Universität Siegen

Wirtschaftsinformatik

Fakultät III

zur Erlangung des akademischen Grades

Doktor rerum politicarum (Dr. rer. pol.)

Daily Encounters with AI: An Inquiry into Users' Sensemaking

M.A. Fatemeh Alizadeh

Erstgutachter	Gunnar Stevens Wirtschaftsinformatik Universität Siegen
Zweitgutachter	Markus Rohde Wirtschaftsinformatik und Neue Medien Universität Siegen
Dekan	*Marc Hassenzahl*

July 30, 2025

To my dear mom, who did not let the Cultural Revolution in Iran — a time when universities were shut down and many were expelled for their beliefs or opposition to the Islamic Republic — stop her. She was the first woman in her family to have a career, and she taught me that our dreams may be delayed, but they never truly die.

And to the fearless women of my country, who faced arrests, lifelong injuries, and even lost their sight during the Women, Life, Freedom revolution in Iran—you are the definition of resilience and hope.

Abstract

We live in times of rapid change, where unprecedented and unexpected events unfold with increasing speed and complexity, often disrupting familiar patterns and norms. In the midst of this uncertainty, AI plays a growing role, shaping how we interact, work, and live. To navigate such ambiguity, humans rely on sensemaking —a process of interpreting the unknown, settling on plausible explanations, and adapting their actions accordingly.

Despite AI's growing influence on nearly every aspect of life, studies repeatedly show that users often have a limited understanding of it, leading to misconceptions and unrealistic expectations. This lack of understanding not only results in frustration when AI systems fail to meet users' needs but also hinders effective interaction and collaboration with these technologies.

This dissertation views AI as a socio-technical umbrella term to explore how users make sense of it across three application domains: AI-assisted decision-making, AI-mediated social platforms, and agentic AI technologies. Guided by three research questions, it focuses on (1) how users make sense of AI in everyday encounters, (2) the empowerment needs that arise from these interactions, and (3) how design can support sensemaking and foster user agency.

Recognizing the importance of context and the situated nature of sensemaking, this work combines various qualitative methods, such as semi-structured interviews, role-playing workshops, and experience sampling.

Building on Weick's sensemaking framework, the findings reveal that sensemaking of AI shares familiar characteristics—it's enactive, driven by plausibility, inherently social, and often triggered by unexpected events. But AI sensemaking also has unique aspects, such as the influence of AI "folk concepts"—users' assumptions and expectations about AI shaped by cultural narratives and societal definitions. These perceptions are also influenced by human traits like trust and intelligence and vary depending on stakeholder roles. AI sensemaking not only guides users' actions but also informs counteractions—how people push back against or adapt to algorithmic systems.

To empower users, this research takes a holistic approach, exploring their needs along three dimensions: feeling, knowing, and doing. It highlights how users' sensemaking is influenced by both episodic power—arising from immediate, specific interactions—and systemic power, which operates at a structural level through institutional norms and opaque algorithmic designs.

For designing AI systems that better support users' sensemaking, the dissertation emphasizes two critical aspects. First, it frames sensemaking as an interactional element, suggesting that AI systems should foster continuous, context-specific engagement by helping users build competencies. Second, it stresses the importance of diversifying user participation in the design process. Involving users as co-creators and knowledge-makers empowers them to engage not just in sensemaking, but also in sense-unmaking, sense-giving, and

sense-breaking—creating a more collaborative and inclusive design process that addresses both individual needs and systemic challenges.

Finally, this dissertation argues for moving beyond Weick's retrospective sensemaking framework to apply the concept of *prospective sensemaking* in the context of AI. This approach emphasizes designing systems that not only help users make sense of past interactions but also enable them to anticipate, adapt to, and shape the uncertainties inherent in their ongoing and future engagements with AI technologies.

Acknowledgments

My PhD journey has probably been the most *uncertain* and unpredictable joinery I've ever taken on. There are many people I would like to thank for helping me navigate and embrace this uncertainty.

First and foremost, I want to thank my supervisor, Gunnar Stevens, for his huge trust in me — treating me like a PhD student while I was still doing my Master's and like a Postdoc during my PhD. You gave me the space to grow and learn at my own pace, and I know how rare and special that privilege is. For that, and for believing and investing in my work, Gunnar, I'm truly grateful to you.

I also want to thank my second supervisor, Markus Rode, for taking the time to read and evaluate my work despite his busy schedule. Markus, your support has meant so much to me, thanks for not letting me down.

A huge shoutout of gratitude and special thanks goes to Dave Randall and Peter Tolmie, my dearest mentors, who stood by my side throughout this journey, rain or shine. Words can't capture the enormity of my gratitude for you both. Your depth of knowledge, your magic with words and articulation—it's been nothing short of inspiring. You saved me more times than I can count, lifting me from the ashes every time I felt like I couldn't go on.

Dave, thank you for being so kind, knowledgeable, and endlessly witty. Thank you for teasing me so charmingly and for reminding me not to take life too seriously whenever I needed it. Your critical and intelligent perspective on almost every topic is really something special. You showed me how to find my way back to life through art and literature when academic life got tough. I've learned so much from you—not just about academia and research, but about life itself.

Peter, you're the reason I was finally able to find peace with my dissertation. Thank you for your constant feedback, for always reading and commenting on my work, for our thoughtful meetings, and for genuinely caring about me. You've been my go-to person for any advice I needed throughout my academic journey. I can't even write this section without tears in my eyes —thank you for everything.

To my amazing colleagues and PhD cohort at the University of Siegen—those who stayed, those who left, those from my group, and those from others who supported me: thank you. A special mention goes to Sima Amirkhani, Jana Krüger, Jenny Berkholz, Konstantin Aal, Timo Jakobi, Sofia Ganzalez, Dennis Lawo, and Margarita Esau-Held.

Sima, you were such an unforgettable part of my PhD journey—as a colleague, buddy, and friend. Thank you for always showing up at campus whenever I was there, for our coffee breaks, and for opening your home to me whenever I needed a place to stay overnight in Siegen. You were the one person I could always count on. Thank you for teaching me to laugh off the problems and always look ahead. I drew so much strength from your "let-it-go" attitude toward both research and everyday life. This journey would have been so lonely without you.

Jana, you have always been my greatest and most selfless support in Germany, offering help even before I asked for it. Thank you for following my work so closely, believing in me, and helping in every way you could. I'll never forget our time working together in the bookstore café, our late-night talks about research, and the constant messages encouraging me to submit my dissertation—even after you left us. You were there for me during the darkest times of my PhD, and I'll always miss having you as a colleague.

Jenny, thanks for always pulling me into coffee breaks whenever I was at campus and for setting up our dissertation writing day in Siegen. I loved your kind, quiet smiles from across the table during our group meetings—they always made things feel a little brighter.

Konstantin, you were always there to support me—listening, helping, and solving every problem imaginable from the very start of my Master's all the way through my PhD. Kosta, you truly deserve all the best in life, and I'm so grateful for everything you've done for me.

Timo, thank you for being the first person to open the door to the PhD world for me. You were my first co-author and supervisor when I was still a Master's student, and I'll always be grateful for the role you played in starting this journey.

I'd like to thank Marietta Krenzer Gräb, Nathanael Klein, Johanna Schirmacher, Merle Berg, and Akin Turan, the incredible department staff who handle all the organizational work. Thank you for always being so helpful and attentive to all my questions and for keeping everything running so smoothly.

Thank you to my colleagues and friends at TU Eindhoven for hosting me so kindly. Minha Lee, thank you for offering me a second home and for finding me a place to stay in Eindhoven, for letting me in your circle of friends, for our walks on TU/e campus with Hannah, and for taking care of me in ways I never expected. Thank you for visiting me in Siegen, for our journey to CUI, for seeing value in my research—and for continuing to collaborate with me even after I left. I'll never forget how you'd always look back at me and say, "up to you," whenever someone asked how long I was staying in Eindhoven. And how at home that made me feel. You are one of my greatest inspirations, and there's nothing more enjoyable than working with you.

I'd also like to thank my Eindhoven friends: Rucha Khot, Yvon Ruitenburg, Gizem Oktay, and especially Yuhui Xu, who was my buddy and closest friend in Eindhoven from the beginning to the end. Yuhui, I loved our hallway chats and missed you so much after I left – and I still do. Rucha and Yuhui, thank you for the collaborative analysis sessions; they were such a highlight. And to all of you, thank you for welcoming me into your group and investing your time and energy into my work.

A big thank you to Douglas Zytko and Motahhare Eslami—two incredible mentors and co-authors who not only shared their knowledge and expertise but also unofficially took me under their wing and guided me along the way.

Doug, I don't even know where to start with my gratitude for you. You are one of a kind: your depth of knowledge, your incredible talent for articulating ideas, and, most of all, your genuine humanity. You've been there for me in ways I'll never forget—from sending that "mental health check" email before my submission to cheering me on during my presentation, sitting in the front

row, taking photos, and sharing them online. And then, right after, asking if we'd collaborate again—it meant so much. Everything about you stands out, Doug. You've been a real highlight of my PhD journey, and I'm so grateful to have had the chance to work with you and learn from you.

Motahhare, I can't quite put into words what you've meant to me and my PhD journey. You took me under your wing like a mom and believed in me so early on—I'll always carry that with me. I'll never forget the time you were meeting with me about our paper while sitting in the emergency room. Who else would have done that? You pushed me forward when I needed it most, working tirelessly on our paper for years to get it published. I still have so much to learn from you, and I'm endlessly grateful for every moment I've spent with you. Thank you for everything—your golden heart and brilliant mind.

A heartfelt thank you to my colleagues at the Fraunhofer Institute, Dominik Pins and Dennis Paul, for making this journey so much more enjoyable. I'm grateful for the way you checked in on me and asked about the state of my dissertation—it always felt good to know you cared. Thanks for our meetings in Sankt Augustin, for the insider jokes that still make me smile, and for stepping up to handle the project work when I needed to focus on writing. You both made a tough process feel lighter, and I really appreciate it.

A big thank you to the Frames mentoring program at the University of Siegen, especially Juni Mohos and Rebecca Weber, for letting me be part of such an amazing program and for all the care and support you've shown me. During the tough times of the Women, Life, Freedom revolution in Iran, you recognized what I was going through and offered so much kindness—I can't tell you how much that meant to me.

And to Albrecht Schmidt, my wonderful mentor, thank you for generously giving your time to my work and for our regular meetings that kept me on track when things got tough. I've learned so much from you, and I'll always be grateful for your support and guidance.

I'd like to thank the funding institutions that made these studies and research projects possible: the Federal Ministry of Justice and Consumer Protection (BMJV), supported by the Parliament of the Federal Republic of Germany through the Federal Office for Agriculture and Food (BLE) under the innovation support program; the Federal Ministry of Education and Research (BMBF); and the European Union (NextGenerationEU).

I want to thank all the co-authors of the papers mentioned in this dissertation for their valuable contributions to my work: Aikatherini Mniestri, Philipp Wintersberger, Alexander Boden, and Volker Wulf. A special thanks to Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux at the Sorbonne HCI Lab—not just for your commitment to this work, but for welcoming me into this journey with you. Working with all of you has been an absolute privilege!

A heartfelt thank you to my study participants for placing their trust in me and for their patience throughout this journey. Your anonymous contributions—sharing your knowledge, experiences, and stories—gave purpose and meaning to this work. This work will forever carry a part of your generosity and insight.

Many parts of this dissertation were written at Onzième Lieu, a co-working space where I truly felt at home. I'm so grateful to the staff there for creating such a warm and supportive environment that made it easier to focus and

keep going. And Adam, your energy is absolutely contagious—you always brought me back to life after a long day of writing. Thank you for that!

A heartfelt and deeply special thank you to my most precious family: Faegheh (my mom), Reza (my dad), Mona (my sister), and Amir (my brother). You've been my rock, my safety net, and my greatest cheerleaders. From listening to my endless complaints during our evening video calls to lifting me up when I felt overwhelmed, you've carried me through this journey with your unconditional love and support. PhD or not, there's no journey in life I could ever take without you by my side. You've been my greatest source of strength for as long as I can remember, and I'm forever grateful for you.

I also want to thank Omma (my grandmother) and Morti (my uncle), who I sadly lost in the past two years. They were always so patient, always believing in me, and waiting for the day I would finally return once this chapter was complete. Though they aren't here to see it, I hope I've made them proud. I carry their love and memory with me every single day.

