginary,” *Science, Technology, & Human Values*, vol. 44, no. 3, pp. 540–563, October 2018. [Online]. Available: <https://doi.org/10.1177/0162243918806061>
- [21] N. Seaver, “Algorithms as culture: Some tactics for the ethnography of algorithmic systems,” *Big Data & Society*, vol. 4, no. 2, November 2017. [Online]. Available: <https://doi.org/10.1177/2053951717738104>
- [22] M. Bovens, “Analysing and assessing accountability: A conceptual framework,” *European Law Journal*, vol. 13, no. 4, pp. 447–468, July 2007. [Online]. Available: <https://doi.org/10.1111/j.1468-0386.2007.00378.x>
- [23] M. Wieringa, “What to account for when accounting for algorithms,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, January 2020, pp. 1 – 18. [Online]. Available: <https://doi.org/10.1145/3351095.3372833>
- [24] A. Bandura, “Social cognitive theory: An agentic perspective,” *Annual Review of Psychology*, vol. 52, no. 1, pp. 1–26, February 2001. [Online]. Available: <https://doi.org/10.1146/annurev.psych.52.1.1>
- [25] V. Jupp, *The SAGE Dictionary of Social Research Methods*. SAGE Publications, Ltd, 2006. [Online]. Available: <https://doi.org/10.4135/9780857020116>
- [26] D. Reed-Danahay, *Auto/Ethnography: Rewriting the Self and the Social*. Berg Publishers, 1997.
- [27] C. Geertz, *The Interpretation of Cultures*. Basic Books, Inc., 1973. [Online]. Available: <https://chairflogicphiloscult.files.wordpress.com/2013/02/clifford-geertz-the-interpretation-of-cultures.pdf>
- [28] J. Kemper and D. Kolkman, “Transparent to whom? No algorithmic accountability without a critical audience,” *Information, Communication & Society*, vol. 22, no. 14, pp. 2081–2096, June 2018. [Online]. Available: <https://doi.org/10.1080/1369118x.2018.1477967>
- [29] (2016) Critical algorithm studies: a reading list – social media collective. [Online]. Available: <https://socialmediacollective.org/reading-lists/critical-algorithm-studies/> (Accessed: 2018-05-30)
- [30] H. W. J. Rittel and M. M. Webber, “Dilemmas in a general theory of planning,” *Policy Sciences*, vol. 4, no. 2, pp. 155–169, June 1973. [Online]. Available: <https://doi.org/10.1007/bf01405730>
- [31] R. Kowalski, “Algorithm = logic + control,” *Communications of the ACM*, vol. 22, no. 7, pp. 424–436, Jul. 1979. [Online]. Available: <https://doi.org/10.1145/359131.359136>

- [32] H. Nissenbaum, “How computer systems embody values,” *Computer*, vol. 34, no. 3, pp. 120–119, March 2001. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/910905/>
- [33] K. Hamilton, K. Karahalios, C. Sandvig, and M. Eslami, “A path to understanding the effects of algorithm awareness,” in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, ser. the extended abstracts of the 32nd annual ACM conference. New York, New York, USA: ACM, April 2014, pp. 631 – 642. [Online]. Available: <https://doi.org/10.1145/2559206.2578883>
- [34] E. Rader and R. Gray, “Understanding user beliefs about algorithmic curation in the facebook news feed,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, April 2015, pp. 173–182. [Online]. Available: <https://doi.org/10.1145/2702123.2702174> (Accessed: 2016-07-19)
- [35] (2021, July) Critical Algorithm Studies in Publications - Dimensions. [Online]. Available: https://app.dimensions.ai/discover/publication?search_mode=content%5C&search_text=%22Critical%20%20Algorithm%20Studies%22%5C&search_type=kws%5C&search_field=full_search%5C (Accessed: 2021-07-30)
- [36] (2021, June) ACM FAcCT - 2021 CFP. [Online]. Available: <https://facctconference.org/2021/cfp.html> (Accessed: 2021-07-23)
- [37] A. Mehrotra and L. E. Celis, “Mitigating bias in set selection with noisy protected attributes,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021, pp. 237–248. [Online]. Available: <https://doi.org/10.1145/3442188.3445887>
- [38] M. Ghadiri, S. Samadi, and S. Vempala, “Socially fair k-means clustering,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021, pp. 438–448. [Online]. Available: <https://doi.org/10.1145/3442188.3445906>
- [39] I. D. Raji, M. K. Scheuerman, and R. Amironesei, “You can't sit with us,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021, pp. 515–525. [Online]. Available: <https://doi.org/10.1145/3442188.3445914>
- [40] C. O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books Limited, 2016.
- [41] C. O’Neil. (2017, November) Opinion | The Ivory Tower Can’t Keep Ignoring Tech. [Online]. Available: <https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html> (Accessed: 2022-06-17)
- [42] D. Moats and N. Seaver, “You Social Scientists Love Mind Games: Experimenting in the divide between data science and critical algorithm studies,” *Big Data & Society*, vol. 6, no. 1, January 2019. [Online]. Available: <https://doi.org/10.1177/2053951719833404>
- [43] T. F. Gieryn, *Cultural Boundaries of Science: Credibility on the Line*. University of Chicago Press, 1999.
- [44] J. Bohman, “Critical Theory,” in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. [Online]. Available: <https://plato.stanford.edu/archives/spr2021/entries/critical-theory/>
- [45] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. (2016, May) Machine Bias. [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (Accessed: 2022-06-17)
- [46] K. Kubler, “State of urgency: Surveillance, power, and algorithms in france’s state of emergency,” *Big Data & Society*, vol. 4, no. 2, November 2017. [Online]. Available: <https://doi.org/10.1177/2053951717736338>
- [47] T. Zarsky, “The trouble with algorithmic decisions,” *Science, Technology, & Human Values*, vol. 41, no. 1, pp. 118–132, October 2015. [Online]. Available: <https://doi.org/10.1177/0162243915605575>
- [48] J. Macdonald. (2015) Algorithmic Studies: A (Brief) Critical Survey. [Online]. Available: <https://algorithmicstudies.uchri.org/literature-survey/> (Accessed: 2021-07-16)

- [49] L. Manovich, *The language of new media*. Cambridge, Massachusetts: The MIT Press, 2001. [Online]. Available: <https://books.google.at/books?id=G-4EyYR-QgMC> (Accessed: 2018-05-30)
- [50] D. Beer, *Popular Culture and New Media*. Palgrave Macmillan UK, 2013. [Online]. Available: <https://doi.org/10.1057/9781137270061>
- [51] J. Danaher, M. J. Hogan, C. Noone, R. Kennedy, A. Behan, A. D. Paor, H. Felzmann, M. Haklay, S.-M. Khoo, J. Morison, M. H. Murphy, N. O'Brolchain, B. Schafer, and K. Shankar, "Algorithmic governance: Developing a research agenda through the power of collective intelligence," *Big Data & Society*, vol. 4, no. 2, September 2017. [Online]. Available: <https://doi.org/10.1177/2053951717726554>
- [52] C. Liu and R. Graham, "Making sense of algorithms: Relational perception of contact tracing and risk assessment during COVID-19," *Big Data & Society*, vol. 8, no. 1, January 2021. [Online]. Available: <https://doi.org/10.1177/2053951721995218>
- [53] T. Schwartz, G. Stevens, T. Jakobi, S. Deneff, L. Ramirez, V. Wulf, and D. Randall, "What people do with consumption feedback: A long-term living lab study of a home energy management system," *Interacting with Computers*, vol. 27, no. 6, pp. 551–576, April 2014. [Online]. Available: <https://doi.org/10.1093/iwc/iwu009>
- [54] J. Danaher, "The threat of algocracy: Reality, resistance and accommodation," *Philosophy & Technology*, vol. 29, no. 3, pp. 245–268, January 2016. [Online]. Available: <https://doi.org/10.1007/s13347-015-0211-1>
- [55] "Programming Globalization: Visions and Revisions," in *Virtual Migration*. Duke University Press, April 2006, pp. 14–36. [Online]. Available: <https://doi.org/10.2307/j.ctv125jms5.5>
- [56] A. Aneesh, "Technologically coded authority: The post-industrial decline in bureaucratic hierarchies," in *International Summer Academy on Technology Studies*. Stanford University, July 2002. [Online]. Available: https://www.researchgate.net/publication/254843955_Technologically_Coded_Authority_The_Post-Industrial_Decline_in_Bureaucratic_Hierarchies
- [57] —, "Global labor: Algocratic modes of organization," *Sociological Theory*, vol. 27, no. 4, pp. 347–370, December 2009. [Online]. Available: <https://doi.org/10.1111/j.1467-9558.2009.01352.x>
- [58] R. Jaleel. (2021, October) Critical Race Theory and the Assault on Antiracist Thinking. [Online]. Available: <https://www.aaup.org/article/critical-race-theory-and-assault-antiracist-thinking> (Accessed: 2022-06-04)
- [59] J. C. P. Lin, "Exposing the chameleon-like nature of racism: a multidisciplinary look at critical race theory in higher education," *Higher Education*, pp. 1–16, May 2022. [Online]. Available: <https://doi.org/10.1007/s10734-022-00879-9>
- [60] N. J. Spalding and T. Phillips, "Exploring the use of vignettes: From validity to trustworthiness," *Qualitative Health Research*, vol. 17, no. 7, pp. 954–962, September 2007. [Online]. Available: <https://doi.org/10.1177/1049732307306187>
- [61] C. A. Pickover, *The Math Book: From Pythagoras to the 57th Dimension, 250 Milestones in the History of Mathematics*. New York / London: Sterling Publishing Co., Inc., 2009.
- [62] A. A. Al-Daffa', "Trigonometry," in *The Muslim Contribution to Mathematics*. London: Routledge, September 2020, pp. 67–79. [Online]. Available: <https://doi.org/10.4324/9781003074793-5>
- [63] B. L. van der Waerden, *A History of Algebra*. Springer Berlin Heidelberg, 1985. [Online]. Available: <https://doi.org/10.1007/978-3-642-51599-6>
- [64] M. G. Ames, "Deconstructing the algorithmic sublime," *Big Data & Society*, vol. 5, no. 1, January 2018. [Online]. Available: <https://doi.org/10.1177/2053951718779194>
- [65] D. Hilbert, W. Ackermann, R. Luce, and L. Hammond, *Principles of Mathematical Logic*, ser. AMS Chelsea Publishing Series. American Mathematical Society, 1999. [Online]. Available: <https://books.google.at/books?id=45ZGMjV9vfcC>
- [66] A. M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem. A Correction," *Proceedings of the London Mathematical Society*, vol. s2-43, no. 1, pp. 544–546, 1938. [Online]. Available: <https://doi.org/10.1112/plms/s2-43.6.544>