To everyone mentioned here and those I may have missed: thank you for being part of this journey. I'm so lucky to have had you by my side. I wish you all the best in both your academic and non-academic lives.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Siegen, July 30th, 2025

M.A. Fatemeh Alizadeh

Contents

Abstract vii

Acknowledgments xi

Declaration xix

Related Publications 3

List of Figures 5

List of Tables 7

1 Introduction and Motivation 9

1.1 Introduction 9

1.2 Research Aim and Approach 13

1.3 My Personal Encounters with AI 19

1.4 Dissertation Overview 21

I Background and Approach 25

2 Related Works 27

2.1 Definitions' of AI vs. Users' Perception 27

2.2 Human Sensemaking 31

2.2.1 Characteristics of Sensemaking 33

2.2.2 Researching Sensemaking 35
2.3 Users' Sensemaking of AI Systems 36
2.3.1 AI-assisted Decision Making 37
2.3.2 AI-Mediated Social Platforms 39
AI Mediation on Social Media 40
AI Meditation on Dating apps 44
2.3.3 Agentic AI 46
2.4 Sensemaking and User Empowerment 49
2.4.1 Empowerment in HCI 51
2.4.2 AI and User Empowerment 53
2.5 Designing for Sensemaking and Empowerment 55
2.5.1 Sensemaking and Interaction Design 56
2.5.2 Sensemaking and Data-enabled Design 60

3 Methodological Approach 63
3.1 Research Design 64
Design Case Studies 65
Co-Design and Research through Design 66
3.2 Methods and Activities 68
3.2.1 Interviews and Distributed Data Collection 68
Science Fiction Film Analysis 71
Application Review Analysis 71
Cognitive Walk-through and Thinking Aloud 72
Experience Sampling Method 73
Role-Playing 74

II How Users Make Sense Of AI 77

4 I Don't Know, Is AI Also Used in Airbags? 81

4.1 Abstract . . . 81
4.2 Introduction . . . 82
4.3 Research Background . . . 84
4.3.1 Public Acceptance and Awareness of AI . . . 85
4.3.2 Designing Explainable Technology . . . 86
4.3.3 Making Sense of Technology . . . 87
4.3.4 Content Curation Algorithms and Recommender Systems 88
4.3.5 Internet of Things (IoT) . . . 89
4.3.6 Robots . . . 90
4.3.7 Conversational Agents . . . 91
4.4 Research Setting . . . 93
4.5 Findings . . . 97
4.5.1 What is AI: A Thing, Characteristic, or Set of Instructions? 97
4.5.2 On the Competencies of AI: From Automation, to Simulation and Superpower . . . 98
4.5.3 On the Expectations of People: When Things Become AI 104
4.5.4 On the Impact of AI: If AI Disappears 110
4.6 Discussion . . . 114
4.6.1 Meaning Ascribed to AI: The Interconnected Folk Concepts . . . 114
4.6.2 Expectation Ascribed to AI: Design for the Unrealizable? 116
4.7 Limitations and Future Work . . . 118
4.8 Conclusion . . . 119

III AI-Assisted Decision Making **121**

5 Trust in AI-assisted Decision Making **125**

5.1 abstract 125

5.2 Introduction 126

5.3 Related Work 129

5.3.1 Background on AI-embedded Systems Assisting Decision Making 130

5.3.2 Human-AI Trust 131

5.3.3 Human-AI Trust and Stakeholders Other Than Users . 133

5.4 Methodology 136

5.4.1 Participants 137

5.4.2 Interview Protocol 139

5.4.3 Analysis of the Interviews 140

5.5 Findings 141

5.5.1 On the Definition of Trust 142

Positive expectations and perceived risk are prerequisites for the emergence of trust, but the nature of risk is debated. 142

Task complexity as a new prerequisite for the emergence of Human-AI trust. 144

Trust is differentiated from trust-related behaviors and trustworthiness. 144

5.5.2 The Role of Inter-personal Relations on Trust 145

The team behind AI plays an important role in (Human-AI) trust 146

The effect of AI certification on Human-AI trust depends on who is behind it. 148
5.5.3 Diverging Opinions on Three Factors impacting Human-AI trust 148
AI practitioners and decision subjects do not share the same view on the role of AI transparency on trust. 149
Interactivity: exploration tool for AI practitioners, means to be included in the loop for decision subjects 151
5.6 Discussion 152
5.6.1 On the Important Role of Inter-personal Relations on Trust Within the Socio-technical System 153
5.6.2 On the Prerequisites for the Existence of Trust 156
5.6.3 On the Notion of Agency over the System 158
5.6.4 Future Work Directions 161

IV AI-Mediated Social Platforms 165

6 Social Media 169
6.1 Abstract 169
6.2 Introduction 170
6.3 Blade Runner: A Closer Look at the Voight Kampff Test 172
6.4 Critique 176
6.4.1 Invisible AI: privacy, transparency, and fairness 177
6.4.2 The metric society 180
6.5 Limitations and Implications for HCI 184
6.6 Conclusion 186

7 Online Dating **189**

7.1 Abstract 189

7.2 Introduction 190

7.3 Related Work 193

7.3.1 Interpersonal Harm through Online Dating 193

7.3.2 Algorithmic Harm through Social Platforms and Users' Folk Theories 196

7.3.3 Online Dating Algorithms and their Impact on User Experience 198

7.4 Method 201

7.4.1 Stage 1: Analyzing Tinder App Reviews 201

Data collection 201

Data analysis 202

7.4.2 Stage 2: In-Depth Interviews 203

Sampling 204

Procedure and Analysis 205

7.5 Findings 206

7.5.1 The Perceived Algorithmic Harms to Intra- and Inter-Personal Relationships (RQ1) 208

Damaging Self-esteem: 208

Sabotaging Potential Relationships: 210

Encouraging Anti-Social Behaviour: 212

Identity-based harm: 215

7.5.2 The "Conflict-of-Interest Theory": Competing Interests and Clashing Values Between the Dating Algorithms and Users (RQ2) 217

Throttling Profile Visibility 218
Manipulating Users' Matches 218
Recommending Large Quantities of Profiles That Will Not Lead to Matches 219
7.5.3 Working With, Through, and Around the Tinder Dating Algorithms to Regain Agency (RQ3) 222
Resetting the algorithm with user account changes . . 222
Knowing Your Competitors 223
Engaging in Counter-Intuitive Behaviors to Disrupt Unfavorable Algorithmic Processes 223
Leveraging Location-based Filtering for Match Variety and Safety 224
Bypassing the Algorithm by Moving to other Platforms 226
7.6 Limitations and reflections 226
7.7 Discussion 228
7.7.1 The Conflict-of-Interest Theory, Its Implications, and Associated User Behavior 229
7.7.2 Synthesizing Perceived Algorithmic Harms in Online Dating with the Broader Algorithmic Harm Literature . 233
7.7.3 The Interplay Between Human-Inflicted and Algorithm-Facilitated Harm 236
7.7.4 Scrutinizing Perceived Algorithmic Harms and Related Algorithmic Processes 238
7.7.5 Implications for Designing Algorithmic Processes in Dating Apps 240
7.8 Future work and conclusion 242

7.9 Appendix 243

V Agentic AI Technologies 245

8 Alexa, We Need to Talk 249

8.1 Abstract 249

8.2 Introduction 250

8.3 Related Work 252

8.3.1 Adopting VAs in Daily Life and Private Spaces 253

8.3.2 Consumer Understanding of Data Collection and Processing 255

8.3.3 Designing for Sense-Making of Data 257

8.4 Data Visualization Prototype 260

8.5 Research Methodology 262

8.5.1 Sample 265

8.6 Results 267

8.6.1 Expectations on the Takeout Format 267

8.6.2 Exploring the Data via the Takeout 268

8.6.3 Exploring the Data via the Prototype 269

8.6.4 Supporting Sense-Making 273

8.7 Discussion 278

8.7.1 Awareness Creates Data Literacy 278

8.7.2 Data Literacy as a Skill that Concerns Everyone 280

8.7.3 Guidelines for VA Data Literacy 282

8.7.4 Limitations 285

8.8 Conclusion 285

9 Alexa Why Did You Do That? 287

9.1 Abstract 287

9.2 Introduction 288

9.3 Background 290

9.3.1 Errors in Interaction with VAs 291

9.3.2 Accountability of VAs 292

9.3.3 Designing for Accountable VAs 294

9.4 Method 295

9.4.1 Procedure 297

ESM: Testing Alexa's Features 297

Co-design workshops: Understanding Users' Expectations 298

Final interviews: Tracking the Changes 299

9.4.2 Analysis 299

9.5 Findings 300

9.5.1 Exploring VA's Accounts over Time (RQ1) 301

Decontextualized single responses lead to abonnement of the feature. 302

"I'm just learning" responses keep users temporary hopeful. 304

9.5.2 Users' Expectations from VA's Accountability Feature (RQ2) 305

Insights from the dialogues 306

Insights from the role-playing 308

9.6 Discussion 309

9.6.1 Explanatory Dialogues as Species of Accounts 310

9.6.2 Design Implications: Towards Contextual Sense-making and Collaborative Repair . . . 313
A checklist for designing accountability through explanatory dialogues. . . . 314
From shared accountability to shared agency. . . . 315
9.6.3 Limitations and Future work . . . 317
9.7 Conclusion . . . 318
9.8 Appendices . . . 320
9.8.1 Appendix 1 . . . 320
9.8.2 Appendix 2 . . . 321

VI Research Outcome **323**

10 Discussion and Implications **325**
10.1 Sensemaking in Everyday life Encounters with AI (RQ1) . . . 326
10.1.1 Sensemaking of AI through Meaning Ascription, and Definitions . . . 328
10.1.2 Sensemaking of AI and Enactment . . . 336
10.2 Sensemaking of AI and User Empowerment (RQ2) . . . 341
10.2.1 Empowering Users by Addressing Emotional Needs . . . 342
10.2.2 Empowering Users Through Knowledge and Understanding . . . 344
10.2.3 Empowering Users in their Actions . . . 345
10.3 Designing for Sensemaking and User Empowerment (RQ3) . . . 347
10.3.1 Framing AI Sensemaking as an Interactional Element in Practice . . . 348
10.3.2 Diversifying Participation for Empowered Sensemaking 350

10.4 Dealing with Uncertainty: Prospective AI Sensemaking (RQ3) 353

11 Conclusion **357**

11.1 Summary of the Thesis 357

11.2 Contributions 358

11.2.1 Sensemaking of AI 359

11.2.2 User Empowerment 361

11.2.3 Designing for Sensemaking: Main Contribution to HCI 363

11.3 Reflections and Closing Thoughts 365

Bibliography **369**

Related Publications

The following published papers were used in each chapter.

Chapter 4

Fatemeh Alizadeh, Gunnar Stevens, Margarita Esau (2021). I Don't Know, Is AI Also Used in Airbags? An Empirical Study of Folk Concepts and People's Expectations of Current and Future Artificial Intelligence. I-com Journal, 20(1), 3-17.

Chapter 5

Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, Baptiste Caramiaux (2024). Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '1

Chapter 6

Alizadeh, F., Mniestri, A., & Stevens, G. (2022, October). Does Anyone Dream of Invisible AI? A Critique of the Making Invisible of AI Policing. In Nordic Human-Computer Interaction Conference.

Chapter 7

Fatemeh Alizadeh, Dennis Lawo, Gunnar Stevens, Douglas Zytko, Motah-

hare Eslami (2024) When the "Matchmaker" Does Not Have Your Interest at Heart: Perceived Algorithmic Harms, Folk Theories, and Users' Counter-Strategies on Tinder. Computer Supported Cooperative Work (CSCW)

Chapter 8

Dominik Pins, Timo Jakobi, Alexander Boden, Fatemeh Alizadeh, Volker Wulf, (2021, June). Alexa, we need to talk: A data literacy approach on voice assistants. In Designing Interactive Systems Conference 2021 (pp. 495-507).

Chapter 9

Fatemeh Alizadeh, Peter Tolmie, Minha Lee, Philip Wintersberger, Dominik Pins, Gunnar Stevens (2024, June). Alexa, why did you do that? A data literacy approach on voice assistants. In Conversational User Interfaces 2024.

It's important to clarify my use of "I" and "me" throughout this thesis. While this dissertation reflects my personal journey, the studies presented here are the result of collaborative efforts. The research involved contributions from academic colleagues, co-authors, and research participants, all of whom played an essential role in shaping this work. However, as I reflect on and write about this journey, it feels more appropriate to use "I" and "me" to capture my own voice, experiences, and perspectives while acknowledging that the work itself is inherently collaborative.

List of Figures

1.1 There is an ongoing struggle to define AI. 10

2.1 Google's No CAPTCHA reCAPTCHA test. 42

2.2 The notifications Instagram users receive when they exhibit bot-like behavior. 44

2.3 The announcement from the Tinder website explaining that the matching algorithm has abandoned the Elo Score. 45

5.1 Schematic representation of the extent to which trust between different stakeholders groups discussed in relation to how it can affect Human-AI trust in the context of decision making. 146

6.1 A sketch of the Voight Kampff (V-K) test in the original Blade Runner movie, focusing on the biometric information collected by the V-K machine during the test 169

7.1 An overview of the perceived algorithmic harm, perceived algorithmic processes, users' folk theory, and users' counteracts. . . 207

8.1 Screenshots of our prototype for supporting consumers in understanding their VA data 261

8.2 Settings and functionalities of the prototype for visualizing the data. 264

8.3 List with identified terms out from the interactions that refer to unknown voice commands 271
8.4 A visualization of VA data 277

9.1 Overview of the process. 296
9.2 Eight scenarios and five dialogues generated for each during Workshop 1 301
9.3 The email that Alexa users in Germany received from Amazon on February 2, 2021, after the new feature had been released (note that the original email was in German). 320

11.1 Dimensions of empowerment in AI sensemaking 363

List of Tables

3.1 Research activities according to section and publication of the thesis 69

4.1 Folk concepts of the meaning of AI and the identified categories. 99

4.2 Folk concepts of an AI-enabled objects and the identified categories. 105

5.1 The different stakeholders related to the AI assisted decision making systems. This article focuses on AI practitioners and Decision subjects, two stakeholders who received less attention in the Human-AI trust literature. 131

5.2 Characterization of AI practitioners, their companies, and AI they work with as reported by the interviewees themselves. “Small” refers to the companies with less than 20 employees, “Large” - with over 1000 employees. Explanation for abbreviations: *XAI* - explainable AI, *R&D* - research and development, *mgr.* - manager, *CEO* - chief executive officer, *CPO* - chief product officer, *CS* - computer science, *eng.* - engineering, *CNNs* - convolutional neural networks, *OR* - operations research, *ML* - machine learning. 135

5.3 Characterization of decision subjects, notably their background and in what context they received a Human-AI decision. 137

5.4 Structure and examples of questions per each group of participants. Data analysis of this paper mostly relies on answers around the definitions and factors of Human-AI trust. A full list of questions is in Appendices A and B. 138

7.1 An overview of the participants' demographics 243

8.1 Participant information of our study 266

9.1 User-designed dialogues in which the VA accounts for users' feelings. 306

9.2 User-designed dialogues in which the VA possesses situational awareness. 307

9.3 The progressivity, as well as co-performance and shared accountability in the user-designed dialogues. 307

9.4 Explanatory dialogue checklist. 315

9.5 Overview of the participants and their involvement in the project 321

11.1 Expanding user roles in AI sensemaking 365

1

Introduction and Motivation

1.1 Introduction

> *Uncertainty is the only certainty there is, and knowing how to live with insecurity is the only security.* – John Allen Paulos

> *We grow accustomed to the Dark –*
> *When light is put away –*
> *As when the Neighbor holds the Lamp*
> *To witness her Goodbye –*
> *A Moment – We uncertain step*
> *For newness of the night –*
> *Then – fit our Vision to the Dark –*
> *And meet the Road – erect* – Emily Dickinson

Starting my first master's degree in artificial intelligence (AI) nearly ten years ago, I often encountered the question from others about what exactly AI was. I found this question puzzling, not only because there was no simple way for me to provide a definition that was accessible to lay people yet reflective of my professional engagement with the field, but also because AI did not represent a specific technology or approach I could easily refer to. Although

the field of AI dates back to 1956 (McCarthy, 2007), the term has actually never represented a specific type of technology (Solomonoff, 1985). For example, during my studies at that time, AI was mostly tied to concepts like computer vision and image recognition. But, this representation was evolving towards expert systems (which became the focus of my first master's thesis).

Later, when I was studying in Germany, AI was rather associated with machine learning and the broader field of data science. And today, as I write this dissertation, the spotlight is on generative and large language models (LLMs). Having said this, I witnessed the evolution of what the term "AI" represented throughout my personal and professional experiences and continued to struggle with defining it for those who asked. That's why Figure 1.1 (Elliott, 2017) has consistently accompanied my PhD presentations. It captures the ongoing struggle to define AI, underscoring that this quest is not merely my personal challenge but a broader one among AI practitioners.

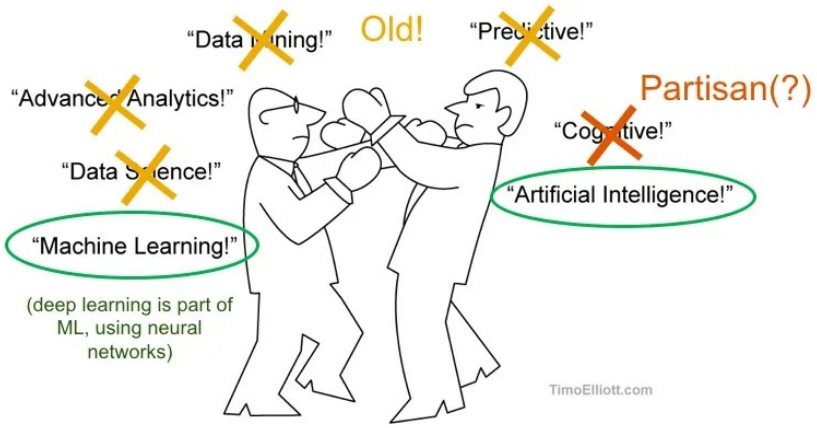


Fig. 1.1.: There is an ongoing struggle to define AI.

When I began my PhD in human-AI interaction (HCI), I was struck by Schank's study which revealed that researchers at the Center for the Learning Sciences

at Carnegie Mellon University were frequently asked by users, “Where is the AI?”. These users were interacting with AI-embedded prototypes for higher education programs but did not realize it (Schank, 1991). What struck me was not just the question itself but that almost three decades later the study still resonated with what I had witnessed. So, why does the question “Where is the AI?” remain so relevant decades later? As I delved into further studies, I kept wondering: Why do subsequent studies continue to show that many users still have a limited understanding of AI, leading to ongoing misconceptions and unrealistic expectations? (e.g., Boden, 2004; Long and Magerko, 2020). This limited understanding not only leads to frustration due to user unmet desires, but also prevents effective interaction and collaboration with AI (Fast and Horvitz, 2017; Long and Magerko, 2020).

In today’s landscape, where general-purpose AI models like ChatGPT [1] and MidJourney [2] are transforming every aspect of our daily lives, the issues of limited understanding, misconceptions, and incorrect expectations surrounding AI are more relevant than ever. Several nations have proposed regulations and AI acts aimed at managing AI’s implications and promoting user empowerment. The European Union has been at the forefront of these efforts, as evidenced by the proposed AI Act, which emphasizes the need for clear guidelines and active oversight by institutions such as the European Commission and the European Parliament to ensure that AI’s integration into society is managed responsibly (the Future of Life Institute (FLI), 2024).

[1] ChatGPT, developed by OpenAI, is a conversational AI model designed to understand and generate human-like text based on user inputs (OpenAI, 2022).

[2] MidJourney is a generative AI platform that creates high-quality images based on textual descriptions, often used for artistic and creative purposes (MidJourney, 2022).

To clarify, in this dissertation I treat AI as a 'sociotechnical construct' (Elliott, 2017) an umbrella term that encompasses both technological capabilities and phenomenological dimensions. For instance, a smartphone camera's autofocus feature might be perceived as AI because it leverages machine learning models to improve accuracy. On the other hand, systems like ELIZA [3], although technically quite simple, are often perceived as AI because they attempt to mimic human behavior. This broad and somewhat ambiguous definition is echoed in the European Commission's attempt to define AI as systems capable of "perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge or processing the information derived from this data, and deciding the best actions to take to achieve a given goal" (Samoili, Lopez, et al., 2020).

By considering AI as an umbrella term rather than focusing on one specific application domain, this dissertation captures the evolving nature of what AI is associated with. This approach reflects the shifting spotlight on AI's representation. I have followed this shifting spotlight in my research, pushing the boundaries of my encounters with AI as a user, and using it as a lens to observe how different systems become prominent in our daily lives under this umbrella term. Some of the systems I look into include health care systems, dating apps, and voice assistants (VAs), which are further discussed in Subsection 1.3.

My approach has arisen from a desire to deal with the feeling of uncertainty, much like a traveler trying to find their way through a fog. At its heart is the

[3] ELIZA is an early natural language processing program developed by Joseph Weizenbaum. It simulates conversation by using pattern matching and scripted responses, often cited as an example of a system perceived as intelligent despite its simplicity (Weizenbaum, 1983).

challenge of making sense of the unknown. We don't really know how AI will unfold before us—perhaps as unexpectedly as a magician pulling a bouquet of flowers or a rabbit from a hat—and we must deal with the consequences. But how can we step onto the stage and move beyond being mere spectators of this show? Can we work with this uncertainty? Can we learn from it for the unforeseen future with AI?

The framework and methodology outlined in this dissertation strive to navigate through this uncertainty, holding up multiple possibilities with the aim of connecting these possibilities to meet the challenges of our present and unforeseeable future with AI.

1.2 Research Aim and Approach

This dissertation lies at the intersection of AI and HCI. During my Ph.D. years, several disciplines have emerged at this intersection, including Human-Centered AI (HCAI) (Shneiderman, 2022), Humane AI (Crowley, OrSullivan, et al., 2019), User-Centered AI (UCAI) (Augustein, Buschek, et al., 2020), and Explainable AI (XAI) (Hagras, 2018) among others. Despite slight differences, they centrally address the question of how to empower users and maintain their control as AI technologies evolve. As Schneiderman puts it "HCAI is a framework that bridges the gap between ethics and practice with specific recommendations for making successful technologies that augment, amplify, empower, and enhance humans rather than replace them" (Shneiderman, 2022).

The current literature identifies several HCAI objectives, including transparent (Zhou, Li, et al., 2022), explainable (Ehsan and Riedl, 2020), ethical

(Hagendorff, 2020), fair (Feuerriegel, Dolata, and Schwabe, 2020), trustworthy (Wickramasinghe, Marino, et al., 2020), responsible (Ghallab, 2019), and sustainable AI (Van Wynsberghe, 2021). These objectives are frequently overlapping and interconnected; achieving one can sometimes necessitate compromises with another (W. Xu, Gao, and Dainoff, 2023). For example, providing transparency about AI-assisted decision-making in grading systems has been shown to diminish students' trust in these systems (Kizilcec, 2016). Therefore, research suggests that designing AI systems often requires a careful balancing act to fulfill as many of these goals as possible (W. Xu, Gao, and Dainoff, 2023).

In this dissertation, I address some of these goals (such as trust and transparency) through the lens of sensemaking. My main aim is to understand how AI is made sense of in daily life and how we can empower users in their encounters with AI. Therefore, I aim to explore:

1. RQ1) How do users make sense of AI in their everyday life encounters?
2. RQ2) What are the needs for empowerment that emerge through users' everyday encounters with AI?
3. RQ3) In what ways can we apply design to empower users in their everyday life encounters with AI?

What mattered most to me was to capture and deeply understand the context of users' sensemaking of AI. When I talk about sensemaking, it is also important to mention what I mean by sensemaking and what contexts I am referring to.

Karl Weick, known as the "father of sensemaking," defines the term as simply "the making of sense" (Weick and Weick, 1995, p. 4). It involves "structur-

ing the unknown" (Waterman, 1990, p. 41) by organizing stimuli into a framework that allows us to "comprehend, understand, explain, attribute, extrapolate, and predict" (Starbuck and Milliken, 1988, p. 51). Beyond this, sensemaking encompasses the interaction of information seeking, meaning ascription, and action (Thomas, Clark, and Gioia, 1993). Additionally, sensemaking mechanisms, as discussed by (Feldman, 1994), involve cultural standards and rules for perceiving, interpreting, believing, and acting within a given context. This process is particularly visible when our expectations are disrupted, prompting us to create explanations for these discrepancies. It helps us transform the complex, ongoing world into a "situation that is comprehended explicitly in words and that serves as a springboard into action" (Weick, Sutcliffe, and Obstfeld, 2005, p. 409). Essentially, sensemaking requires us to articulate the unknown because trying to explain it can reveal the extent of our understanding.

It's important to recognize that sensemaking doesn't aim to find a single "correct" answer. Instead, it's about developing an evolving picture that becomes more detailed through data collection, action, experience, and conversation (Ancona, 2012). This approach helps us act when our usual understanding of the world shifts (Weick, Sutcliffe, and Obstfeld, 2005). Hence, sensemaking provides a sense of stability that helps keep fear at bay.

AI is one of the most significant forces reshaping our world today. As a vague, evolving, and often unclear concept, AI simultaneously dominates societal discourse and becomes increasingly embedded in everyday life through so-called *AI technologies*. This creates a need for people to make sense of AI on two interconnected levels: first, by understanding the specific systems or tech-

nologies they encounter that are presented as AI, and second, by engaging with the broader idea of AI that appears in societal discussions and personal exchanges. These two levels of sensemaking are deeply connected—they shape and reinforce each other, blending the concrete experiences of AI with its more abstract interpretations.

This dissertation addresses AI sensemaking through various approaches, including the lens of symbolic interactionism, by examining the meanings people ascribe to AI. Symbolic interactionists emphasize the methods and practices people use to make sense of their world, considering the context-specific and reflexive nature of social actions. As explained by Blumer, "people act towards things (such as physical objects, people as well as abstract ideas) on the basis of meanings they ascribe to them." (Blumer, 1986, p. 2). I refer to these meanings as folk concepts, as this term expresses the folk-conceptualization of the subject matter.

Another way I approach sensemaking is through a social and practice-oriented lens, focusing on the appropriation of technologies as a social process. This involves incorporating artifacts into one's life (Draxler, Stevens, et al., 2012; Stevens, Pipek, and Wulf, 2009), influencing behavior and evoking new practices, thought patterns, and design approaches through reciprocal effects (Rohde, Brödner, et al., 2017). This practice-oriented focus is informed by the Information Systems department at the University of Siegen and their school of thought. The Siegen school emphasizes interaction and practice-oriented theory, particularly through the concept of living labs. Living labs are practical approaches for ICT design, focusing on long-term co-creation within real-life environments. They provide sustained access to application fields, allowing for continuous involvement of users and stakeholders, and

fostering deep engagement with everyday practices (Ogonowski, 2021). This approach not only aids in understanding and evolving technology use but also promotes emancipatory actions and organizational change through collaborative efforts (Ahmadi, 2023).