- [67] A. Church, "An Unsolvable Problem of Elementary Number Theory," *American Journal of Mathematics*, vol. 58, no. 2, p. 345, April 1936. [Online]. Available: <https://doi.org/10.2307/2371045>
- [68] Y. Gurevich, "Algorithms: A Quest for Absolute Definitions," *Bulletin of the European Association for Theoretical Computer Science Number 81*, pp. 195–225, 2003. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/164-algorithms-quest-absolute-definitions/>
- [69] D. E. Knuth, *The Art of Computer Programming - Volume 1: Fundamental Algorithms*. Boston: Addison-Wesley Professional, 1997.
- [70] E. Trist, *The Social Engagement of Social Science*. University of Pennsylvania Press, 1993, vol. Volume II.
- [71] F. W. Taylor, *The Principles of Scientific Management*. London: Harper & Brothers, 1911. [Online]. Available: <https://archive.org/stream/principlesofscie00taylrich#page/n5/mode/2up>
- [72] J. A. Konstan and J. Riedl, "Recommender systems: from algorithms to user experience," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 101–123, March 2012. [Online]. Available: <https://doi.org/10.1007/s11257-011-9112-x>
- [73] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA and London, England: Harvard University Press, January 2015. [Online]. Available: <https://doi.org/10.4159/harvard.9780674736061>
- [74] S. Zuboff, *The Age of Surveillance Capitalism*. New York, NY: PublicAffairs, January 2019.
- [75] G. Deleuze and F. Guattari, *Kafka: Toward a Minor Literature*, ser. Theory and History of Literature. Univ of Minnesota Press, 1986.
- [76] —, *Anti-Oedipus: Capitalism and Schizophrenia*. University Of Minnesota Press, 1983.
- [77] —, *Mille plateaux: Capitalisme et Schizophrénie*. Editions de Minuit, 1976, vol. 2.
- [78] J. W. Phillips, "Agencement/assemblage," *Theory, Culture & Society*, vol. 23, no. 2-3, pp. 108–109, May 2006. [Online]. Available: <https://doi.org/10.1177/026327640602300219>
- [79] G. Deleuze and C. Parnet, *Dialogues II*, ser. European Perspectives: A Series in Social Thought and Culture. Columbia University Press, 2007.
- [80] B. Latour, "Technology is society made durable," *The Sociological Review*, vol. 38, no. 1_suppl, pp. 103–131, May 1990. [Online]. Available: <https://doi.org/10.1111/j.1467-954x.1990.tb03350.x>
- [81] J. Law and J. Hassard, *Actor Network Theory and After*. Wiley-Blackwell, May 1999.
- [82] L. Winner, "Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology," *Science, Technology, & Human Values*, vol. 18, no. 3, pp. 362–378, July 1993. [Online]. Available: <https://doi.org/10.1177/016224399301800306>
- [83] M. Müller and C. Schurr, "Assemblage thinking and actor-network theory: conjunctions, disjunctions, cross-fertilisations," *Transactions of the Institute of British Geographers*, vol. 41, no. 3, pp. 217–229, March 2016. [Online]. Available: <https://doi.org/10.1111/tran.12117>
- [84] J. Law, "Actor network theory and material semiotics," in *The New Blackwell Companion to Social Theory*. Wiley-Blackwell, March 2009, pp. 141–158. [Online]. Available: <https://doi.org/10.1002/9781444304992.ch7>
- [85] M. Müller, "Assemblages and Actor-networks: Rethinking Socio-material Power, Politics and Space," *Geography Compass*, vol. 9, no. 1, pp. 27–41, January 2015. [Online]. Available: <https://doi.org/10.1111/gec3.12192>
- [86] D. McQuillan, "Data Science as Machinic Neoplatonism," *Philosophy & Technology*, vol. 31, no. 2, pp. 253–272, August 2017. [Online]. Available: <https://doi.org/10.1007/s13347-017-0273-3>
- [87] R. Kitchin, "Thinking critically about and researching algorithms," *Information, Communication & Society*, vol. 20, no. 1, pp. 14–29, February 2016. [Online]. Available: <https://doi.org/10.1080/1369118x.2016.1154087>

- [88] S. Jasanoff, *States of Knowledge: The Co-Production of Science and the Social Order*, ser. International Library of Sociology. Taylor & Francis, July 2004. [Online]. Available: <https://books.google.at/books?id=IYQArKR0ETwC>
- [89] S. Barocas, S. Hood, and M. Ziewitz, "Governing algorithms: A provocation piece," *SSRN Electronic Journal*, 2013. [Online]. Available: <https://doi.org/10.2139/ssrn.2245322>
- [90] S. L. Thomas, D. Nafus, and J. Sherman, "Algorithms as fetish: Faith and possibility in algorithmic work," *Big Data & Society*, vol. 5, no. 1, January 2018. [Online]. Available: <https://doi.org/10.1177/2053951717751552>
- [91] O. H. Gandy, "Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems," *Ethics and Information Technology*, vol. 12, no. 1, pp. 29–42, June 2009. [Online]. Available: <https://doi.org/10.1007/s10676-009-9198-6>
- [92] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm*. Microsoft Research, 2009. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>
- [93] D. Beer, "Power through the algorithm? Participatory web cultures and the technological unconscious," *New Media & Society*, vol. 11, no. 6, pp. 985–1002, September 2009. [Online]. Available: <https://doi.org/10.1177/1461444809336551>
- [94] D. K. Citron, "Technological Due Process," *University of Maryland School of Law Legal Studies Research Paper*, vol. 1, no. 26, pp. 1249–1313, September 2007. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1012360
- [95] J. Trippi, "How Technology Has Restored the Soul of Politics," *MIT Technology Review*, 2012. [Online]. Available: <https://www.technologyreview.com/2012/12/18/84609/how-technology-has-restored-the-soul-of-politics>
- [96] E. Oddleifson, "The effects of modern data analytics in electoral politics," *Political Science Undergraduate Review*, vol. 5, no. 1, pp. 46–52, April 2020. [Online]. Available: <https://doi.org/10.29173/psur130>
- [97] Y. Benkler, R. Faris, and H. Roberts, *Mammon's Algorithm*. Oxford University Press, October 2018. [Online]. Available: <https://doi.org/10.1093/oso/9780190923624.003.0009>
- [98] L. Susemichel and J. Wajcman, "Siri & die Jungs von Silicon-Valley," *an.schläge: Das feministische Magazin*, no. 5, 2017. [Online]. Available: <https://anschlaege.at/siri-die-jungs-von-silicon-valley/> (Accessed: 2020-11-1)
- [99] K. Fiscella, D. Tancredi, and P. Franks, "Adding socioeconomic status to Framingham scoring to reduce disparities in coronary risk assessment," *American Heart Journal*, vol. 157, no. 6, pp. 988–994, June 2009. [Online]. Available: <https://doi.org/10.1016/j.ahj.2009.03.019>
- [100] G. A. Kaplan and J. E. Keil, "Socioeconomic factors and cardiovascular disease: a review of the literature." *Circulation*, vol. 88, no. 4, pp. 1973–1998, October 1993. [Online]. Available: <https://doi.org/10.1161/01.cir.88.4.1973>
- [101] M. Seemann. (2018, March) Digitaltechnologie Blockchain: Eine als Technik getarnte Ideologie. [Online]. Available: <https://www.deutschlandfunkkultur.de/digitaltechnologie-blockchain-eine-als-technik-getarnte-100.html> (Accessed: 2022-06-17)
- [102] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems*, vol. 14, no. 3, pp. 330–347, July 1996. [Online]. Available: <https://doi.org/10.1145/230538.230561>
- [103] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89. [Online]. Available: <https://doi.org/10.1109/DSAA.2018.00018>
- [104] M. Skirpan and M. Gorelick, "The Authority of "Fair" in Machine Learning," *arXiv.org*, vol. cs.CY, June 2017. [Online]. Available: <http://arxiv.org/abs/1706.09976v2>

- [105] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin, “FairTest: Discovering unwarranted associations in data-driven applications,” in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, April 2017, p. 401–416. [Online]. Available: <https://doi.org/10.1109/eurosp.2017.29>
- [106] B. Hutchinson and M. Mitchell, “50 years of test (un)fairness,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, January 2019, pp. 49–58. [Online]. Available: <https://doi.org/10.1145/3287560.3287600>
- [107] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, January 2019, pp. 329–338. [Online]. Available: <https://doi.org/10.1145/3287560.3287589>
- [108] B. Taskesen, J. Blanchet, D. Kuhn, and V. A. Nguyen, “A statistical test for probabilistic fairness,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021, pp. 648–665. [Online]. Available: <https://doi.org/10.1145/3442188.3445927>
- [109] A. Z. Jacobs and H. Wallach, “Measurement and Fairness,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021, pp. 375–385. [Online]. Available: <https://doi.org/10.1145/3442188.3445901>
- [110] M. Andrus, E. Spitzer, J. Brown, and A. Xiang, “What We Can’t Measure, We Can’t Understand,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021, pp. 249–260. [Online]. Available: <https://doi.org/10.1145/3442188.3445888>
- [111] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press, January 2018.
- [112] A. Adensamer and L. D. Klausner, “Ich weiß, was du nächsten Sommer getan haben wirst,” *Zeitschrift für kritik - recht - gesellschaft*, no. 3, p. 419, 2019. [Online]. Available: <https://doi.org/10.33196/juridikum201903041901>
- [113] K. McGrory and N. Bedi. (2020, September) Targeted. [Online]. Available: <https://projects.tampabay.com/projects/2020/investigations/police-pasco-sheriff-targeted/intelligence-led-policing/> (Accessed: 2022-06-17)
- [114] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, “Runaway Feedback Loops in Predictive Policing,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 160–171. [Online]. Available: <https://proceedings.mlr.press/v81/ensign18a.html>
- [115] B. J. Jefferson, “Predictable policing: Predictive crime mapping and geographies of policing and race,” *Annals of the American Association of Geographers*, vol. 108, no. 1, pp. 1–16, May 2017. [Online]. Available: <https://doi.org/10.1080/24694452.2017.1293500>
- [116] A. G. Ferguson, “Policing Predictive Policing,” *Washington University Law Review*, vol. 94, no. 1109, 2017. [Online]. Available: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5/
- [117] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, July 2021. [Online]. Available: <https://doi.org/10.1145/3457607> (Accessed: 2021-08-04)
- [118] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries,” *Frontiers in Big Data*, vol. 2, p. 13, July 2019. [Online]. Available: <https://doi.org/10.3389/fdata.2019.00013> (Accessed: 2021-08-15)
- [119] H. Suresh and J. Guttag, “A framework for understanding sources of harm throughout the machine learning life cycle,” in *Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO ’21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: <https://doi.org/10.1145/3465416.3483305>
- [120] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, “The ethics of algorithms: Mapping the debate,” *Big Data & Society*, vol. 3, no. 2, December 2016. [Online]. Available: <https://doi.org/10.1177/2053951716679679>