In pursuing the aims of my PhD, I also encountered research trajectories that led me to various university departments, including the HCI Lab at Sorbonne University in Paris, France (HCI Lab, 2021), the HCI Institute at Carnegie Mellon University (CMU) in Pittsburgh, USA (Human-Computer Interaction Institute, 2021), and the Industrial Design (ID) department at the Eindhoven University of Technology in the Netherlands (Eindhoven University, 2024). This broadened my approach to understanding context in the sensemaking process and incorporating multi-stakeholder perspectives, folk theorization, and data-enabled methods as research approaches.

The HCI Lab at Sorbonne University focuses on designing interactive systems that facilitate sensemaking through user-centered methodologies and participatory design (HCI Lab, 2021). Their work emphasizes the importance of context and the social dimensions of technology use (e.g., Caramiaux and Fdili Alaoui, 2022; Mohammadzadeh, Françoise, et al., 2024; Teyssier, Bailly, et al., 2017). This approach has allowed me to broaden my exploration of meaning ascription by going beyond the direct users of AI-enabled systems. I adopted a multi-stakeholder approach, incorporating the perspectives of both those involved in the development of AI systems and those affected by their use.

At Carnegie Mellon University's HCI Institute, research includes exploring how people develop folk theories about technology and AI (e.g., Eslami, Karahalios, et al., 2016; Eslami, Rickman, et al., 2015; Karizat, Delmonaco, et

al., 2021). Folk theories refer to the informal, often simplistic understandings that users form about how technology works (DeVito, Gergle, and Birnholtz, 2017). These theories influence how users interact with technology and make sense of its behavior. Collaborating with the institute allowed me to learn how to apply folk theorization as a lens to capture users' sensemaking, particularly through their interpretations of algorithmic processes and their coping strategies on dating platforms.

The most influential experience was, perhaps, my research visit to the Eindhoven University of Technology. Engaging with the design scholars there broadened my perspective beyond the social aspects of sensemaking to include moral dimensions. This engagement led to exploring how AI can serve as interactive mirrors, reflecting our interactions and identities back to us, and influencing our self-perception (M. Lee, 2021). In this context, I approach sensemaking through research through design (RtD), which entails "investigative experiments of complex matters of human relationships with technologies in the intimate contexts of everyday life, such as in the home" (Reddy, 2020, p. 105). Specifically, I discuss cases in which users of voice assistants engage in explorative experiments with the data collected by these devices in their home settings to make sense of their own interactions and usage behavior. Later, I combine this with the broader discourse on datafication and how emerging agentic technologies might change our sensemaking of ourselves and our relationships with others, especially in intimate settings like the home.

Therefore, the approaches in this dissertation are shaped by the experiences and knowledge I gained from these diverse learning environments and the

academic networks I engaged with. I hope they not only accurately reflect my experience but also convey the depth and breadth of my learning journey.

1.3 My Personal Encounters with AI

As mentioned earlier, my research does not focus on a single AI application domain. This is because, from the perspective of everyday users (myself included), AI has never represented just one specific technology. My choices were shaped by two key factors: first, my personal experiences as an AI user, and second, my involvement in research projects and collaborations as an AI researcher. Broadly speaking, I explored three main application domains: 1) AI-assisted decision-making systems, 2) AI-mediated social platforms, and 3) agentic AI technologies. Below, I provide some context to clarify my positionality and the rationale behind these choices.

1. **AI-Assisted Decision-Making Systems**: Living in Germany, I encountered various instances of AI-assisted decision-making in my everyday life, from background checks for renting a residence to securing a mobile contract. These decisions were often presented to me as automated outcomes based on the information the system had about me. As a result, I consider myself a "decision subject"—a person directly affected by AI decision-making. As a migrant, one question that naturally stood out to me was: *How can I make sense of these decisions, and to what extent can I trust them to ensure I'm not being discriminated against because of my background?* This question became a central focus during my collaboration with the HCI Sorbonne lab and is explored in the study presented in *Part III*. This part of the work looks at AI-assisted

decision-making from two perspectives: the people affected by these systems and those involved in designing them.

2. **AI-Mediated Social Platforms**: The second application domain I explore is AI-mediated social platforms, a focus that emerged during the COVID-19 pandemic, when much of our social interactions—especially with strangers—shifted to online platforms. During this time, I experienced firsthand, and observed within my social circle, the harm these AI-mediated platforms can cause. Building on these personal experiences, I examined users' empowerment needs on Instagram and collaborated with researchers from Carnegie Mellon University and the University of Michigan-Flint to explore how people make sense of the potential harms caused by algorithmic mediation in relationships and how they navigate these challenges on the Tinder dating platform. (Part IV).

3. **Agentic AI Technologies**: The final stage of my PhD was significantly shaped by the emergence of large language models (LLMs) as agentic AI technologies, which began to play a prominent role in various aspects of our everyday lives. My involvement in two research projects, focusing on the sensemaking of data collected by home assistants, allowed me to explore VAs as a specific type of agentic AI within the home—where our private lives unfold. I investigated how households make sense of the data collected by these agents and how they perceive the accountability of these systems when things go wrong (Part V).

Drawing from my personal encounters with AI across different application domains not only accounts for the evolving concept of AI—from mediator to assistant and even to more autonomous agents—but also underscores what

AI has come to represent during my PhD years. This experience positioned me as an 'insider' in the field, offering several advantages that are well-documented in research on studying familiar settings. For example, being an insider facilitated easier access to the field and a deeper understanding of the processes involved. I could better grasp the nuanced reactions of participants who shared similar experiences (De Tona et al., 2006). My familiarity with the application domains and shared experiences made it easier to recruit participants, increasing both their comfort and mine, and helping to build rapport more quickly (Kacen and Chaitin, 2006). My personal experience, combined with a certain level of knowledge about the subject, allowed me to address specific topics more effectively, such as users' expectations or the challenges they face (Berger, 2015). It also made me more attuned to issues that might not have been immediately obvious, ensuring that important topics were not overlooked.

1.4 Dissertation Overview

This thesis is organized into six main parts, spanning eleven chapters. *Part I* sets the stage with two chapters that position my work and provide the necessary background. Chapter 2 reviews existing literature on AI definitions, frameworks for sensemaking, and sensemaking of AI across the application domains addressed in this dissertation. It also covers concepts of user empowerment and approaches to designing for users' sensemaking and user empowerment. Chapter 3 outlines my methodological approach, focusing on design case studies, co-design, and RtD.

Part II, *part III*, *Part IV*, and *Part V* present key conference articles that explore different case studies on sensemaking from various perspectives discussed earlier.

I start with, *Part II* where I lay the groundwork using symbolic interactionism to understand how users perceive and make sense of the term "AI." From there, I transition into a more focused examination of *"Human-AI Trust"* within AI-assisted decision-making systems *part III*. These two sections highlight sensemaking as it relates to rules for perception, assigning meaning, and constructing definitions.

Next, in *Part IV*, capture sensemaking through users' *folk theories* and *counter-strategies* in response to algorithmic harm and bias. Here, I focus on two specific contexts: social media platforms (Chapter 6), and dating apps (Chapter 7).

Finally, *Part V*, explores sensemaking within agent technologies—specifically voice assistants—through processes of data and information seeking. This section is divided into two chapters set in the *home context*. The first chapter (Chapter 8) examines the data collected by VAs in domestic spaces, while the second (Chapter 9) focuses on sensemaking through users' direct dialogues with these systems.

Part VI presents the outcomes of my research across two chapters. I begin Chapter 10 with a cross-comparison of Weick's framework against my findings across various application domains (Section 10.1). I then discuss users' need for *empowerment* along three key dimensions: feeling, knowing, and doing (Section 10.2). This leads into Section 10.3 where I present the central message of this dissertation:

Users' sensemaking of AI systems is not static—it is an active, continuous process that involves ongoing learning, adaptation, and engagement. Through repeated interactions and routinized behaviors, users interpret, construct, and reshape their understanding of AI systems. Sensemaking, therefore, should be seen as interactional. Users are not passive recipients of AI outputs; they are dynamic actors—sensemakers, sense-unmakers, sense-givers, and even sense-breakers—who actively shape their relationship with these systems.

In the final chapter, I provide a summary of the thesis (Section 11.1), highlight its main contributions (Section 11.2), and share some final reflections (Section 11.3).

I hope that this work will inspire future research and practical applications that prioritize user empowerment and effective sensemaking in the evolving landscape of AI.

Part I

Background and Approach

2 Related Works

Over the course of the work reflected in this thesis, I have needed to draw upon a number of bodies of literature. These incorporate (1) definitions and narratives around "AI" versus users' perception of it, (2) Different frameworks and approaches to sensemaking (3) how users make sense of and understand AI within the three application domains explored in this work, (3) various perspectives on user empowerment, and (5) current strategies for designing systems that support users' sensemaking of AI. In this chapter, I'll dive into these areas of research to lay the groundwork for the contributions of this thesis.

2.1 Definitions' of AI vs. Users' Perception

Since its advent, AI has been defined in various ways, reflecting its complexity and impact across different domains. To position my research in making sense of AI as it is perceived I first explore various definitions from influential figures in the field and emphasize the challenge in attributing meaning to AI due to the lack of consensus in its definition.

Research on AI began in the mid-20th century, with Alan Turing's foundational ideas, particularly his 1950 "Turing Test," which proposed that a machine could be considered intelligent if a human interacting with it

couldn't distinguish it from another human. Though criticized, Turing's concept has significantly shaped AI development, influencing the creation of machines capable of mimicking human intelligence (Oppy and Dowe, 2003). In 1956 McCarthy coined the term "Artificial intelligence" as "*the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable*." (McCarthy et al., 2007). Similarly, Marvin Minsky defined "AI" as "*the science of making machines do things that would require intelligence if done by men.*" (Stonier, 1992).

While these definitions are relevant to the broader scientific understanding of AI, they do not apply to the narrow or weak AI systems that users interact with in their daily lives, such as Siri and ChatGPT. These systems are designed for specific tasks, targeting a limited subset of cognitive abilities, and cannot operate beyond their pre-defined limitations (Filina, 2024).

From a technical standpoint, AI has been defined as a branch of computer science focused on creating algorithms that emulate intelligent human behavior, including machine learning and deep learning (Bahl, 2020). Nilsson offers a broader view, describing AI as, "*AI in a broad sense is the intelligent behavior of artificially created objects, which, in turn, includes perception, reasoning, interaction and action in complex environments.*" (Nilsson, 1971). Similarly, the European Union (EU) defines AI systems as capable of "*perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge or processing the information derived from this data, and deciding the best actions to take to achieve a given goal.*" (Samoili, Lopez, et al., 2020).

While these definitions stem from different sources, they touch on similar concepts, such as perception, reasoning, and decision-making. However, to fully grasp what AI is based on these definitions, one must have a foundational understanding of concepts such as *data acquisitions* and *machine reasoning* which are often inaccessible to users without a technical background.

Apart from scientific and technical definitions of AI, popular media, such as films and literature, contribute to these narratives by drawing parallels between AI and human cognition, reinforcing myths about AI's potential to replicate or surpass human intelligence (March–Russell, 2020). One prominent narrative is the transhumanist apocalyptic view, which envisions AI as leading to an Artificial General Intelligence that could either liberate humanity from physical limitations or render it obsolete (Cohen, 2024). Cultural narratives also shape perceptions of AI, often portraying it as a double-edged sword that can both enhance and threaten human capabilities (Gáti, 2023).

Studies have shown that such narratives also influence public perceptions of AI. For example, Nader et al. examined how AI is depicted in entertainment media, influencing public perceptions of AI, and found those who believe AI is realistically depicted in entertainment media were more likely to see AIs as potential emotional partners or apocalyptic robots than to imagine AIs taking over jobs or operating as surveillance tools (Nader, Toprac, et al., 2024). Schwarz and Unselt, analyzed how AI is framed in YouTube videos, focusing on societal threat and efficacy. The study revealed that these narratives shape public understanding by highlighting certain aspects of AI, such as efficiency and innovation, while downplaying potential risks and impact in society (Schwarz and Unselt, 2024).

The lack of a clear, accessible definition of AI for the general public, combined with media-driven narratives that often diverge from the realities of AI, presents an important challenge to how users make sense of AI in their everyday lives. Several studies have shown that even today, many AI users do not realize they are interacting with AI systems (Davies, 2020; Eslami, Rickman, et al., 2015; Eslami, Vaccaro, et al., 2019; Long and Magerko, 2020; Pegasystems, 2019). For example, a HubSpot study involving over 1,400 consumers worldwide found that 63% of those who claimed not to use AI tools were, in fact, unknowingly using AI technologies, as identified by experts (Santiago, 2023). Similarly, Pegasystems confirmed these findings in a study of 5,000 consumers, where only 34% acknowledged having used AI technology, while 84% were found to have used such technologies according to the study authors (Pegasystems, 2019). This highlights a clear discrepancy between the understanding of AI among laypeople and experts.

Davies argued that public awareness of AI depends on the visibility of its application (Davies, 2020). In line with this, researchers at the Center for the Governance of AI at the University of Oxford found that most users believed that virtual assistants, smart speakers, driverless cars, social robots, and autonomous drones use AI, but did not recognize that Facebook photo tagging, Google Search, Netflix or Amazon recommendations, or Google Translate also rely on technologies classified as AI in the literature (B. Zhang and Dafoe, 2019). Similarly, results from a Northstar survey of 3,804 consumers in 2020 indicated a significant difference in recognition between "invisible AI" (i.e., AI algorithms behind the scenes) and "visible AI" (i.e., tangible AI devices) (Davies, 2020). Here, 90% of respondents knew that voice assistants are

AI-enabled, but almost one-third did not consider social media to be AI-based technology.

The challenges highlighted above form the foundation for this dissertation, which aims to explore how users make sense of AI in their daily encounters and identify their needs for empowerment, especially given how users can be disadvantaged by algorithmic biases, lack of transparency, and concerns around fairness and trust (see Section 2.4.2).

2.2 Human Sensemaking

The field of sensemaking has developed into a rich, multidisciplinary area of study, drawing from and contributing to several distinct disciplines. For example, in information science, sensemaking has been recognized as a catalyst for shifting the research focus from system-centered to user-centered phenomena (Turner, Allen, et al., 2023). In communication studies, it has reoriented the focus from simple sender-receiver transactions to the complex constructions shaped by time, place, and perspective. As Naumer et al.,put it:

> "The sense-making moment is the point in time-space when a person experiences a gap while moving through time-space. The situation and outcome, as experienced, are informed by the nature of the situation, its history, its constraints, its relevant external power structures and other situational, contextual, and personal factors." (Naumer, Fisher, and Dervin, 2008, p. 3)

However, the interdisciplinary expansion of sensemaking has also led to challenges, particularly in defining its constructs and propositions clearly. As Maitlis and Christianson noted: "The last decade has seen something of a proliferation of sensemaking-related constructs, which are not always clearly defined" (Maitlis and Christianson, 2014, p. 108).

While there is no universally accepted definition of sensemaking in the literature, it generally assumes that humans exist in an ever-changing reality, requiring a continuous process of bridging information gaps to guide our actions (Turner, Allen, et al., 2023). Naumer et al. further describe this gap-bridging process as:

> "The person bridges this gap by experiencing questions and muddles that lead them to construct bridges consisting of ideas, thoughts, emotions, feelings, hunches, and memories. Sometimes these 'bridges' are repetitions from the past; sometimes they are entirely new; sometimes they are deliberate and planned; sometimes capricious; sometimes unconscious at the time of action but brought to consciousness in interviewing talk; sometimes tactic and unarticulated but alluded to in examples and stories." (Naumer, Fisher, and Dervin, 2008, p. 3)

Sensemaking has two primary ontological roots: individual and social (Turing, 2009). At the individual level, it is seen as a personal cognitive process involving frameworks like mental maps and interpretive schemes (Maitlis and Christianson, 2014), while the social level emphasizes the creation of reality through communication, such as conversations and storytelling (Heath and Johansen, 2018). These two levels are interdependent, with social interactions informing individual understanding and vice versa.

As a field of practice, sensemaking aims to create meaningful, context-specific practices and frameworks, moving away from the traditional reliance on "expertise imposed on users" (Naumer, Fisher, and Dervin, 2008). This approach involves agents using sensemaking methods to navigate their environments, balancing structure and agency to develop relevant frameworks for their specific context (Turner, Allen, et al., 2023). In the literature, this is represented as "[humans] users as theorists and knowledge-makers in their worlds" (Dervin, 2015; Naumer, Fisher, and Dervin, 2008). Sensemaking equips individuals with the tools and methods needed to act as creative and innovative autonomous agents.

2.2.1 Characteristics of Sensemaking

The characteristics of sensemaking have been described in various ways depending on the theoretical perspective. At its core, sensemaking is characterized by cognitive processes such as acquiring, interpreting, understanding, and taking action (Heath and Johansen, 2018).

Dervin and Naumer differentiated key characteristics related to sensemaking, such as information-related behaviors, diverse behaviors, internal versus external actions, and cognitive processes. These include distinctions like **Internal (cognitive) vs. External (seeking)** behaviors, **Processing vs. Retrieving** for information-related activities, and **Knowing vs. Understanding** for cognitive tasks (Dervin and Naumer, 2009).

Klein proposed a set of characteristics for his Cognitive Task Analysis (CTA) approach, which emphasizes the importance of understanding individuals' internal cognitive processes. His framework examines how people think, what

they know, how they organize and structure information, and the strategies they use to determine what information to seek for deeper understanding (Klein, 2000).

Weick introduced characteristics of sensemaking within organizational communication, describing it as a process that is **grounded in identity construction**, **retrospective**, **enactive**, and inherently **social** (Weick, Sutcliffe, and Obstfeld, 2005). He emphasized that sensemaking is an **ongoing** activity, focused on **extracted cues**, and driven by **plausibility** rather than striving for complete accuracy. This framework is particularly useful in dynamic and uncertain environments, where traditional decision-making models may fall short. This makes it particularly applicable to fields like crisis management and high-reliability organizations, where the ability to quickly interpret and respond to unexpected events is crucial (Maitlis and Christianson, 2014).

By integrating the social and psychological dimensions of sensemaking, Weick's framework offers a robust perspective on how meaning is constructed, accounting for factors such as identity, context, and interactions that shape individuals' understanding of their environments (H. Kaur, Adar, et al., 2022). In this work, I apply Weick's framework to the sensemaking of AI, shifting the focus from solely enhancing the technical aspects of AI systems (e.g., through explanations) to examining how AI systems are experienced and interpreted by various stakeholders in their everyday lives and across diverse contexts.

Applying Weick's framework to AI sensemaking necessitates its extension to address the unique characteristics of AI, including its opacity and dynamic interactions. This extension is a key contribution of this work and will be explored further in *part VI*.

2.2.2 Researching Sensemaking

According to (Naumer, Fisher, and Dervin, 2008), sensemaking is a distinctive research approach that aims to understand how individuals make sense of their experiences and communication, characterized by several key hallmarks. The methodology adopts a **Communication Praxis** perspective, viewing research itself as a form of communication that must be conducted dialogically. The **Certainty vs. Uncertainty** aspect of the methodology acknowledges that individuals move fluidly between certainty and uncertainty, which shapes both their identities and understanding. The **Focus on Processes** aspect prioritizes actions and interactions rather than static descriptions. Another significant hallmark is **Power as a Core Concept**. Here, power is treated as a tangible force that both enables and constrains sensemaking. The methodology takes a phenomenological approach to explore how individuals navigate power dynamics, emphasizing systems responsive to everyday experiences rather than those relying solely on expert-imposed frameworks. Moreover, the methodology views **Humans as Theorists**, positioning individuals as active agents who construct their own understanding of the world.

The methodologies for researching sensemaking go beyond theory building and encompass a wide range of qualitative, quantitative, and mixed-method research techniques (Turner, Allen, et al., 2023). Common methods include case studies, ethnography, conversation and discursive analysis, interviews, narrative analysis, grounded theory, mathematical modeling, social network analysis, action research, and storytelling (Heath and Johansen, 2018).

Different research focuses have been employed to investigate how sensemaking occurs, including the examination of cognitive, emotional, and physical

processes, as well as the interplay between structure and agency (Dervin, 2015). Some studies have identified iterative learning processes that address problems, such as searching for, establishing, modifying representations, and interpreting encoded information (Russell, Stefik, et al., 1993). Other research emphasizes events that trigger sensemaking, like environmental shifts, crises, and identity threats, which lead to ambiguous situations requiring sensemaking responses (Maitlis and Christianson, 2014).

Building on this foundation, the next section will explore sensemaking research relevant to the question addressed in this thesis, specifically focusing on how users make sense of AI technologies as they become increasingly prominent in everyday life.

2.3 Users' Sensemaking of AI Systems

As discussed in the prior sections, despite the widespread use of AI in daily interactions with technology, its opaque nature often leaves users unaware that they are interacting with it (Santiago, 2023). This lack of awareness can hinder users' ability to use AI effectively, collaborate with it, and critically evaluate its outputs. Misconceptions about AI can also result in poorly directed regulatory efforts and disappointment when AI developments do not meet public expectations (Long and Magerko, 2020).

Researchers in the HCI community have begun addressing public misconceptions of AI by investigating how people make sense of AI (e.g., Druga, Williams, et al., 2017) and exploring how to design more understandable technology (e.g., Gunning, Stefik, et al., 2019). For example, by examining users' mental models, which are internal representations that help users

understand and predict AI behavior (e.g., Bauer, von Zahn, and Hinz, 2023), and folk theories, which are informal, often simplistic explanations developed to make sense of complex AI phenomena (e.g., DeVito, Gergle, and Birnholtz, 2017). However, existing studies on mental models and folk theories related to AI are highly fragmented due to variations across different AI application domains (Alizadeh, Stevens, and Esau, 2021).

To provide context for this thesis, I offer an overview of prior research on users' sensemaking of AI within the three key application domains addressed here: 1) AI-assisted decision making, 2) AI-mediated social platforms, and 3) Agentic AI technologies.

2.3.1 AI-assisted Decision Making

"**AI-embedded systems assisting decision making**" are systems that analyze data to derive information used to facilitate human decision making (Castelluccia and Le Métaye, 2019; Schmude, Koesten, et al., 2023). Usually, such systems provide assistance to human decision makers in the form of one or multiple recommendations. When the system is not fully automated, meaning agentic, it is the human who has the final say in making decisions. If the AI's recommendation differs from the decision maker's initial opinion, the decision maker finds themselves in conflict between their initial opinion and the new information received, necessitating a choice between their opinion and the recommendation to make a better decision (Yaniv and Kleinberger, 2000). Making a better decision based on a recommendation means being able to interpret the quality of the recommendation. However, it can sometimes be difficult to understand how a system arrived at a certain conclusion

due to their "black box" nature (Adadi and Berrada, 2018a; Lagioia and Contissa, 2020). This, in turn, obfuscates understanding why a certain AI recommendation was produced, anticipating potential biases in decision making, and identifying the reasons for wrong predictions (Scherer, 2015; Yu and Kohane, 2019). For instance, (Yuan, Bi, et al., 2023), explored how users perceive biases in AI systems, particularly in high-stakes situations such as loan approvals. The study found that users often struggle to detect AI biases due to the opaque nature of these black-box systems. Similarly, Saxena et al. examined the use of algorithms in high-stakes public services like child welfare and criminal justice, revealing that users frequently do not understand the algorithmic processes, leading to inconsistent and unreliable decisions (Saxena, Badillo-Urquiola, et al., 2021b).

When one is uncertain about how to correctly assess the quality of a recommendation (Sniezek and Van Swol, 2001), one can rely on their level of **trust** towards it to decide whether to stick to one's own opinion or to follow the system (Van Swol and Sniezek, 2005; X. Wang and Du, 2018). But the black box nature of such systems can lead to reliance on AI without comprehending its decision making rationale (Kolkman, Bex, et al., 2024).

As AI systems are increasingly used in decision-making processes that affect people's lives, such as public safety (Kalhan, 2013), hiring (Ajunwa, Friedler, et al., 2016) or loan approval (O'Dwyer, 2018), it is important to understand how users perceive and attribute trust to these systems (Board, 2019; Commission, 2020; Fund, 2019; G20, 2019; KPMG, 2019; of Science and Policy, 2020; Services, 2019; Taskforce, 2019; Villani, Bonnet, Rondepierre, et al., 2018). As a result, several studies have focused on understanding the factors that influence Human-AI trust (e.g., Ma, Lei, et al., 2023; Ooge, Kato,

and Verbert, 2022; X. Wang and Yin, 2023; Yin, Wortman Vaughan, and Wallach, 2019; Y. Zhang, Liao, and Bellamy, 2020). These studies primarily examine trust from the perspective of users—individuals who interact with AI-assisted decision-making systems and rely on their recommendations to make decisions (Lai, Chen, et al., 2023). However, less is known about how other stakeholders involved in, or impacted by, the design, deployment, and use of these systems perceive and understand trust in AI. The views of these stakeholders can be shaped by their specific roles and contexts, making their perspectives valuable for a broader understanding of Human-AI trust.

To address this gap, part of this work explores the perspectives of two key groups in risk-sensitive fields such as finance, law, management, and medicine on Human-AI trust. The first group, **AI practitioners**, includes individuals involved in the design and deployment of these systems, ranging from AI developers to project managers. The second group, **decision subjects**, comprises people who do not interact directly with the systems incorporating AI but are affected by the decisions made by users based on the systems' recommendations. By investigating these diverse perspectives in Chapter 5, I aim to contribute to a more comprehensive understanding of trust in AI-assisted decision-making.

2.3.2 AI-Mediated Social Platforms

AI-mediated social platforms are digital environments where AI facilitates social interactions and connections. These platforms use AI to analyze user data, personalize content, and enhance user engagement. some examples include social media and messaging apps (Eslami, Rickman, et al., 2015;

J. T. Hancock, Naaman, and Levy, 2020; Wu and Kelly, 2020), online marketplaces (Jhaver, Karpfen, and Antin, 2018), and dating platforms (Courtois and Timmermans, 2018).

In this dissertation, I focus on two specific types of platforms: social media and dating apps, with broader implications for other social platforms. The following subsections provide an overview of research related to users' sensemaking of AI within these specific categories.

AI Mediation on Social Media

Social media platforms employ algorithmic systems for various purposes, including content curation, bot detection, recommendations, and many more.