- [121] T. C. C. Austria, “RIS - G77/2018 - Entscheidungstext zu § 2 Abs. 2 Z 3 des Bundesgesetzes über die Regelung des Personenstandswesens - Verfassungsgerichtshof (VfGH),” December 2021. [Online]. Available: https://www.ris.bka.gv.at/Dokument.wxe?Abfrage=Vfgh&Dokumentnummer=JFT_20180615_18G00077_00 (Accessed: 2021-21-27)
- [122] D. A. Norman, “The ‘problem’ with automation: inappropriate feedback and interaction, not ‘over-automation’,” *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 327, no. 1241, pp. 585–593, April 1990. [Online]. Available: <https://doi.org/10.1098/rstb.1990.0101>
- [123] C. E. Billings, *Aviation Automation: The Search for A Human-centered Approach*, ser. Human Factors in Transportation. CRC Press, 1997.
- [124] K. Christoffersen and D. Woods, “How to make automated systems team players,” in *Advances in Human Performance and Cognitive Engineering Research*. Elsevier, 2002, pp. 1–12. [Online]. Available: [https://doi.org/10.1016/s1479-3601\(02\)02003-9](https://doi.org/10.1016/s1479-3601(02)02003-9)
- [125] E. Rader, K. Cotter, and J. Cho, “Explanations as mechanisms for supporting algorithmic transparency,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. the 2018 CHI Conference. ACM, April 2018, pp. 1 – 13. [Online]. Available: <https://doi.org/10.1145/3173574.3173677>
- [126] R. F. Kizilcec, “How much information?” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. 2016 CHI Conference. ACM, May 2016, pp. 2390 – 2395. [Online]. Available: <https://doi.org/10.1145/2858036.2858402>
- [127] J. Burrell, “How the machine ‘thinks’: Understanding opacity in machine learning algorithms,” *Big Data & Society*, vol. 3, no. 1, January 2016. [Online]. Available: <https://doi.org/10.1177/2053951715622512>
- [128] G. Skraaning and G. A. Jamieson, “Human performance benefits of the automation transparency design principle,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 63, no. 3, pp. 379–401, December 2019. [Online]. Available: <https://doi.org/10.1177/0018720819887252>
- [129] M. Ananny and K. Crawford, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,” *New Media & Society*, vol. 20, no. 3, pp. 973–989, December 2016. [Online]. Available: <https://doi.org/10.1177/1461444816676645>
- [130] M. Janssen, R. Matheus, J. Longo, and V. Weerakkody, “Transparency-by-design as a foundation for open government,” *Transforming Government: People, Process and Policy*, vol. 11, no. 1, pp. 2–8, March 2017. [Online]. Available: <https://doi.org/10.1108/tg-02-2017-0015>
- [131] B. Mittelstadt, C. Russell, and S. Wachter, “Explaining explanations in AI,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’19. ACM, January 2019, pp. 279 – 288. [Online]. Available: <https://doi.org/10.1145/3287560.3287574>
- [132] R. Brauneis and E. P. Goodman, “Algorithmic transparency for the smart city,” *SSRN Electronic Journal*, 2017. [Online]. Available: <https://doi.org/10.2139/ssrn.3012499>
- [133] N. Diakopoulos and M. Koliska, “Algorithmic transparency in the news media,” *Digital Journalism*, vol. 5, no. 7, pp. 809–828, July 2016. [Online]. Available: <https://doi.org/10.1080/21670811.2016.1208053>
- [134] R. S. Geiger, “Beyond opening up the black box: Investigating the role of algorithmic systems in wikipedia organizational culture,” *Big Data & Society*, vol. 4, no. 2, September 2017. [Online]. Available: <https://doi.org/10.1177/2053951717730735>
- [135] A. Kunze, S. J. Summerskill, R. Marshall, and A. J. Filtness, “Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces,” *Ergonomics*, vol. 62, no. 3, pp. 345–360, February 2019. [Online]. Available: <https://doi.org/10.1080/00140139.2018.1547842>
- [136] H. Nissenbaum, “Computing and accountability,” *Communications of the ACM*, vol. 37, no. 1, pp. 72–80, January 1994. [Online]. Available: <https://doi.org/10.1145/175222.175228>
- [137] N. Diakopoulos, “Algorithmic Accountability Reporting: On The Investigation Of Black Boxes,” *Tow Center for Digital Journalism*, pp. 1 – 37, February 2014. [Online]. Available: http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf

- [138] —, “Accountability in algorithmic decision making,” *Communications of the ACM*, vol. 59, no. 2, pp. 56–62, January 2016. [Online]. Available: <https://doi.org/10.1145/2844110>
- [139] T. Z. Zarsky, “Governmental Data Mining and its Alternatives,” *Penn State Law Review*, vol. 116, no. 2, pp. 285–330, December 2011. [Online]. Available: <https://ssrn.com/abstract=1983326>
- [140] —, “Transparent Predictions,” *University Of Illinois Law Review*, vol. 2013, no. 4, p. 1503, 2013. [Online]. Available: <https://ssrn.com/abstract=2324240>
- [141] R. E. Bellman, *Dynamic programming*. Princeton, NJ: Princeton University Press, October 1957.
- [142] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, October 2012. [Online]. Available: <https://doi.org/10.1145/2347736.2347755>
- [143] —, *The master algorithm*. London, England: Basic Books, September 2015.
- [144] V. Belle and I. Papantonis, “Principles and practice of explainable machine learning,” *Frontiers in Big Data*, vol. 4, July 2021. [Online]. Available: <https://doi.org/10.3389/fdata.2021.688969>
- [145] Z. C. Lipton, “The mythos of model interpretability,” *Queue*, vol. 16, no. 3, pp. 31–57, June 2018. [Online]. Available: <https://doi.org/10.1145/3236386.3241340>
- [146] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, “Recent advances in convolutional neural networks,” *Pattern Recognition*, vol. 77, pp. 354–377, May 2018. [Online]. Available: <https://doi.org/10.1016/j.patcog.2017.10.013>
- [147] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 International Conference on Engineering and Technology (ICET)*. IEEE, August 2017, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/icengtechnol.2017.8308186>
- [148] B. Herman, “The promise and peril of human evaluation for model interpretability,” in *Proceedings of NIPS 2017 Symposium on Interpretable Machine Learning*. arXiv, 2017. [Online]. Available: <https://arxiv.org/abs/1711.07414>
- [149] T. Miller, P. Howe, and L. Sonenberg, “Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences,” in *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, 2017. [Online]. Available: <http://people.eng.unimelb.edu.au/tmiller/pubs/explanation-inmates.pdf>
- [150] J. Singh, C. Millard, C. Reed, J. Cobbe, and J. Crowcroft, “Accountability in the IoT: Systems, law, and ways forward,” *Computer*, vol. 51, no. 7, pp. 54–65, July 2018. [Online]. Available: <https://doi.org/10.1109/mc.2018.3011052>
- [151] R. Binns, “Algorithmic accountability and public reason,” *Philosophy & Technology*, vol. 31, no. 4, pp. 543–556, May 2017. [Online]. Available: <https://doi.org/10.1007/s13347-017-0263-5>
- [152] W. Schulz, K. Turk, B. d. l. Chapelle, J. Hörnle, T. Kersevan-Smokvina, M. Kettemann, D. Nielandt, A. Nedyak, P. Podvinskis, T. Schneider, S. Stalla-Bourdillon, D. Voorhoof, and B. Wagner, “Algorithms and Human Rights,” Council of Europe, Tech. Rep. DGI(2017)12, March 2018. [Online]. Available: <https://www.coe.int/en/web/freedom-expression/algorithms-and-human-rights>
- [153] D. Neyland, “Bearing account-able witness to the ethical algorithmic system,” *Science, Technology, & Human Values*, vol. 41, no. 1, pp. 50–76, July 2015. [Online]. Available: <https://doi.org/10.1177/0162243915598056>
- [154] M. Veale, M. V. Kleek, and R. Binns, “Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. the 2018 CHI Conference. ACM, April 2018, pp. 1 – 14. [Online]. Available: <https://doi.org/10.1145/3173574.3174014>
- [155] M. Loi and M. Spielkamp, “Towards Accountability in the Use of Artificial Intelligence for Public Administrations,” in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, July 2021. [Online]. Available: <https://doi.org/10.1145/3461702.3462631>

- [156] M. Bovens, “Two concepts of accountability: Accountability as a virtue and as a mechanism,” *West European Politics*, vol. 33, no. 5, pp. 946–967, August 2010. [Online]. Available: <https://doi.org/10.1080/01402382.2010.486119>
- [157] J. G. Koppell, “Pathologies of accountability: ICANN and the challenge of “multiple accountabilities disorder”,” *Public Administration Review*, vol. 65, no. 1, pp. 94–108, January 2005. [Online]. Available: <https://doi.org/10.1111/j.1540-6210.2005.00434.x>
- [158] L. Floridi and J. Sanders, “On the morality of artificial agents,” *Minds and Machines*, vol. 14, no. 3, pp. 349–379, August 2004. [Online]. Available: <https://doi.org/10.1023/b:Amind.0000035461.63578.9d>
- [159] M. J. Dubnick, “Situating Accountability: Seeking Salvation for the Core Concept of Modern Governance,” 2007. [Online]. Available: <http://mjdubnick.dubnick.net/papersrw/2007/situacct.pdf>
- [160] S. Passi and S. J. Jackson, “Trust in Data Science,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–28, November 2018. [Online]. Available: <https://doi.org/10.1145/3274405>
- [161] D. Hyde. (2013, September) Who to blame when ‘computer says no’. [Online]. Available: <https://www.telegraph.co.uk/finance/personalfinance/bank-accounts/10324271/Who-to-blame-when-computer-says-no.html> (Accessed: 2022-01-19)
- [162] “Little Britain - Computer says no!” January 2022. [Online]. Available: <https://memegenerator.net/instance/71645180/little-britain-computer-says-no> (Accessed: 2022-01-28)
- [163] C. O’Kelly and M. J. Dubnick, “Dissecting the semantics of accountability and its misuse,” in *Quality of Governance*. Springer International Publishing, November 2019, pp. 45–79. [Online]. Available: https://doi.org/10.1007/978-3-030-21522-4_3
- [164] D. F. Thompson, “Moral responsibility of public officials: The problem of many hands,” *American Political Science Review*, vol. 74, no. 4, pp. 905–916, December 1980. [Online]. Available: <https://doi.org/10.2307/1954312>
- [165] P. Day, R. Klein, and Nuffield Trust, *Auditing the auditors*, ser. Nuffield Trust S. Norwich, England: TSO, September 2001.
- [166] R. Mulgan, “Comparing accountability in the public and private sectors,” *Australian Journal of Public Administration*, vol. 59, no. 1, pp. 87–97, March 2000. [Online]. Available: <https://doi.org/10.1111/1467-8500.00142>
- [167] U. P. P. C. (USACM), “Statement on Algorithmic Transparency and Accountability,” *Association for Computing Machinery*, pp. 1 – 2, 2017.
- [168] A. C. on Professional Ethics, *ACM Code of Ethics and Professional Conduct*. ACM, 2018. [Online]. Available: <https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf>
- [169] A. Ressayguier and R. Rodrigues, “AI ethics should not remain toothless! A call to bring back the teeth of ethics,” *Big Data & Society*, vol. 7, no. 2, July 2020. [Online]. Available: <https://doi.org/10.1177/2053951720942541>
- [170] The European Parliament and the Council of the European Union, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),” 2016. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [171] J. Thamkittikasem, B. Saunders, and K. Jin, “ADS Task Force Report,” NYC ADS TaskForce, New York, Tech. Rep., 2019. [Online]. Available: <https://www1.nyc.gov/site/adstaskforce/index.page>
- [172] Treasury Board of Canada. (2019, December) Directive on Automated Decision-Making. [Online]. Available: <https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592> (Accessed: 2022-01-17)