Research has shown that users are not always consciously aware of these algorithmic processes, which can negatively impact their interpersonal relationships (e.g., Eslami, Karahalios, et al., 2016; Eslami, Vaccaro, et al., 2019; Jhaver, Karpfen, and Antin, 2018; Raval and Dourish, 2016). For instance, studies on Facebook's news feed curation algorithm have revealed that users frequently do not realize the algorithmic influence at play. This lack of awareness leads them to mistakenly blame their friends and family for the disappearance of posts, rather than recognizing it as a result of algorithmic selection (Eslami, Karahalios, et al., 2016).

Conversely, in some cases users are aware of these process and construct *folk theories* to explain how algorithms work (e.g., Alizadeh, Stevens, and Esau, 2021; DeVito, Birnholtz, et al., 2018; DeVito, Gergle, and Birnholtz, 2017; Karizat, Delmonaco, et al., 2021; T. Liao and Tyson, 2021; Siles,

Segura-Castillo, et al., 2020; Toff and Nielsen, 2018). These folk theories, which are intuitive, informal explanations that individuals develop to make sense of technological systems, have been found to guide user behavior and resistance to algorithmic changes (DeVito, Gergle, and Birnholtz, 2017). Devito et al. conducted several studies on the interplay of folk theories and self-representation practices on social media to explore how folk theories help guide users' behavior in managing impressions of others (DeVito, Birnholtz, et al., 2018; DeVito, Gergle, and Birnholtz, 2017; DeVito, 2021). Karizat, Delmonaco, et al., 2021, applied Identity Strainer Theory to refer to users' beliefs that "*an algorithm filters out and suppresses certain social identities*" (e.g., LGBTQ+ identity). The study also illustrated that users changed their behavior to resist the perceived algorithmic suppression and associated harm. An analysis of 102,827 tweets of a hashtag (i.e., #RIPTwitter) in a Twitter timeline revealed that users' folk theories influence their resistance to algorithmic change (DeVito, Gergle, and Birnholtz, 2017).

Social media platforms also utilize AI algorithms for automated detection of bots and fake accounts. Bots are computer algorithms that autonomously generate content and interact with users, attempting to emulate and potentially influence human behavior (Ferrara, Varol, et al., 2016). Social bots have been used for harmful purposes, such as undermining political discourse, manipulating financial markets, stealing personal data, and spreading misinformation (Ferrara, Varol, et al., 2016). Detecting bots is, therefore, a critical research goal. However, as social bots become more sophisticated, the line between human-like and bot-like behavior becomes increasingly blurred (Akyon and Kalfaoglu, 2019; Guerar, Merlo, et al., 2018). This ambiguity has led to the rise of reverse Turing tests, which determine whether a

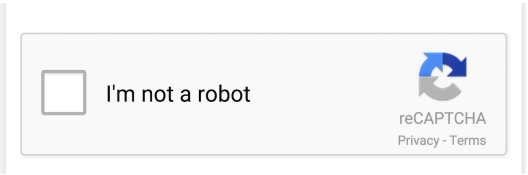


Fig. 2.1.: Google's No CAPTCHA reCAPTCHA test.

user is genuinely human. Google's No CAPTCHA reCAPTCHA, for instance, claims to be "tough on bots" and "easy on humans" (Google, 2024). These algorithms operate invisibly in the background, assessing a user's behavior across multiple dimensions. For the user, the process is seemingly simple: they encounter a checkbox asking them to confirm that they are human (see Figure 2.1, Gafni and Nagar, 2016). Yet behind this simplicity lies sophisticated risk assessment software that, as Google explains, "actively considers a user's entire engagement with the CAPTCHA—before, during, and after—to determine whether that user is a human" (Agarwal, 2023).

Both the No CAPTCHA reCAPTCHA and its enterprise counterpart, reCAPTCHA Enterprise, are marketed as offering a "frictionless user experience" while extending fraud protection across websites (Fortunati, Manganelli, et al., 2019). Consequently, more and more organizations are deploying invisible audits of their platform activities (Alizadeh, Mniestri, and Stevens, 2022b).

While such systems are not immune to error, they provide no help or support to users who are often not even aware of the test, let alone able to understand and judge the fairness of final decision. Researchers have shown that users need multiple attempts to pass a CAPTCHA test, and that a false negative in the bot detection algorithm is inevitable (meaning that a human user will be incorrectly identified as a bot and blocked or banned from the platform) (Rauchfleisch and Kaiser, 2020; Waters, 2021).

On Instagram, the battle begins as soon as a user loses access to their account (Alizadeh and Stevens, 2020). Figure 2.2 displays the notifications Instagram users receive when they exhibit *bot-like* behavior, such as rapidly liking, following, or commenting (Alizadeh and Stevens, 2020). These notifications often result in restricted access to certain platform features, regardless of whether the user's actions were accurately flagged (Register, Qin, et al., 2023). As shown in the figure, the notifications vary depending on the type of restriction applied to the account. However, users are not provided with an explanation for why a specific type of restriction was chosen, leaving them without clarity about the nature of their flagged behavior.

Once the algorithm flags their content, users, especially those who are entrepreneurial through the platform in the form of promotions and sponsorship deals, are targeted by dealers. These dealers offer creators the opportunity to recover their accounts for a fee. While some of them are simply scammers, others inexplicably have enough access to Instagram's moderation infrastructure to restore the account and "right" all the wrongs decided by the algorithm (Alizadeh, Mniestri, and Stevens, 2022b; Waters, 2021). These practices expose the opacity of algorithmic moderation systems, amplifying users' sense of powerlessness and dependence on external actors to navigate platforms they once controlled.

In Chapter 6, I critique this new generation of invisible AI policing on social media by juxtaposing it with the science fiction imagining of the Turing test in the film *Blade Runner*. By drawing parallels between speculative fiction and current practices, I highlight the growing dissonance between security measures and user agency. This part of my work contributes to the scholarly discourse around biometric CAPTCHA tests and invisible AI

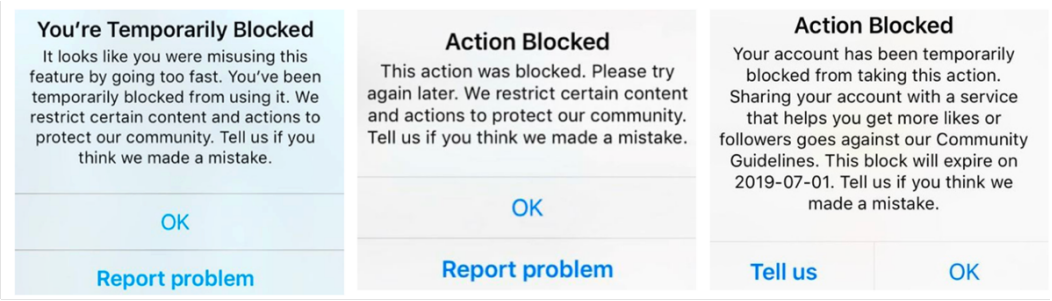


Fig. 2.2.: The notifications Instagram users receive when they exhibit bot-like behavior.

policing by proposing future trajectories for HCI research to design systems that preserve user security and privacy while fostering meaningful control and agency.

AI Meditation on Dating apps

AI-mediated social platforms are not limited to social media; they also include dating apps, where AI algorithms play a crucial role in facilitating connections. For instance, research has explored how users of dating apps develop algorithmic folk theories to make sense of what they experience on these platforms (Nader and Lee, 2022), to explain how matching algorithms work and increase their attractiveness rating (S. A. Huang, Hancock, and Tong, 2022; Nader and Lee, 2022), and even to manage the subsequent stages of relationship formation (Sharabi, 2021; Tong, Hancock, and Slatcher, 2016). Tinder has long been speculated to base its matching algorithm on "Elo scores" (Junkie, 2020; K. L. Smith, 2020) - a measure of one's attractiveness based on the number of "likes" one receives in conjunction with how selective one is in giving out "likes" oneself (the most attractive users would therefore receive the most likes while giving out the least likes). In March 2019, the company announced they had abandoned the Elo score (see Figure

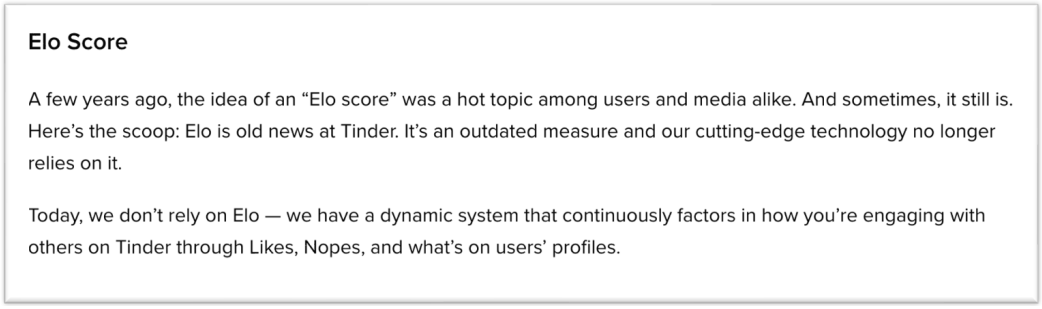


Fig. 2.3.: The announcement from the Tinder website explaining that the matching algorithm has abandoned the Elo Score.

2.3); however, users are still in the dark about how the algorithm works (Tinderpress, 2022a).

To infer the inner workings of the matching algorithm, Courtois and Timmermans used an experience sampling method and measured users’ swiping and liking ratio, along with resultant number of matches, (Courtois and Timmermans, 2018). The study concluded that Tinder “prohibits its major assets of attractive profiles and liked profiles to run out too soon” in order to convert non-paying users to premium subscription plans (p. 13).

Nader and Lee observed that users’ belief in a desirability score (similar to the Elo score speculation) led to various strategies to increase one’s score, such as creating new profiles or adjusting search filters (Nader and Lee, 2022). Abel, Pei, et al., 2022, built upon this by conducting 22 in-depth user interviews. Their findings revealed diverse theories among users about how Tinder’s algorithms operate. They attributed Tinder’s profile curation not only to desirability scores but also to factors like user feedback, romantic compatibility, and the profitability of choices. Similarly, in their interview study, Huang et al. identified folk theories surrounding the Tinder matching algorithm. They noted perceptions of the algorithm as facilitating a form of

'shopping', being subject to chance and randomness, and functioning as a bracket system for scoring users (S. A. Huang, Hancock, and Tong, 2022).

These examples illustrate the diverse ways users engage in sensemaking of AI systems on dating apps, often by developing folk theories to explain algorithmic outcomes. However, despite these insights, the explicit study of user perceptions regarding potential algorithmic harms, unfavorable outcomes, and the coping strategies employed to navigate such platforms has remained underexplored.

In Chapter 7, I address this gap by examining the nature of algorithmic harm on Tinder, the folk theories users construct, and the strategies they employ to mitigate such harm.

2.3.3 Agentic AI

Agentic AI refers to AI systems that operate autonomously, making decisions and taking actions without human intervention (Y. Huang, 2024). An **agent** in this context is an entity, such as a robot, conversational agent, or voice assistant, that interacts with users and performs tasks independently, enhancing efficiency and user experience (Dragoni, Sernani, Calvaresi, et al., 2018). Humans tend to anthropomorphize such computational artifacts to rationalize actions and behaviors they cannot reasonably explain. This tendency can lead people to overestimate the capability and knowledge of autonomous systems (Phillips, Ososky, et al., 2011). Research has shown that users' mental models of agents are shaped by factors like appearance, physical attributes, dialogue, personality traits, language, and origin, but often

lack a clear understanding of their mechanical and conceptual functionalities (Kiesler and Goetz, 2002; Phillips, Ososky, et al., 2011).

For example, Kiesler and Goetz showed that robots' facial features, such as the dimensions of the forehead and chin, impact the perception of intelligence (Kiesler and Goetz, 2002). Consequently, people developed high expectations of the robots that were not fulfilled, harming collaboration and relationship building. Anthropomorphism can support users in their approach and reactions to robots and prevent initial rejection. However, it might also lead to misconceptions, which can have detrimental consequences for the outcome and even endanger human lives, such as in military settings or critical workspaces (Phillips, Ososky, et al., 2011). Therefore, humans need a clear and accurate understanding beyond anthropomorphism of how agents collect and process data and make decisions.