- [173] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building ethics into artificial intelligence," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 5527–5533. [Online]. Available: <https://doi.org/10.24963/ijcai.2018/779>
- [174] D. Estlund, "On following orders in an unjust war," *Journal of Political Philosophy*, vol. 15, no. 2, pp. 213–234, June 2007. [Online]. Available: <https://doi.org/10.1111/j.1467-9760.2007.00277.x>
- [175] V. Tadros, "Against following orders," in *To Do, To Die, To Reason Why*. Oxford University Press, July 2020, pp. 57–86. [Online]. Available: <https://doi.org/10.1093/oso/9780198831549.003.0004>
- [176] D. K. Citron and F. Pasquale, "The Scored Society - Due Process for Automated Predictions," *Washington Law Review*, vol. 89, no. 2014-8, January 2014. [Online]. Available: <https://ssrn.com/abstract=2376209>
- [177] B. Wagner, "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems," *Policy & Internet*, vol. 11, no. 1, pp. 104–122, January 2019. [Online]. Available: <https://doi.org/10.1002/poi3.198>
- [178] M. Bovens and S. Zouridis, "From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control," *Public Administration Review*, vol. 62, no. 2, pp. 174–184, January 2002. [Online]. Available: <https://doi.org/10.1111/0033-3352.00168>
- [179] J. Heller, *Catch-22*. London, United Kingdom: Prentice Hall & IBD, 1996.
- [180] S. Hodkinson, "The accountability vacuum," in *Safe as houses*. Manchester University Press, May 2019, pp. 159–190. [Online]. Available: <https://doi.org/10.7228/manchester/9781526141866.003.0006>
- [181] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish, "Algorithmic impact assessments and accountability," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, March 2021, pp. 735–746. [Online]. Available: <https://doi.org/10.1145/3442188.3445935>
- [182] Y. Clarke, "Algorithmic Accountability Act of 2019," 2019. [Online]. Available: <https://www.congress.gov/bill/116th-congress/house-bill/2231/text>
- [183] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, "An empirical study of the impact of modern code review practices on software quality," *Empirical Software Engineering*, vol. 21, no. 5, pp. 2146–2189, April 2015. [Online]. Available: <https://doi.org/10.1007/s10664-015-9381-9>
- [184] L. MacLeod, M. Greiler, M.-A. Storey, C. Bird, and J. Czerwonka, "Code reviewing in the trenches: Challenges and best practices," *IEEE Software*, vol. 35, no. 4, pp. 34–42, July 2018. [Online]. Available: <https://doi.org/10.1109/ms.2017.265100500>
- [185] J. Czerwonka, M. Greiler, and J. Tilford, "Code reviews do not find bugs. how the current code review best practice slows us down," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, vol. 2. IEEE, May 2015, pp. 27–28. [Online]. Available: <https://doi.org/10.1109/icse.2015.131>
- [186] E. Doğan and E. Tüzün, "Towards a taxonomy of code review smells," *Information and Software Technology*, vol. 142, February 2022. [Online]. Available: <https://doi.org/10.1016/j.infsof.2021.106737>
- [187] A. Decan, T. Mens, and E. Constantinou, "On the impact of security vulnerabilities in the npm package dependency network," in *Proceedings of the 15th International Conference on Mining Software Repositories*. ACM, May 2018, pp. 181–191. [Online]. Available: <https://doi.org/10.1145/3196398.3196401>
- [188] J. R. Herkert, "Ways of thinking about and teaching ethical problem solving: Microethics and macroethics in engineering," *Science and Engineering Ethics*, vol. 11, no. 3, pp. 373–385, 2005. [Online]. Available: <https://doi.org/10.1007/s11948-005-0006-3>
- [189] T. Barnes, D. Garcia, E. K. Hawthorne, M. A. Pérez-Quiñones, M. Skirpan, N. Beard, S. Bhaduri, C. Fiesler, and T. Yeh, "Ethics Education in Context," *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pp. 940–945, 2018. [Online]. Available: <https://doi.org/10.1145/3159450.3159573>

- [190] J. L. Hess and G. Fore, “A Systematic Literature Review of US Engineering Ethics Interventions,” *Science and Engineering Ethics*, vol. 24, no. 2, pp. 551–583, 2018. [Online]. Available: <https://doi.org/10.1007/s11948-017-9910-6>
- [191] A. Kramer, “Facebook’s hiring crisis: Engineers are turning down offers,” *Protocol*, October 2021. [Online]. Available: <https://www.protocol.com/workplace/facebook-docs-hiring-recruiting-crisis> (Accessed: 2022-10-26)
- [192] P. Soderling, “How to get engineers to work for you instead of Facebook or Google,” *Quartz*, December 2014. [Online]. Available: <https://qz.com/302490/how-to-get-engineers-to-work-for-you-instead-of-facebook-or-google> (Accessed: 2022-12-26)
- [193] N. Bowles, “‘I Don’t Really Want to Work for Facebook.’ So Say Some Computer Science Students.” *N.Y. Times*, November 2018. [Online]. Available: <https://www.nytimes.com/2018/11/15/technology/jobs-facebook-computer-science-students.html> (Accessed: 2022-10-26)
- [194] R. Caplan and d. boyd, “Isomorphism through algorithms: Institutional dependencies in the case of Facebook,” *Big Data & Society*, vol. 5, no. 1, p. 2053951718757253, 2018. [Online]. Available: <https://doi.org/10.1177/2053951718757253>
- [195] I. Rahwan, “Society-in-the-loop: programming the algorithmic social contract,” *Ethics and Information Technology*, vol. 20, no. 1, pp. 5–14, August 2017. [Online]. Available: <https://doi.org/10.1007/s10676-017-9430-8>
- [196] G. J. Brandsma and T. Schillemans, “The Accountability Cube: Measuring Accountability,” *Journal of Public Administration Research and Theory*, vol. 23, no. 4, pp. 953–975, September 2012. [Online]. Available: <https://doi.org/10.1093/jopart/mus034>
- [197] M. Bovens, T. Schillemans, and P. T. Hart, “Does Public Accountability Work? An Assessment Tool,” *Public Administration*, vol. 86, no. 1, pp. 225–242, March 2008. [Online]. Available: <https://doi.org/10.1111/j.1467-9299.2008.00716.x>
- [198] K. Crawford, “Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics,” *Science, Technology, & Human Values*, vol. 41, no. 1, pp. 77–92, June 2015. [Online]. Available: <https://doi.org/10.1177/0162243915589635>
- [199] I. Nonaka and H. Takeuchi, *The knowledge-creating company: How Japanese companies create the dynamics of innovation*, ser. Long Range Planning. Oxford University Press, 1995.
- [200] A. Sinclair, “The chameleon of accountability: Forms and discourses,” *Accounting, Organizations and Society*, vol. 20, no. 2-3, pp. 219–237, 1995. [Online]. Available: [https://doi.org/10.1016/0361-3682\(93\)e0003-y](https://doi.org/10.1016/0361-3682(93)e0003-y)
- [201] A. Meijer, “Transparency,” in *The Oxford Handbook of Public Accountability*. Oxford University Press, 05 2014. [Online]. Available: <https://doi.org/10.1093/oxfordhb/9780199641253.013.0043>
- [202] A. Vedder and L. Naudts, “Accountability for the use of algorithms in a big data environment,” *International Review of Law, Computers & Technology*, vol. 0, no. 2, pp. 1 – 19, 03 2017. [Online]. Available: <https://doi.org/10.1080/13600869.2017.1298547>
- [203] S. Herbek, H. A. Eisl, M. Hurch, A. Schator, S. Sabutsch, G. Rauchegger, A. Kollmann, T. Philippi, P. Dragon, E. Seitz, and S. Repas, “The Electronic Health Record in Austria: a strong network between health care and patients,” *European Surgery*, vol. 44, no. 3, pp. 155–163, 2012. [Online]. Available: <https://doi.org/10.1007/s10353-012-0092-9>
- [204] W. O. Hackl, A. Hoerbst, and E. Ammenwerth, “Why the Hell Do We Need Electronic Health Records??” *Methods of Information in Medicine*, vol. 50, no. 01, pp. 53–61, 2011. [Online]. Available: <https://doi.org/10.3414/me10-02-0020>
- [205] M. Bovens and T. Schillemans, “Meaningful accountability,” in *The Oxford Handbook of Public Accountability*. Oxford University Press, 05 2014. [Online]. Available: <https://doi.org/10.1093/oxfordhb/9780199641253.013.0038>

- [206] W. Blischke, *Product Warranty Handbook*. Taylor & Francis, 1995. [Online]. Available: <https://books.google.at/books?id=0A1Vh1Iwc1IC>
- [207] S. W. Waller, J. G. Brady, R. Acosta, J. Fair, and J. Morse, “Consumer Protection in the United States: An Overview,” *European Journal of Consumer Law*, January 2011. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1000226
- [208] G. Fitzpatrick, “Mind the gap: Modelling the human in human-centric computing,” in *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, vol. 00. IEEE, October 2018, pp. 3–3. [Online]. Available: <https://doi.org/10.1109/vlhcc.2018.8506554>
- [209] W. Pedrycz, “Granular computing for data analytics: a manifesto of human-centric computing,” *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 6, pp. 1025–1034, November 2018. [Online]. Available: <https://doi.org/10.1109/jas.2018.7511213>
- [210] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, “Power to the people: The role of humans in interactive machine learning,” *AI Magazine*, vol. 35, no. 4, pp. 105–120, December 2014. [Online]. Available: <https://doi.org/10.1609/aimag.v35i4.2513>
- [211] T. Kulesza, “Toward end-user debugging of machine-learned classifiers,” in *2010 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, September 2010, pp. 253 – 254. [Online]. Available: <https://doi.org/10.1109/vlhcc.2010.45>
- [212] H. Werthner, E. A. Lee, H. Akkermans, M. Vardi, C. Ghezzi, N. Magnenat-Thalmann, H. Nowotny, L. Hardman, O. Stock, J. Larus, M. Aiello, E. Nardelli, M. Stampfer, C. Frauenberger, M. Ortiz, P. Reichl, V. Schiaffonati, C. Tsigkanos, W. Aspray, M. E. d. Bruijn, M. Strassnig, J. Neidhardt, N. Forgo, M. Hauswirth, G. G. Parker, M. Sertkan, A. Stanger, P. Knees, G. Tamburrini, H. Tellioglu, F. Ricci, and I. Nalis-Neuner, “Vienna Manifesto on Digital Humanism,” 2019. [Online]. Available: <https://dighum.ec.tuwien.ac.at/dighum-manifesto/>
- [213] A. Giddens, *The Constitution of Society*. Polity, 1986.
- [214] J. Butler, *Gender Trouble*, 1st ed. London, England: Routledge, 1990. [Online]. Available: <https://doi.org/10.4324/9780203902752>
- [215] A. Bandura, “Human Agency in Social Cognitive Theory,” *American Psychologist*, vol. 44, no. 9, pp. 1175–1184, 1989. [Online]. Available: <https://doi.org/10.1037/0003-066x.44.9.1175>
- [216] J. Maze, “Normativity versus normalisation: reassembling actor-network theory through Butler and Foucault,” *Culture, Theory and Critique*, vol. 61, no. 4, pp. 389–403, 2020. [Online]. Available: <https://doi.org/10.1080/14735784.2020.1780623>
- [217] A. Bandura, *Self efficacy*. New York, NY: W.H. Freeman, 1997.
- [218] —, “Toward a Psychology of Human Agency,” *Perspectives on Psychological Science*, vol. 1, no. 2, pp. 164–180, 2006.
- [219] —, “Social foundations of thought and action: A social cognitive theory.” *Social foundations of thought and action: A social cognitive theory.*, pp. xiii, 617–xiii, 617, 1986.
- [220] J. Häkli, “The subject of citizenship – Can there be a posthuman civil society?” *Political Geography*, vol. 67, pp. 166–175, 2018. [Online]. Available: <https://doi.org/10.1016/j.polgeo.2017.08.006>
- [221] J. Bennett, “The Agency of Assemblages and the North American Blackout,” *Public Culture*, vol. 17, no. 3, pp. 445–466, 2005. [Online]. Available: <https://doi.org/10.1215/08992363-17-3-445>
- [222] S. R. Krause, “Bodies in Action: Corporeal Agency and Democratic Politics,” *Political Theory*, vol. 39, no. 3, pp. 299–324, 2011. [Online]. Available: <https://doi.org/10.1177/0090591711400025>
- [223] M. Noorman, “Computing and moral responsibility,” in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018. [Online]. Available: <https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility/>

- [224] J. Bentham and R. Smith, *The Rationale of Reward*, ser. The Making of the Modern Law. John and H. L. Hunt, 1825.
- [225] J. S. Mill and R. Crisp, *Utilitarianism*. Oxford University Press, USA, January 1998.
- [226] L. Alexander and M. Moore, “Deontological ethics,” in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2016. [Online]. Available: <https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/>
- [227] H. Nissenbaum, “Accountability in a computerized society,” *Science and Engineering Ethics*, vol. 2, no. 1, pp. 25–42, March 1996. [Online]. Available: <https://doi.org/10.1007/bf02639315>
- [228] S. Wachter, B. Mittelstadt, and L. Floridi, “Transparent, explainable, and accountable AI for robotics,” *Science Robotics*, vol. 2, no. 6, p. eaan6080, May 2017. [Online]. Available: <https://doi.org/10.1126/scirobotics.aan6080>
- [229] P. Ferris. An introduction to explainable AI, and why we need it. [Online]. Available: <https://medium.freecodecamp.org/an-introduction-to-explainable-ai-and-why-we-need-it-a326417dd000> (Accessed: 2021-05-02)
- [230] J. Ladd, *Computers and Moral Responsibility: A Framework for an Ethical Analysis*. San Diego, CA, USA: Academic Press Professional, Inc., 1991, p. 664–675. [Online]. Available: <https://dl.acm.org/doi/10.5555/117868.117917>
- [231] M. Taddeo and L. Floridi, “The debate on the moral responsibilities of online service providers,” *Science and Engineering Ethics*, vol. 22, no. 6, pp. 1575–1603, November 2015. [Online]. Available: <https://doi.org/10.1007/s11948-015-9734-1>
- [232] B. Goodman and S. Flaxman, “European Union Regulations on Algorithmic Decision-Making and a ‘Right to Explanation’,” *AI Magazine*, vol. 38, no. 3, pp. 50–57, October 2017. [Online]. Available: <https://doi.org/10.1609/aimag.v38i3.2741>
- [233] A. Christin, A. Rosenblat, and d. boyd, “Courts and predictive algorithms,” *Data & Society*, 2017. [Online]. Available: <https://datasociety.net/output/data-civil-rights-courts-and-predictive-algorithms/> (Accessed: 2022-06-17)
- [234] T. Brennan, W. Dieterich, and B. Ehret, “Evaluating the predictive validity of the compas risk and needs assessment system,” *Criminal Justice and Behavior*, vol. 36, no. 1, pp. 21–40, October 2008. [Online]. Available: <https://doi.org/10.1177/0093854808326545>
- [235] R. Parasuraman and V. Riley, “Humans and automation: Use, misuse, disuse, abuse,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 2, pp. 230–253, June 1997. [Online]. Available: <https://doi.org/10.1518/00187209778543886>
- [236] M. Cummings, “Automation bias in intelligent time critical decision support systems,” in *AIAA 1st Intelligent Systems Technical Conference*. American Institute of Aeronautics and Astronautics, June 2004. [Online]. Available: <https://doi.org/10.2514/6.2004-6313>
- [237] J. Niklas, K. Sztandar-Sztanderska, K. Szymielewicz, A. Baczko-Dombi, and A. Walkowiak, “Profiling the unemployed in Poland: Social and political implications of algorithmic decision making,” *Fundacja Panoptykon, Tech. Rep.*, October 2015. [Online]. Available: https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf
- [238] D. G. Johnson, “Computer systems: Moral entities but not moral agents,” *Ethics and Information Technology*, vol. 8, no. 4, pp. 195–204, November 2006. [Online]. Available: <https://doi.org/10.1007/s10676-006-9111-5>
- [239] D. G. Johnson and T. M. Powers, “Computer systems and responsibility: A normative look at technological complexity,” *Ethics and Information Technology*, vol. 7, no. 2, pp. 99–107, June 2005. [Online]. Available: <https://doi.org/10.1007/s10676-005-4585-0>
- [240] L. Dogruel, “Too much information!? Examining the impact of different levels of transparency on consumers’ evaluations of targeted advertising,” *Communication Research Reports*, vol. 36, no. 5, pp. 383–392, October 2019. [Online]. Available: <https://doi.org/10.1080/08824096.2019.1684253>

- [241] G. Fitzpatrick, *The Locales Framework*. Springer Netherlands, 2003. [Online]. Available: <https://doi.org/10.1007/978-94-017-0363-5>
- [242] S. L. Star, "The Ethnography of Infrastructure," *American Behavioral Scientist*, vol. 43, no. 3, pp. 377–391, 1999. [Online]. Available: <https://doi.org/10.1177/00027649921955326>
- [243] S. J. Cunningham and M. Jones, "Autoethnography," in *Proceedings of the 6th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction making CHI natural - CHINZ '05*. ACM Press, 2005, pp. 1–8. [Online]. Available: <https://doi.org/10.1145/1073943.1073944>
- [244] C. C. Cain and E. Trauth, "Black men in IT," *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, vol. 48, no. 2, pp. 35–51, April 2017. [Online]. Available: <https://doi.org/10.1145/3084179.3084184>
- [245] A. Lucero, A. Desjardins, C. Neustaedter, K. Höök, M. Hassenzahl, and M. E. Cecchinato, "A sample of one," in *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. ACM, June 2019, pp. 385–388. [Online]. Available: <https://doi.org/10.1145/3301019.3319996>
- [246] A. Lucero, "Living without a mobile phone," in *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, June 2018, pp. 765–776. [Online]. Available: <https://doi.org/10.1145/3196709.3196731>
- [247] K. Höök, "Transferring qualities from horseback riding to design," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10*. ACM Press, 2010, pp. 226–235. [Online]. Available: <https://doi.org/10.1145/1868914.1868943>
- [248] N. K. Denzin and Y. S. Lincoln, *The SAGE Handbook of Qualitative Research*. SAGE, April 2011.
- [249] A. Ploder and J. Stadlbauer, "Strong Reflexivity and Its Critics," *Qualitative Inquiry*, vol. 22, no. 9, pp. 753–765, 2016. [Online]. Available: <https://doi.org/10.1177/1077800416658067>
- [250] P. Mayring, *Qualitative Inhaltsanalyse: Grundlagen und Techniken*, 12th ed. Beltz, 2015.
- [251] J. Saldaña, *The Coding Manual for Qualitative Researchers*, 3rd ed. SAGE Publications Ltd, 2015.
- [252] AlDanial, S. Snel, Jolkdarr, C. Beckmann, MichaelDimmitt, Roman, G. Chaves, J. Wilk, BoB Rudis, Asrmchq, A. Gough, J. Tang, J. Dursi, Achary, A. Ali, C. Ebberson, L. David, D. Ulrich, Erkmos, L. Brinkhoff, LoganDark, T. Irländer, W. Rösler, B1f6c1c4, S. Houzé, A. Ryan, A. Shinn, A. Mastrean, A. Molinaro, and A. Turner, "Aldanial/cloc: 1.92," December 2021. [Online]. Available: <https://zenodo.org/record/5760077>
- [253] G. A. Bowen, "Document analysis as a qualitative research method," *Qualitative Research Journal*, vol. 9, no. 2, pp. 27–40, August 2009. [Online]. Available: <https://doi.org/10.3316/qrj0902027>
- [254] D. Silverman, *Interpreting Qualitative Data*, 5th ed. SAGE, February 2015.
- [255] I. Wagner, *Critical Reflections on Participation in Design*. Oxford University Press, April 2018. [Online]. Available: <https://doi.org/10.1093/oso/9780198733249.003.0008>
- [256] N. B. Hansen, C. Dindler, K. Halskov, O. S. Iversen, C. Bossen, D. A. Basballe, and B. Schouten, "How participatory design works," in *Proceedings of the 31st Australian Conference on Human-Computer Interaction*. ACM, December 2019. [Online]. Available: <https://doi.org/10.1145/3369457.3369460>
- [257] S. Kemmis and R. McTaggart, "Participatory Action Research," in *Strategies of Qualitative Inquiry*. SAGE, 2008, p. 283.
- [258] F. Kensing and J. Greenbaum, "Heritage: having a say," in *Routledge International Handbook of Participatory Design*. Routledge, October 2012, pp. 41–56. [Online]. Available: <https://doi.org/10.4324/9780203108543-9>
- [259] P. Ehn, *Work-Oriented Design of Computer Artifacts*. Stockholm: Arbetslivscentrum, 1989.
- [260] P. Ehn and M. Kyng, *Cardboard Computers: Mocking-It-up or Hands-on the Future*. USA: L. Erlbaum Associates Inc., 1992, p. 169–196.