Besides anthropomorphism (Kuzminykh, Sun, et al., 2020) many studies in HCI explored voice interactions and the adoption of domestic agents. In particular, some studies have investigated the perception and understanding of attributed intelligence (Druga, Williams, et al., 2017; Garg and Sengupta, 2020; Kuzminykh, Sun, et al., 2020; S. Lee, Kim, and Lee, 2019; Y. Xu and Warschauer, 2020). Researchers have used methods of drawing (Kuzminykh, Sun, et al., 2020; S. Lee, Kim, and Lee, 2019; Y. Xu and Warschauer, 2020) interactive tasks (Druga, Williams, et al., 2017; Garg and Sengupta, 2020; Kuzminykh, Sun, et al., 2020), and interviews (Druga, Williams, et al., 2017; Garg and Sengupta, 2020) to explore the reasoning and explanations of the behavior of the voice assistants that children of different ages create. Xu and Warschauer showed that children used behavioral references like listening and talking to justify cognitive properties, reciprocity as an indicator of psy-

chological properties, and biological references like mechanical causality or fantasy reasoning to explain behavioral properties (Y. Xu and Warschauer, 2020). They tended to allocate conversational agents rather to a continuum between humans and artifacts than to a distinct category. Especially younger children who were less aware of the underlying concepts attempted to make sense of computational artifacts by personifying them. When children acknowledged the devices to be more intelligent than themselves (Garg and Sengupta, 2020; Y. Xu and Warschauer, 2020) they were more likely to trust and believe the information provided, which contributed to improved learning (Garg and Sengupta, 2020). To examine the state of intelligence or consciousness, the study of Druga et al. indicated that younger children asked personal questions such as "what is your favorite color?" and older children asked the device to perform actions of which they knew humans were capable (Druga, Williams, et al., 2017). With prior experience with technology or an engineering background, more thoughtful reasoning was applied (S. Lee, Kim, and Lee, 2019). Nonetheless, voice and tone affected the perceived friendliness (Druga, Williams, et al., 2017), and available input modalities communicated the expected level of intelligence (Druga, Williams, et al., 2017; Garg and Sengupta, 2020). Kuhl et al. explored the relationship between users' mental models and behaviors, the study highlights that users' understanding of personalized agents' learning algorithms varies largely based on their backgrounds, such as training, domain knowledge, and demographic characteristics (Kuhl, Lobana, and Meske, 2020).

In summary, while agentic AI systems have the potential to enhance user experience and efficiency, their effectiveness relies heavily on users' mental models and anthropomorphic tendencies. Research suggests that users often

personify AI agents, which can lead to an overestimation of their capabilities.

As part of this work (chapter 9), I aim to investigate how agentic AI technologies—specifically voice assistants—can effectively support users in making sense of the agent's behavior, particularly in situations where it behaves in unexpected ways or does not meet user expectations.

2.4 Sensemaking and User Empowerment

Empowerment is a multifaceted concept with diverse interpretations. It can be understood as a process, method, or end goal; a worldview, ideology, or paradigm; an approach to action, or even a symbol or metaphor (Schneider, Eiband, et al., 2018). As Rappaport, credited with popularizing the term, notes: "empowerment is a little bit like obscenity; you have trouble defining it but you know it when you see it" (Rappaport, 1985).

At its core, empowerment is rooted in the concept of power. In sociology and organization theory, power is often categorized into **episodic** and **systemic** forms (Clegg and Haugaard, 2009; Fleming and Spicer, 2014). Episodic power involves deliberate actions by individuals or groups to influence or control others, while systemic power operates subtly through institutional norms and shared knowledge structures, shaping how actors perceive and act within the world (Clegg, 2023; Fleming and Spicer, 2014).

The term "user empowerment" in HCI has been closely associated with accessibility, focusing on design practices that enable users, particularly those

with disabilities, to engage effectively with technologies (e.g., Cha, Kim, et al., 2021; Ladner, 2015; Trewin, Basson, et al., 2019).

In the context of AI, however, empowerment has evolved into a broader concept within HCAI. This approach emphasizes designing systems that enhance and amplify human capabilities rather than replace them (Usmani, Happonen, and Watada, 2023). Shneiderman, for instance, suggests that HCAI bridges ethical considerations and practice by providing recommendations for creating systems that augment and support human abilities (Shneiderman, 2022).

Sensemaking offers another perspective on empowerment by addressing the dynamics of power and control in users' interactions with systems. Sensemaking, as proposed by Dervin, provides a framework for addressing these dynamics by emphasizing that "knowledge made today is rarely perfectly suited to application tomorrow" and that "attending to the unmaking of sense is as important as attending to its making." (Dervin and Naumer, 2009). Dervin highlights two barriers to this process: the entrenched belief in definitive answers and hidden power structures that constrain dissent. Sensemaking challenges these barriers by valuing "deviant observations" that disrupt dominant narratives. Dervin argues that "no matter how closed a system, somewhere, someone is making deviant observations" that can benefit the entire system if these perspectives are safely incorporated (Dervin and Naumer, 2009). By mandating attention to power in its methods, sensemaking turns moments of interpretation into opportunities for disagreement and change (Schildt, Mantere, and Cornelissen, 2020).

These ideas connect with Zimmerman's definition of psychological empowerment, as "a process in which people gain understanding and control over

personal, social, economic, or political forces in order to take action to better their lives" (Zimmerman, 1995). Together, these perspectives extend the concept of user empowerment beyond accessibility or augmented abilities. They underscore the importance of designing systems that not only enhance capabilities but also address power imbalances, enabling users to challenge and shape their interactions with technology. Since sensemaking involves the "situated formation of shared understandings" (Gioia and Chittipeddi, 1991; Hobbes and Missner, 2016; Maitlis, 2005; Weick and Weick, 1995), it provides a useful framework for examining how power operates.

Schildt et al. extend the distinction between episodic and systemic power to the context of sensemaking, demonstrating how power not only shapes the content of sensemaking—what is discussed and prioritized—but also its form, including who participates, how discussions unfold, and what frameworks are employed (Schildt, Mantere, and Cornelissen, 2020). This distinction is particularly relevant for understanding user empowerment in AI systems, as it highlights both the immediate, interaction-level effects of power and the broader structural constraints embedded in system design.

To position my work, which uses sensemaking as a lens to explore user empowerment needs, I first examine various interpretations of empowerment within HCI and review the current state of research on user empowerment in interactions with AI systems.

2.4.1 Empowerment in HCI

Empowering people through technology is of increasing concern in the HCI community. Schneider et al. reviewed 54 CHI full papers using the terms

empower and *empowerment* and proposed a framework to analyze notions of empowerment in HCI research. Their analysis identified four key categories to classify existing approaches (Schneider, Eiband, et al., 2018).

- **Concept of Power:** Empowerment in HCI can be viewed through the lenses of "power-to" (the ability to act) and "power-over" (challenging authority). Systems that enable individuals to build their own technology reflect power-to, while those supporting communities in challenging systemic power structures align with power-over.
- **Psychological Component:** Drawing from Zimmerman's theory of psychological empowerment, empowerment manifests in three dimensions:
  - *Feeling*: For instance, systems that boost users' self-esteem through positive feedback.
  - *Knowing*: For instance, systems that teach skills that improve users' opportunities, such as career advancement.
  - *Doing*: For example, systems that enhance efficiency by automating tasks or enabling users to accomplish more in less time.
- **Persistence of Empowerment:** Empowerment can be transient or persistent. For instance, a prosthesis empowers its user only during use (transient), while educational tools or voting systems grant persistent empowerment that extends beyond immediate interaction.
- **Design Mindset:** Sanders distinguishes between two design approaches: an *expert mindset*, where users are treated as subjects or consumers, and a *participatory mindset*, where users are co-creators in the design pro-

cess. A participatory mindset is crucial for meaningful empowerment, as it emphasizes user initiative over reliance on authority.

Among the different approaches to empowerment categorized above, those based on psychological components—or what Schneider refers to as "holistic approaches"—not only address interpersonal, interactional, and behavioral aspects but also emphasize the interconnectedness of feeling, knowing, and doing in fostering successful empowerment (Schneider, Eiband, et al., 2018). Therefore, within the scope of this work, I draw on holistic approaches to explore users' empowerment needs in a more comprehensive manner.

2.4.2 AI and User Empowerment

User empowerment has become a central goal of the human-centered AI discipline and is also reflected in regulatory frameworks such as the EU's AI Act (Valdez, Heine, et al., 2024). These initiatives aim to ensure that AI technologies enhance human agency, protect fundamental rights, and promote equitable outcomes (Usmani, Happonen, and Watada, 2023).

Despite these efforts, AI systems have been shown to pose hidden risks, including harm and bias. The term *algorithmic harm* encompasses a wide range of issues, including racial, gender, and identity bias (e.g., Buolamwini, 2022; Gerrard and Thornham, 2020; Hamilton, 2019; Noble, 2018), discrimination (e.g., Citron and Pasquale, 2014; Kantor, 2022), privacy violations, surveillance, news filtering, and electoral impacts (e.g., Tufekci, 2015), and so on. Scholars categorize these harms into two main types (Barocas, Crawford, et al., 2017):

- **Allocational harms**, which occur when resources or opportunities are unfairly distributed or withheld.
- **Representational harms**, which involve the misrepresentation or marginalization of social groups, as highlighted in studies by (Andalibi, Pyle, et al., 2023a; Karizat, Delmonaco, et al., 2021; Katzman, Wang, et al., 2023).

Further classification is provided by Saurwein and Spencer-Smith, who identify roles for algorithms in perpetuating harm on social media platforms, such as (1) deficient tools that lead to errors, (2) instruments for manipulation, (3) amplifiers of problematic content, (4) enablers of harmful behavior, and (5) tools of platform power (Saurwein and Spencer-Smith, 2021). The AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) repository documents over 1,000 cases of algorithmic harm and has developed a taxonomy to analyze them (AIAAIC, 2023).

Algorithmic biases are also evident in biometric technologies, which identify individuals based on physical or behavioral traits (A. K. Jain and Kumar, 2012). As Pugliese argues, biometric technologies are deeply embedded with societal biases related to race, gender, disability, class, and age (Pugliese, 2012). For instance, Joy Buolamwini, a Ghanaian-American MIT graduate student, exposed racial bias in facial recognition software that failed to recognize her dark-skinned face while accurately identifying her lighter-skinned peers. Ironically, when she wore a white mask, the system successfully recognized the mask's features, further highlighting the discriminatory nature of the software.

Cases like this underscore Kate Crawford's warning that the greatest threat posed by AI systems is not their potential to surpass human intelligence but their capacity to "hard-code sexism, racism, and other forms of discrimination into the digital infrastructure of our societies" (D'ignazio and Klein, 2023).

These examples highlight the need to address user empowerment in everyday AI interactions. Part of this work uses *sensemaking* as a lens to explore where and how users require empowerment in everyday life situations.

2.5 Designing for Sensemaking and Empowerment

HCI researchers have employed various approaches to design for sensemaking and empowerment in AI interactions (e,g, Castelli, Ogonowski, et al., 2017; Eslami, Rickman, et al., 2015; Reddy, 2020). Following the Sanders design mindset for empowerment in this thesis, I apply two specific approaches: **interaction design** and **data-enabled design** to facilitate different levels of user participation in the sensemaking process (see Section 3.1). Sensemaking through interaction design involves creating interfaces and experiences that enhance users' understanding of AI systems (Greenwald, Krakowski, et al., 2024; Morrison, Cutrell, et al., 2021). Sensemaking through data-enabled design takes a broader approach by integrating real-time data into the design process. This method involves capturing contextual, behavioral, and experiential data from users interacting with AI systems (Bogers, van Kollenburg, et al., 2016).

In the following subsections, I provide an overview of previous research that has applied these approaches to design for AI sensemaking and user empowerment.

2.5.1 Sensemaking and Interaction Design

Sensemaking can be approached in designing users' interactions with AI through multiple methods, but for the purpose of this thesis, I specifically examine **accountability/explainability-focused design** and **transparency-focused design**.

Accountability-focused design ensures that users can trace the decision-making processes of AI systems, making it clear how and why certain outcomes are produced (Gammelgård-Larsen, van Berkel, et al., 2024). For instance, Hocutt's examination of Alexa's decision-making process in source selection revealed a default preference for retail solutions. He underscored the importance of providing users with reasons for source selection beyond mere citations or URLs (Hocutt, 2021). Porcheron et al. explored accountability in the design of user requests, arguing that responses from digital assistants should serve as resources for users to gauge the system's state and inform further interaction. They described accountability as the achievement of effective interaction through these responses (Porcheron, Fischer, et al., 2018). Echoing this, Cho and Rader argued for designing responses to serve as "evidence" that coordinates a conversational process, as opposed to optimized outputs (J. Cho and Rader, 2020). This approach would allow users to continually explore a system's functionality and capabilities over the long term. Additionally, the authors emphasized the importance of mental models

in understanding how users perceive and interact with VAs, suggesting that well-designed responses can help users form accurate mental models of the system's capabilities and limitations (J. Cho and Rader, 2020). Kianzad et al. described accountability-aware design as design that considers context, "specifically task complexity, type of user(s), and ambiguity of the interactions" (Kianzad, Kim, et al., n.d., p. 7, 32). This study also showed that recommendations and instructions from the system generate a sense of "shared accountability" among the users. Seymour et al. investigated designing accountability through trust and privacy explanations, recommending the use of analogies to enhance user understanding (Seymour and Van Kleek, 2021).

A significant body of research has also focused on designing explainable AI (XAI) to support users' understanding of AI-based systems and combat incomplete mental models (e.g., Hoffman, Mueller, et al., 2018; Lim, Dey, and Avrahami, 2009; Ribera and Lapedriza, 2019). XAI refers to AI systems whose activities can be effortlessly comprehended and examined by humans (Hagras, 2018). By refining users' mental models and resolving their misconceptions, XAI promises to improve user performance (Molnar, 2020).

However, XAI has been criticized for several limitations. Some researchers have criticized XAI for leading to less efficient systems with less flexible and capable outputs (Molnar, 2020; Ribera and Lapedriza, 2019). To address this issue, XAI researchers have argued that explanations are not always necessary and have identified specific application domains, such as healthcare and military contexts, where they can provide significant benefits (Eslami, Karahalios, et al., 2016). Molnar suggests that human-friendly explanations

should focus on anomalies, explaining system behavior that deviates from users' expectations rather than offering complete, exhaustive explanations of all causes (Molnar, 2020).