- [261] D. Novick, "Testing documentation with "low-tech" simulation," in *18th Annual Conference on Computer Documentation. ipcc sigdoc 2000. Technology and Teamwork. Proceedings. IEEE Professional Communication Society International Professional Communication Conference and ACM Special Interest Group on Documentation Conference*. IEEE, 2000, pp. 55–68. [Online]. Available: <https://doi.org/10.1109/ipcc.2000.887261>
- [262] S. Bødker and K. Grønbaek, "Cooperative prototyping: users and designers in mutual activity," *International Journal of Man-Machine Studies*, vol. 34, no. 3, pp. 453–478, March 1991. [Online]. Available: [https://doi.org/10.1016/0020-7373\(91\)90030-b](https://doi.org/10.1016/0020-7373(91)90030-b)
- [263] P. Atkinson and A. Coffey, "Analysing documentary realities," *Qualitative Research*, pp. 77–92, 2011-01.
- [264] A. Szigetvari. (2018, October) AMS-Vorstand Kopf: "Was die EDV gar nicht abbilden kann, ist die Motivation". [Online]. Available: <https://www.derstandard.at/story/2000089096795/ams-vorstand-kopf-menschliche-komponente-wird-entscheidend-bleiben> (Accessed: 2022-04-01)
- [265] ——. (2018, October) AMS bewertet Arbeitslose künftig per Algorithmus. [Online]. Available: <https://www.derstandard.at/story/2000089095393/ams-bewertet-arbeitslose-kuenftig-per-algorithmus> (Accessed: 2022-04-01)
- [266] L. Poirier, "Reading datasets: Strategies for interpreting the politics of data signification," *Big Data & Society*, vol. 8, no. 2, July 2021. [Online]. Available: <https://doi.org/10.1177/20539517211029322>
- [267] L. Bartlett and F. Vavrus, "Comparative case studies: An innovative approach," *Nordic Journal of Comparative and International Education (NJCIE)*, vol. 1, no. 1, July 2017. [Online]. Available: <https://doi.org/10.7577/njcie.1929>
- [268] E. Babbie, *The practice of social research*, 13th ed. Belmont, CA: Wadsworth Publishing, 2012.
- [269] H. S. Becker, "How to do qualitative research?" *International Journal of Communication*, vol. 3, p. 9, 2009.
- [270] "Eingabeübersicht EnerCoach: Brauchwarmwasser (BWW)," Energiestadt, Tech. Rep., 2020. [Online]. Available: https://www.local-energy.swiss/dam/jcr:c8d497a8-9a6c-43c7-9654-3cd710f8284b/Eingabeuebersicht_Brauchwarmwasser_EnerCoach_1.pdf
- [271] IPCC, "Summary for Policymakers," in *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, P. Shukla, J. Skea, R. Slade, A. A. Khourdajie, R. v. Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley, Eds. Cambridge, UK: Cambridge University Press, 2022. [Online]. Available: <https://www.ipcc.ch/report/ar6/wg3/>
- [272] "Transforming our world: the 2030 agenda for sustainable development," United Nations, New York, USA, Tech. Rep., 2015. [Online]. Available: <http://digitallibrary.un.org/record/1654217>
- [273] S. Gabrielli, P. Forbes, A. Jylhä, S. Wells, M. Sirén, S. Hemminki, P. Nurmi, R. Maimone, J. Masthoff, and G. Jacucci, "Design challenges in motivating change for sustainable urban mobility," *Computers in Human Behavior*, vol. 41, no. C, pp. 416–423, December 2014. [Online]. Available: <https://doi.org/10.1016/j.chb.2014.05.026>
- [274] J. Meurer, D. Lawo, L. Janßen, and V. Wulf, "Designing mobility eco-feedback for elderly users," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM, May 2016, pp. 921–926. [Online]. Available: <https://doi.org/10.1145/2851581.2851599>
- [275] G. Broll, H. Cao, P. Ebben, P. Holleis, K. Jacobs, J. Koolwaaij, M. Luther, and B. Souville, "Tripzoom," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia - MUM '12*, ser. ACM. ACM Press, December 2012. [Online]. Available: <https://doi.org/10.1145/2406367.2406436>
- [276] J. Froehlich, T. Dillahunt, P. Klasnja, J. Mankoff, S. Consolvo, B. Harrison, and J. A. Landay, "UbiGreen," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. ACM. ACM, April 2009. [Online]. Available: <https://doi.org/10.1145/1518701.1518861>

- [277] M. Frejus and D. Martini, "Taking into account user appropriation and development to design energy consumption feedback," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, ser. the 33rd Annual ACM Conference Extended Abstracts. ACM, April 2015, pp. 2193 – 2198. [Online]. Available: <https://doi.org/10.1145/2702613.2732718>
- [278] A. Spagnoli, N. Corradi, L. Gamberini, E. Hoggan, G. Jacucci, C. Katzeff, L. Broms, and L. Jonsson, "Eco-feedback on the go: Motivating energy awareness," *Computer*, vol. 44, no. 5, pp. 38–45, May 2011. [Online]. Available: <https://doi.org/10.1109/mc.2011.125>
- [279] P. Bertoldi, T. Serrenho, and P. Zangheri, "Consumer Feedback Systems: How Much Energy Saving Will They Deliver and for How Long?" *ACEEE Summer Study on Energy Efficiency in Buildings*, pp. 1 – 13, August 2016. [Online]. Available: https://aceee.org/files/proceedings/2016/data/papers/12_769.pdf
- [280] Y. Strengers, "Beyond demand management: co-managing energy and water practices with australian households," *Policy Studies*, vol. 32, no. 1, pp. 35–58, January 2011. [Online]. Available: <https://doi.org/10.1080/01442872.2010.526413>
- [281] T. Schwartz, S. Deneff, G. Stevens, L. Ramirez, and V. Wulf, "Cultivating energy literacy," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. the SIGCHI Conference. ACM, April 2013. [Online]. Available: <https://doi.org/10.1145/2470654.2466154>
- [282] S. Georg and L. Justesen, "Counting to zero: accounting for a green building," *Accounting, Auditing & Accountability Journal*, vol. 30, no. 5, pp. 1065–1081, June 2017. [Online]. Available: <https://doi.org/10.1108/aaaj-04-2013-1320>
- [283] D. Casado-Mansilla, J. L. de Armentia, P. Garaizar, and D. L. de Ipiña, "To switch off the coffee-maker or not," in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. ACM, April 2014, pp. 2425–2430. [Online]. Available: <https://doi.org/10.1145/2559206.2581152>
- [284] C. Fischer, "Feedback on household electricity consumption: a tool for saving energy?" *Energy Efficiency*, vol. 1, no. 1, pp. 79–104, February 2008. [Online]. Available: <https://doi.org/10.1007/s12053-008-9009-7>
- [285] J. Kjeldskov, M. B. Skov, J. Paay, D. Lund, T. Madsen, and M. Nielsen, "Eco-forecasting for domestic electricity use," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, April 2015, pp. 1985–1988. [Online]. Available: <https://doi.org/10.1145/2702123.2702318>
- [286] H. Hasselqvist, C. Bogdan, M. Romero, and O. Shafqat, "Supporting energy management as a cooperative amateur activity," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, April 2015, pp. 1483–1488. [Online]. Available: <https://doi.org/10.1145/2702613.2732724>
- [287] R. Yun, A. Aziz, P. Scupelli, B. Lasternas, C. Zhang, and V. Loftness, "Beyond eco-feedback," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, April 2015, pp. 1989–1992. [Online]. Available: <https://doi.org/10.1145/2702123.2702268>
- [288] J. Froehlich, L. Findlater, M. Ostergren, S. Ramanathan, J. Peterson, I. Wragg, E. Larson, F. Fu, M. Bai, S. Patel, and J. A. Landay, "The design and evaluation of prototype eco-feedback displays for fixture-level water usage data," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12. ACM, May 2012, pp. 2367–2376. [Online]. Available: <https://doi.org/10.1145/2207676.2208397>
- [289] J. Pierce, W. Odom, and E. Bleviss, "Energy aware dwelling," in *Proceedings of the 20th Australasian Conference on Computer-Human Interaction Designing for Habitus and Habitat - OZCHI '08*, ser. ACM. ACM Press, December 2008. [Online]. Available: <https://doi.org/10.1145/1517744.1517746>
- [290] M. Promann, "Examining the role visual graph structures play in collective awareness and cooperative decisions," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, April 2018, pp. 1–6. [Online]. Available: <https://doi.org/10.1145/3170427.3180303>
- [291] R. H. Thaler and C. R. Sunstein, "Nudge: improving decisions about health, wealth, and happiness," *Choice Reviews Online*, vol. 46, no. 02, pp. 46–0977–46–0977, October 2008. [Online]. Available: <https://doi.org/10.5860/choice.46-0977>

- [292] J. Meurer, D. Lawo, L. Janßen, and V. Wulf, “Designing mobility eco-feedback for elderly users,” in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. the 2016 CHI Conference Extended Abstracts. ACM, May 2016, pp. 921–926. [Online]. Available: <https://doi.org/10.1145/2851581.2851599>
- [293] H. Brynjarsdottir, M. Håkansson, J. Pierce, E. Baumer, C. DiSalvo, and P. Sengers, “Sustainably unpersuaded,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. the 2012 ACM annual conference. ACM, May 2012, pp. 947–956. [Online]. Available: <https://doi.org/10.1145/2207676.2208539>
- [294] W. C. Turner and S. A. Parker, “Energy accounting: A new field develops,” *IIE Transactions*, vol. 16, no. 2, pp. 135–143, June 1984. [Online]. Available: <https://doi.org/10.1080/07408178408974678>
- [295] A. Consonni and J. Lesourd, “Industrial energy accounting and control systems: A survey,” *Energy Conversion and Management*, vol. 26, no. 3-4, pp. 357–361, January 1986. [Online]. Available: [https://doi.org/10.1016/0196-8904\(86\)90017-8](https://doi.org/10.1016/0196-8904(86)90017-8)
- [296] B. J. Cornwall, “How to Organize and Communicate Your Energy Data. A Guide to Energy Accounting.” California Energy Extension Service, Sacramento, Tech. Rep., 1984. [Online]. Available: <https://files.eric.ed.gov/fulltext/ED251934.pdf>
- [297] L. Shao, G. Chen, Z. Chen, S. Guo, M. Han, B. Zhang, T. Hayat, A. Alsaedi, and B. Ahmad, “Systems accounting for energy consumption and carbon emission by building,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 19, no. 6, pp. 1859–1873, June 2014. [Online]. Available: <https://doi.org/10.1016/j.cnsns.2013.10.003>
- [298] J. Warnken, M. Bradley, and C. Guilding, “Exploring methods and practicalities of conducting sector-wide energy consumption accounting in the tourist accommodation industry,” *Ecological Economics*, vol. 48, no. 1, pp. 125–141, January 2004. [Online]. Available: <https://doi.org/10.1016/j.ecolecon.2003.08.007>
- [299] J. Emblemsvåg, *Reengineering Capitalism*. Springer International Publishing, 2016. [Online]. Available: <https://doi.org/10.1007/978-3-319-19689-3>
- [300] S. Kellner, “Flexible online energy accounting in TinyOS,” in *Real-World Wireless Sensor Networks*. Springer Berlin Heidelberg, 2010, pp. 62–73. [Online]. Available: https://doi.org/10.1007/978-3-642-17520-6_6
- [301] S. Lee, W. Jung, Y. Chon, and H. Cha, “EnTrack,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*. ACM Press, 2015. [Online]. Available: <https://doi.org/10.1145/2750858.2807531>
- [302] B. Schminke, “Overview of the current state of research on characteristics and algorithms of energy management systems in households and buildings,” *International Journal of Energy Research*, vol. 45, no. 10, pp. 14 194–14 206, May 2021. [Online]. Available: <https://doi.org/10.1002/er.6738>
- [303] C. Chen, S. Duan, T. Cai, B. Liu, and G. Hu, “Smart energy management system for optimal microgrid economic operation,” *IET Renewable Power Generation*, vol. 5, no. 3, p. 258, 2011. [Online]. Available: <https://doi.org/10.1049/iet-rpg.2010.0052>
- [304] T. Fiedler and P.-M. Mircea, “Energy management systems according to the ISO 50001 standard — Challenges and benefits,” in *2012 International Conference on Applied and Theoretical Electricity (ICATE)*. IEEE, October 2012. [Online]. Available: <https://doi.org/10.1109/icate.2012.6403411>
- [305] K. McGlenn, B. Yuce, H. Wicaksono, S. Howell, and Y. Rezgui, “Usability evaluation of a web-based tool for supporting holistic building energy management,” *Automation in Construction*, vol. 84, pp. 154–165, December 2017. [Online]. Available: <https://doi.org/10.1016/j.autcon.2017.08.033>
- [306] L. Energycap. (2022, April) EnergyCAP Energy Management & Utility Bill Accounting Software Solution. [Online]. Available: <https://www.energycap.com> (Accessed: 2022-04-15)
- [307] (2022, April) Sustainability & Energy Data Management Solutions - WatchWire. [Online]. Available: <https://watchwire.ai> (Accessed: 2022-04-15)

- [308] D. A. Work. (2022, April) Energy Accounting Software & Meter Data Management System | Power Costs, Inc. (PCI). [Online]. Available: <https://www.powercosts.com/solutions/energy-accounting> (Accessed: 2022-04-15)
- [309] C. Lebiere, P. Pirolli, R. Thomson, J. Paik, M. Rutledge-Taylor, J. Staszewski, and J. R. Anderson, "A functional model of sensemaking in a neurocognitive architecture," *Computational Intelligence and Neuroscience*, vol. 2013, no. 4124, pp. 1–29, 2013. [Online]. Available: <https://doi.org/10.1155/2013/921695>
- [310] G. Klein, B. Moon, and R. Hoffman, "Making sense of sensemaking 1: Alternative perspectives," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 70–73, July 2006. [Online]. Available: <https://doi.org/10.1109/mis.2006.75>
- [311] G. Wood, R. Day, E. Creamer, D. van der Horst, A. Hussain, S. Liu, A. Shukla, O. Iwaka, M. Gaterell, P. Petridis, N. Adams, and V. Brown, "Sensors, sense-making and sensitivities: UK household experiences with a feedback display on energy consumption and indoor environmental conditions," *Energy Research & Social Science*, vol. 55, pp. 93–105, September 2019. [Online]. Available: <https://doi.org/10.1016/j.erss.2019.04.013>
- [312] E. E. Awards, "European Energy Award Factsheet," European Energy Awards, Tech. Rep., 2020. [Online]. Available: https://www.european-energy-award.org/fileadmin/Documents/Download/FS_eea_2020_hoch.pdf
- [313] A. Dix, J. Finlay, G. D. Abowd, and R. Beale, *Human-computer interaction*, 3rd ed., 2004.
- [314] "sia | schweizerischer ingenieur- und architektenverein," February 2021. [Online]. Available: <https://www.sia.ch/de> (Accessed: 2021-02-12)
- [315] B. Dervin, "Sense-making theory and practice: an overview of user interests in knowledge seeking and use," *Journal of Knowledge Management*, vol. 2, no. 2, pp. 36–46, December 1998. [Online]. Available: <https://doi.org/10.1108/13673279810249369>
- [316] H. A. Reijers and I. T. P. Vanderfeesten, "Cohesion and coupling metrics for workflow process design," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, pp. 290–305. [Online]. Available: https://doi.org/10.1007/978-3-540-25970-1_19
- [317] U. Feige, "A threshold of $\ln n$ for approximating set cover," *Journal of the ACM (JACM)*, vol. 45, no. 4, pp. 634–652, 1998. [Online]. Available: <https://doi.org/10.1145/285055.285059>
- [318] C. D. Hundhausen, S. A. Douglas, and J. T. Stasko, "A meta-study of algorithm visualization effectiveness," *Journal of Visual Languages & Computing*, vol. 13, no. 3, pp. 259–290, June 2002. [Online]. Available: <https://doi.org/10.1006/jvlc.2002.0237>
- [319] A. F. Blackwell, "Pictorial representation and metaphor in visual language design," *Journal of Visual Languages & Computing*, vol. 12, no. 3, pp. 223–252, June 2001. [Online]. Available: <https://doi.org/10.1006/jvlc.2001.0207>
- [320] R. Luck, "What is it that makes participation in design participatory design?" *Design Studies*, vol. 59, pp. 1–8, Nov. 2018. [Online]. Available: <https://doi.org/10.1016/j.destud.2018.10.002>
- [321] B. Wimmer. (2018, October) Der AMS-Algorithmus ist ein „Paradebeispiel für Diskriminierung“. [Online]. Available: <https://futurezone.at/netzpolitik/der-ams-algorithmus-ist-ein-paradebeispiel-fuer-diskriminierung/400147421> (Accessed: 2022-04-30)
- [322] R. E. Bucklin, D. R. Lehmann, and J. D. C. Little, "From Decision Support to Decision Automation: A 2020 Vision," *Marketing Letters*, vol. 9, no. 3, pp. 235–246, 1998. [Online]. Available: <http://www.jstor.org/stable/40216167>
- [323] M. Spielkamp, "Automating Society," AlgorithmWatch, Tech. Rep., January 2019. [Online]. Available: https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf
- [324] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing algorithms: Research methods for detecting discrimination on internet platforms," May 2014. [Online]. Available: <http://www-personal.umich.edu/~csandvig/research/AuditingAlgorithms--Sandvig--ICA2014DataandDiscriminationPreconference.pdf>

- [325] J. Cobbe, M. S. A. Lee, and J. Singh, “Reviewable automated decision-making,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021. [Online]. Available: <https://doi.org/10.1145/3442188.3445921>
- [326] L. Edwards and M. Veale, “Enslaving the algorithm: From a “right to an explanation” to a “right to better decisions”?” *IEEE Security & Privacy*, vol. 16, no. 3, pp. 46–54, May 2018. [Online]. Available: <https://doi.org/10.1109/msp.2018.2701152>
- [327] M. Kuziemski and G. Misuraca, “AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings,” *Telecommunications Policy*, vol. 44, no. 6, p. 101976, July 2020. [Online]. Available: <https://doi.org/10.1016/j.telpol.2020.101976>
- [328] B. Marczak, J. Scott-Railton, S. McKune, B. A. Razzak, and R. Deibert, “Hide and Seek: Tracking NSO Group’s Pegasus Spyware to Operations in 45 Countries,” *Citizen Lab, Tech. Rep. 113*, 2018. [Online]. Available: <https://citizenlab.ca/2018/09/hide-and-peek-tracking-nso-groups-pegasus-spyware-to-operations-in-45-countries/> (Accessed: 2022-06-17)
- [329] O. Penz, B. Sauer, M. Gaitsch, J. Hofbauer, and B. Glinsner, “Post-bureaucratic encounters: Affective labour in public employment services,” *Critical Social Policy*, vol. 37, no. 4, pp. 540–561, January 2017. [Online]. Available: <https://doi.org/10.1177/0261018316681286>
- [330] C. Pollitt and G. Bouckaert, *Public management reform: A Comparative Analysis — New Public Management, Governance, and the Neo-Weberian State*. London, England: Oxford University Press, 2011.
- [331] O. Penz, B. Glinsner, M. Gaitsch, J. Hofbauer, and B. Sauer, “Affektive Interaktionsarbeit in der öffentlichen Arbeitsvermittlung in Österreich, Deutschland und der Schweiz,” *AIS-Studien*, 2015. [Online]. Available: <https://www.ssoar.info/ssoar/handle/document/64812>
- [332] A. S. Pascual, “Reshaping Welfare States: Activation Regimes in Europe,” in *Reshaping Welfare States and Activation Regimes in Europe*, A. S. Pascual and L. Magnusson, Eds. European Interuniversity Press, 2007.
- [333] S. Jasanoff, “Future Imperfect: Science, Technology, and the Imaginations of Modernity,” pp. 1 – 49, October 2014.
- [334] S. Desiere, K. Langenbacher, and L. Struyven, “Statistical profiling in public employment services,” vol. 224, pp. 1 – 29, February 2019. [Online]. Available: <https://doi.org/10.1787/b5e5f16e-en>
- [335] A.-K. Will, “The German statistical category “migration background”: Historical roots, revisions and shortcomings,” *Ethnicities*, vol. 19, no. 3, pp. 535 – 557, March 2019. [Online]. Available: <http://journals.sagepub.com/doi/10.1177/1468796819833437>
- [336] Parlament der Bundesrepublik Österreich, “Bundesgesetz zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten (Datenschutzgesetz – DSGVO),” 2022. [Online]. Available: <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=bundesnormen&Gesetzesnummer=10001597>
- [337] —, “Bundesgesetz über das Arbeitsmarktservice (Arbeitsmarktservicegesetz – AMSG),” 2022. [Online]. Available: <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008905>
- [338] B. Wimmer. (2018, October) AMS-Chef: “Mitarbeiter schätzen Jobchancen pessimistischer ein als der Algorithmus”. [Online]. Available: <https://futurezone.at/netzpolitik/ams-chef-mitarbeiter-schaetzen-jobchancen-pessimistischer-ein-als-der-algorithmus/400143839> (Accessed: 2022-05-13)
- [339] (2022, May) Anderl fordert mehr AMS Personal. [Online]. Available: https://www.arbeiterkammer.at/service/presse/AMS_PERSONAL_Presseaussendung.html (Accessed: 2022-05-13)
- [340] R. Böheim, R. Eppel, and H. Mahringer, “Die Auswirkungen einer Verbesserung der Betreuungrelation für Arbeitslose in der Arbeitsvermittlung des AMS,” *Austrian Institute of Economic Research, Vienna, Tech. Rep. 2017/393-1/S/WIFO-Projektnummer: 9814*, 2017. [Online]. Available: https://www.wifo.ac.at/jart/prj3/wifo/resources/person_dokument/person_dokument.jart?publikationsid=61298&mime_type=application/pdf

- [341] G. E. O'Brien, *Psychology of work and unemployment*. John Wiley & Sons, 1986.
- [342] R. Liem and J. H. Liem, "Psychological effects of unemployment on workers and their families," *Journal of Social Issues*, vol. 44, no. 4, pp. 87–105, January 1988. [Online]. Available: <https://doi.org/10.1111/j.1540-4560.1988.tb02093.x>
- [343] P. Norlander, G. C. Ho, M. Shih, D. J. Walters, and T. L. Pittinsky, "The role of psychological stigmatization in unemployment discrimination," *Basic and Applied Social Psychology*, vol. 42, no. 1, pp. 29–49, November 2019. [Online]. Available: <https://doi.org/10.1080/01973533.2019.1689363>
- [344] D. S. Dougherty, J. M. Rick, and P. Moore, "Unemployment and social class stigmas," *Journal of Applied Communication Research*, vol. 45, no. 5, pp. 495–516, September 2017. [Online]. Available: <https://doi.org/10.1080/00909882.2017.1382708>
- [345] G. C. Bowker and S. L. Star, *Sorting things out*, ser. Inside technology. London, England: MIT Press, November 1999.
- [346] D. Weichselbaumer, "Discrimination against female migrants wearing headscarves," *SSRN Electronic Journal*, 2016. [Online]. Available: <https://doi.org/10.2139/ssrn.2842960>
- [347] H. Hofer, G. Titelbach, and D. Weichselbaumer, "Diskriminierung von MigrantInnen am österreichischen Arbeitsmarkt," pp. 1 – 119, January 2014. [Online]. Available: <https://irihs.ihs.ac.at/id/eprint/2246/1/IHSPR6311119.pdf>
- [348] V. Gaigg and M. Simoner. (2018, June) Verfassungsgerichtshof bestätigt Recht auf drittes Geschlecht. [Online]. Available: <https://www.derstandard.at/story/2000082511550/verfassungsgerichtshof-bestaetigt-recht-auf-drittes-geschlecht> (Accessed: 2022-05-22)
- [349] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *SSRN Electronic Journal*, 2016. [Online]. Available: <https://doi.org/10.2139/ssrn.2477899>
- [350] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im)possibility of fairness," vol. cs.CY, September 2016. [Online]. Available: <http://arxiv.org/abs/1609.07236v1>
- [351] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from berkeley," *Science*, vol. 187, no. 4175, pp. 398–404, February 1975. [Online]. Available: <https://doi.org/10.1126/science.187.4175.398>
- [352] M. Wattenberg, F. Viegas, and M. Hardt, "Attacking discrimination with smarter ML," 2018. [Online]. Available: <https://research.google.com/bigpicture/attacking-discrimination-in-ml/> (Accessed: 2022-06-17)
- [353] P. Gajane and M. Pechenizkiy, "On Formalizing Fairness in Prediction with Machine Learning," *arXiv.org*, vol. cs.LG, p. arXiv:1710.03184, October 2017. [Online]. Available: <https://arxiv.org/abs/1710.03184>
- [354] E. Tagiou, Y. Kanellopoulos, C. Aridas, and C. Makris, "A tool supported framework for the assessment of algorithmic accountability," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, July 2019. [Online]. Available: <https://doi.org/10.1109/iisa.2019.8900715>
- [355] (2022, October) ACM FAccT - 2022 Accepted Papers. [Online; accessed 3. Nov. 2022]. [Online]. Available: <https://facctconference.org/2022/acceptedpapers.html> (Accessed: 2022-11-03)
- [356] N. Anand. (2017) The banality of infrastructure. [Online]. Available: <https://items.ssrc.org/the-banality-of-infrastructure/>
- [357] S. Joyce, D. Neumann, V. Trappmann, and C. Umney, "A Global Struggle: Worker Protest in the Platform Economy," *SSRN Electronic Journal*, 2020. [Online]. Available: <https://doi.org/10.2139/ssrn.3540104>
- [358] O. Howitt and C. Baraniuk. (2016, April) These unlucky people have names that break computers. [Online]. Available: <https://www.bbc.com/future/article/20160325-the-names-that-break-computer-systems> (Accessed: 2022-05-26)
- [359] R. Munroe. (2022, May) Exploits of a Mom. [Online]. Available: <https://xkcd.com/327> (Accessed: 2022-05-26)

- [360] G. Wright, G. Cairns, and R. Bradfield, "Scenario methodology: New developments in theory and practice: Introduction to the special issue," *Technological Forecasting and Social Change*, vol. 80, no. 4, pp. 561–565, 2013, scenario Method: Current developments in theory and practice. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S004016251200296X>
- [361] N. J. Rowland and M. J. Spaniol, "Social foundation of scenario planning," *Technological Forecasting and Social Change*, vol. 124, pp. 6–15, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0040162517301890>
- [362] M. Aoyama, "Persona-and-scenario based requirements engineering for software embedded in digital consumer products," in *13th IEEE International Conference on Requirements Engineering (RE'05)*. IEEE, 2005, pp. 85–94. [Online]. Available: <https://doi.org/10.1109/re.2005.50>
- [363] P. J. White and F. Devitt, "Creating personas from design ethnography and grounded theory," *J. Usability Studies*, vol. 16, no. 3, p. 156–178, May 2021. [Online]. Available: <https://dl.acm.org/doi/pdf/10.5555/3532758.3532760>
- [364] T. Markussen and E. Knutz, "The poetics of design fiction," in *Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces - DPPI '13*. ACM Press, 2013. [Online]. Available: <https://doi.org/10.1145/2513506.2513531>
- [365] A. Buhmann, J. Paßmann, and C. Fieseler, "Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse," *Journal of Business Ethics*, vol. 163, no. 2, pp. 265–280, June 2019. [Online]. Available: <https://doi.org/10.1007/s10551-019-04226-4>
- [366] A. Henriksen, S. Enni, and A. Bechmann, "Situated accountability: Ethical principles, certification standards, and explanation methods in applied AI," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, July 2021. [Online]. Available: <https://doi.org/10.1145/3461702.3462564>
- [367] M. Busuioc, "Accountable Artificial Intelligence: Holding Algorithms to Account," *Public Administration Review*, vol. 81, no. 5, pp. 825–836, November 2020. [Online]. Available: <https://doi.org/10.1111/puar.13293>
- [368] D. F. Engstrom and D. E. Ho, "Algorithmic Accountability in the Administrative State," *Yale Journal on Regulation*, vol. 37, no. 800, 2020. [Online]. Available: <http://hdl.handle.net/20.500.13051/8311>
- [369] S. Bødker and C. N. Klokrose, "The Human–Artifact Model: An Activity Theoretical Approach to Artifact Ecologies," *Human–Computer Interaction*, vol. 26, no. 4, pp. 315–371, 2011.
- [370] A. N. Leont'ev, "The problem of activity in psychology," *Soviet Psychology*, vol. 13, no. 2, pp. 4–33, 1974. [Online]. Available: <https://doi.org/10.2753/RPO1061-040513024>
- [371] —, *Activity, consciousness and personality*. Old Tappan, NJ: Prentice Hall, 1979.
- [372] L. A. Suchman, *Learning in doing: Social, cognitive and computational perspectives: Plans and situated actions: The problem of human-machine communication*, 2nd ed. Cambridge, England: Cambridge University Press, 1987.
- [373] V. Kaptelinin and B. A. Nardi, *Acting with Technology: Activity Theory and Interaction Design*. London, England: MIT Press, 2006.
- [374] A. L. Institute, "Examining the Black Box: Tools for assessing algorithmic systems," Ada Lovelace Institute, Tech. Rep., April 2020. [Online]. Available: <https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/>
- [375] J. Thebault-Spieker, L. Terveen, and B. Hecht, "Toward a geographic understanding of the sharing economy," *ACM Transactions on Computer-Human Interaction*, vol. 24, no. 3, pp. 1–40, July 2017. [Online]. Available: <https://doi.org/10.1145/3058499>
- [376] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios, "Quantifying search bias," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, February 2017. [Online]. Available: <https://doi.org/10.1145/2998181.2998321>

- [377] L. Chen, A. Misllove, and C. Wilson, “Peeking beneath the hood of uber,” in *Proceedings of the 2015 Internet Measurement Conference*. ACM, October 2015. [Online]. Available: <https://doi.org/10.1145/2815675.2815681>
- [378] L. Chen, R. Ma, A. Hannák, and C. Wilson, “Investigating the impact of gender on rank in resume search engines,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, April 2018. [Online]. Available: <https://doi.org/10.1145/3173574.3174225>
- [379] B. Rieder, A. Matamoros-Fernández, and Ò. Coromina, “From ranking algorithms to ‘ranking cultures’,” *Convergence: The International Journal of Research into New Media Technologies*, vol. 24, no. 1, pp. 50–68, January 2018. [Online]. Available: <https://doi.org/10.1177/1354856517736982>
- [380] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Misllove, and A. Rieke, “Discrimination through optimization,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–30, November 2019. [Online]. Available: <https://doi.org/10.1145/3359301>
- [381] A. Hannák, C. Wagner, D. Garcia, A. Misllove, M. Strohmaier, and C. Wilson, “Bias in online freelance marketplaces,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, February 2017. [Online]. Available: <https://doi.org/10.1145/2998181.2998327>
- [382] I. D. Raji and J. Buolamwini, “Actionable auditing,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, January 2019. [Online]. Available: <https://doi.org/10.1145/3306618.3314244>
- [383] “Guidance on the AI auditing framework,” UK Information Commissioner’s Office, Tech. Rep., 2020. [Online]. Available: <https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf>
- [384] “Regulation of the European Parliament and of the Council laying down harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts,” April 2021. [Online]. Available: <https://artificialintelligenceact.eu/the-act/>
- [385] L. Floridi, M. Holweg, M. Taddeo, J. A. Silva, J. Mökander, and Y. Wen, “capAI - a procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act,” *SSRN Electronic Journal*, 2022. [Online]. Available: <https://doi.org/10.2139/ssrn.4064091>
- [386] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mökander, and L. Floridi, “Ethics as a Service: A Pragmatic Operationalisation of AI Ethics,” *Minds and Machines*, vol. 31, no. 2, pp. 239–256, 2021. [Online]. Available: <https://doi.org/10.1007/s11023-021-09563-w>
- [387] J. Morley, J. Cowls, M. Taddeo, and L. Floridi, “Ethical guidelines for COVID-19 tracing apps,” *Nature*, vol. 582, no. 7810, pp. 29–31, May 2020. [Online]. Available: <https://doi.org/10.1038/d41586-020-01578-0>
- [388] S. Kacianka and A. Pretschner, “Designing accountable systems,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021. [Online]. Available: <https://doi.org/10.1145/3442188.3445905>
- [389] P. Chodron and E. H. Sell, *Comfortable with uncertainty: 108 teachings on cultivating fearlessness and compassion*. Boston, MA: Shambhala Publications, October 2008.
- [390] Contributors to Wikimedia projects. (2022, April) Kilowatt-hour - Wikipedia. [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Kilowatt-hour&oldid=1083539352> (Accessed: 2022-04-22)
- [391] ——. (2022, April) Joule - Wikipedia. [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Joule&oldid=1084025480> (Accessed: 2022-04-22)