Another critique of XAI is its emphasis on practitioners, often overlooking the needs of general users, which may result in explanations that are difficult for non-experts to understand (Weitz, Schiller, et al., 2019). Ehsan et al. have advocated for human-centered XAI methods, emphasizing the importance of creating actionable frameworks and evaluation methods to bridge gaps in algorithmic opacity and address user needs effectively (Ehsan, Wintersberger, et al., 2021).

Although accountability and explainability are often treated as separate strands of research, part of this work aims to explore the intersection of these concepts by examining how users make sense of an AI agent's behavior. Specifically, I investigate whether providing explanations during user interactions can fulfill users' need for accountability.

Transparency-focused design, on the other hand, aims to make the inner workings and limitations of AI systems visible to users. By exposing the uncertainties and potential biases in AI outputs, users are better equipped to interpret and interact with these systems effectively (Morrison, Cutrell, et al., 2021). Seamful design is a key strategy in this context, as it deliberately highlights the imperfections and boundaries of AI capabilities, fostering a more informed and critical user engagement (Chalmers, MacColl, and Bell, 2003). Seamful design emphasizes configurability, user appropriation, and the revelation of complexity, ambiguity, or inconsistency (Ehsan, Liao, et al., 2024). It contrasts with seamless design, which prioritizes clarity, simplicity, and ease of use (Inman and Ribes, 2019). By exposing the seams,

or the inherent limitations and uncertainties in AI systems, users can better understand the system's behavior and limitations, leading to more effective and empowered interactions (Chalmers, MacColl, and Bell, 2003). For example, Eslami et al. developed a system called FeedVis to reveal the difference between algorithmically curated and unadulterated news feeds to users (Eslami, Rickman, et al., 2015). This exposure led to users experiencing surprise and anger initially, but eventually, it bolstered their overall feelings of control and engagement with the platform . Another study by Eslami et al., examined how users discovered and managed biases in hotel rating algorithms (Eslami, Vaccaro, et al., 2017). Users engaged in practices like reverse-engineering the rating algorithm and raising awareness among other users, demonstrating how seamful design can empower users to detect and respond to algorithmic biases.

However, the concept of algorithmic transparency has also been criticized for shortcomings. In particular, Ananny and Crawford argue that transparency in its literal meaning as *being able to see a system* is not synonymous with understanding algorithms: We can *see without knowing* (Ananny and Crawford, 2018). Moreover, transparency can also bring its own problems, especially when not carefully designed. The wrong level of transparency can burden and confuse users, making it difficult for them to interact with the system (Eslami, Vaccaro, et al., 2019). Too much transparency may even have undesirable consequences, such as allowing users to strategically manipulate the system and reduce competitiveness (Palmiotto, 2020).

While transparency aims to reveal the inner workings and limitations of AI systems, it does not guarantee that users will effectively comprehend or interpret this information. Therefore, part of this thesis explores how

transparency-focused design can move beyond merely exposing information to actively enhancing users' interpretation and understanding, ensuring that transparency leads to meaningful and empowering interactions.

2.5.2 Sensemaking and Data-enabled Design

Data collection processes in AI systems are often opaque, leaving users with limited awareness and understanding of how their data is collected and used (Alizadeh, Jakobi, et al., 2019; Jakobi, Stevens, et al., 2020; Merrill, Chuang, and Cheshire, 2019). Additionally, individual users typically lack the tools and expertise needed to manage data sensitivity effectively (Castelli, Ogonowski, et al., 2017; J. E. Fischer, Crabtree, et al., 2016; Tolmie, Crabtree, Rodden, Colley, and Luger, 2016). These challenges are particularly relevant in the context of smart home technologies, which operate within private spaces. Research by Tolmie at al. shows that smart home users require more than simple notifications to understand data practices—they need comprehensive support for exploring and interpreting their data (Tolmie, Crabtree, Rodden, Colley, and Luger, 2016).

Data-enabled design integrates data analytics and user-centered design to improve the interpretability and usability of IoT systems. This approach employs iterative prototyping and user feedback loops to refine designs, making them more adaptable and suitable for real-world use. The goal is to help users make sense of complex data through intuitive interfaces and interactions, fostering informed decision-making and enhancing user satisfaction (Reddy, 2020).

Researchers have applied data-enabled design to enhance data literacy and help users understand system behaviors, including data disclosure (e.g., Castelli, Ogonowski, et al., 2017; Jakobi, Stevens, et al., 2018). For example, Seymour developed natural language processing tools to assist users in identifying patterns in sensitive home data (Seymour, 2019). Other approaches focus on privacy measures, such as storing data locally on devices (Yao, Basdeo, et al., 2019) or minimizing data collection (D. A. Muller and Lévy, 2019). Tools have also been developed to notify users when a device shares data (Seymour, 2019) or provide summaries of shared information (Balebako, Jung, et al., 2013).

Research in the field of usable privacy (A. Adams and Sasse, 1999; Brodie, Karat, et al., 2005; S. Garfinkel and Lipford, 2014) aims to enable users to manage data disclosure by offering effective privacy controls (A. Adams and Sasse, 1999; Cranor, 2008; Whitten and Tygar, 1999). This includes tools for accessing and managing personal data as well as understanding how the data is processed (S. Garfinkel and Lipford, 2014; Van Kleek, Liccardi, et al., 2017). Studies have often focused on sensor data, such as energy consumption, system status, and presence detection, with existing tools supporting sensemaking of smart home data (Castelli, Ogonowski, et al., 2017; J. E. Fischer, Crabtree, et al., 2016; Kurze, Bischof, et al., 2020). Moreover, privacy-aware, data-enabled objects have been developed to empower users to control their data release actively or passively, such as sensor-augmented research tools like observational cameras with coverable lenses (Y.-T. Cheng, Funk, et al., 2019, 2023).

Overall, research on data-enabled design underscores the need for flexible and transparent tools that help users make sense of complex data. Despite

progress in smart home data visualization and privacy-aware design, gaps remain in providing tools to support sensemaking of interaction data collected over extended periods. In Part V of this thesis, I address these gaps by introducing a data visualization tool to help users interpret and make sense of the data collected by their voice assistants.

3

Methodological Approach

This chapter covers the methodological framing for researching how users make sense of AI in their everyday lives. Arriving at the methodological framework requires acknowledging the situation under which this thesis was conducted—specifically, the challenges posed by COVID-19. Conducting empirical research during the pandemic required a shift in the research process due to reduced in-person interactions and increased reliance on online methods.

However, the primary methodological challenge lay in the evolving nature of AI, where users' sensemaking was shaped not only by the shifting perceptions of AI across different application domains but also by the continuous changes in the underlying algorithmic processes within each specific context. This required a methodology that could both accommodate the limitations imposed by the pandemic and capture users' sensemaking as a dynamic and ongoing process.

In this chapter, I detail how I addressed these challenges by:

1. Drawing from multiple research frameworks to account for users' dynamic sensemaking from different perspectives (3.1).

2. Shifting to methods such as distributed data collection, experience sampling, and review analysis, which allowed for the collection of rich empirical data with minimal in-person interaction (3.2).

The ultimate question I aim to answer with my methodological choices is: How can research adapt to such unprecedented conditions while ensuring the depth and quality of qualitative inquiry?

3.1 Research Design

As I mentioned earlier, to explore how users make sense of AI, I approached the topic from a mix of different angles and disciplines. These include symbolic interactionism (how people ascribe meaning to AI), sociology (the context of experiences and users' everyday practices), cognitive sciences (folk theorization), and design (looking at data as an interaction material). By integrating these diverse perspectives, I was able to capture the complex and dynamic nature of sensemaking from different viewpoints. This multidisciplinary approach also guided my selection of methods and research activities for the 6 papers presented in this thesis. However, broadly speaking, the methods and research activities in this thesis are primarily inspired by three main frameworks: Design Case Study (DCS) (Wulf, Müller, et al., 2015), co-design (Sanders and Stappers, 2008), and Research through Design (Giaccardi, 2019).

Up next, I'll describe these three frameworks to provide a big-picture understanding of the foundation for the specific methods and research activities I applied. After that, I'll dive into each method and research activity I carried out in the different studies presented in this work (see Subsection 3.2).

Design Case Studies

Since humans' sensemaking of everyday encounters with specific phenomena is highly contextual and situated within daily activities (e.g., Fitzgerald and Palincsar, 2019; Lowe and Rod, 2018), DCS provide an excellent tool for exploring this process. DCS emphasize the importance of context and the situated nature of social practices (Wulf, Müller, et al., 2015). DCS approach unfolds in three main stages (Wulf, Rohde, et al., 2011):

The first stage involves a detailed examination of the existing social practices before any intervention takes place. This includes documenting the tools and media in use and understanding their application within the given context. The second stage shifts focus to the development of innovative ICT artifacts, examining both the product itself and the design process behind it. This stage also considers the involvement of stakeholders, the design methods employed, and the emergence of key concepts. A critical aspect of this phase is understanding how the anticipated shifts in social practices informed the design choices made. The final stage tracks the introduction of these ICT artifacts into the real-world setting, observing how they are adopted, adapted, and potentially redesigned within their intended environment. This stage allows for an analysis of how specific features and design decisions impact social practices (Wulf, Pipek, et al., 2018).

As a result, DCS provide a thorough method for examining how social practices interact with the design of ICT artifacts. They are especially valuable for comparing different contexts, as multiple case studies allow researchers to identify recurring patterns, distinguish differences, and draw broader conclusions across various settings (Wulf, Müller, et al., 2015).

While the overall structure of most studies reported in this thesis was inspired by the DCS framework, particularly in conducting an initial phase within each application domain to gather in-depth data and understand existing social practices, the studies do not strictly adhere to all stages of the DCS framework. Instead, they focus on capturing the context in which encounters with AI occur, aiming to provide a full picture of how AI technologies are understood in daily life. The insights gained from this process are then used to formulate design implications.

Co-Design and Research through Design

The design chapter of this thesis draws inspiration from two key design frameworks: Co-design and RtD. While both frameworks emphasize the concept of "participation", enabling users to act as active contributors rather than passive recipients in the design process, they do so based on differing interpretations of participation. These differences center around factors such as the level of agency granted to participants and the perceived value of their involvement (Bardzell, Bardzell, and Blythe, 2018).

Co-design, is a participatory approach where users are actively involved as partners throughout the design process. This method emphasizes collaboration and shared decision making, ensuring that the users' needs and perspectives are integral to the development of the final product. Users in co-design are often engaged in various stages, from initial idea generation to prototype testing and final implementation (Sanders and Stappers, 2008).

Research through Design, on the other hand, is an iterative process of research exploration in which the inquiry is introduced through a design implementation and then re-framed and re-focused based on reflections

on the outcomes of that inquiry (Ylirisku, Halttunen, et al., 2009). In this context, users participate in experiments or design practice that is typically carried out with a mindset focused on use. This implies that the results cannot be entirely predicted or planned (Koskinen, Zimmerman, et al., 2013). Therefore, researchers must consistently re-frame and re-focus their research questions through a process of reflection-in-action (Schön, 2017).

Hence, the primary difference between co-design and RtD lies in the depth and nature of user involvement. In Co-design, users are treated as equal partners, actively contributing to all stages of the design process, which often results in more user-centered and contextually relevant solutions (Dietrich, Trischler, et al., 2017). In contrast, RtD generally involves users in more specific, often evaluative roles, focusing on gathering data to inform the design process rather than on co-creating the solution (Busciantella-Ricci and Scataglini, 2024).

In this context, I approach sensemaking through user participation in the co-design process to create contextually and situationally relevant solutions (Chapter 9), as well as through user engagement in design experiments initiated by a design implementation to enable productive encounters with data (Chapter 8). These design experiments provided a space to explore questions like, "What if we introduced this type of engagement with the data collected by voice assistants". Thus, RtD allowed me to critically explore users' sensemaking by integrating technology to examine and question the values and consequences of the design process.

3.2 Methods and Activities

In this subsection I provide an overview of the research activities undertaken in the studies covered in this thesis (see Table 3.1). The selection of specific methods and activities for each study was informed not only by the research frameworks discussed in the previous subsection but also by the constraints imposed by the COVID-19 pandemic. The following sections elaborate on these methodological considerations and offer a comprehensive account of how the research was conducted.

3.2.1 Interviews and Distributed Data Collection

Interviews provide an opportunity for an "interchange of views" between two individuals engaged in a conversation on a topic of mutual interest (Kvale, 2009). They enable researchers to delve into complex phenomena through direct, interactive engagement with individuals, yielding rich and detailed data (Hong and Cross Francis, 2020). There are several types of interviews, each suited to different research needs.

- **Semi-structured interviews:** Use a flexible framework with open-ended questions, allowing for in-depth exploration while maintaining a degree of consistency across interviews (Dietrich, Trischler, et al., 2017). Generally, these guidelines ensure data comparability, provide an overview of topics, and help structure the interview flow, serving as a reference to cover key subjects. Open-ended questions, in particular, create an environment where interviewees feel at ease sharing their personal perspectives. Unlike closed-ended questions used in quantitative surveys, the goal here is to encourage participants to provide detailed

Tab. 3.1.: Research activities according to section and publication of the thesis

Chpt.	Paper (Short Title)	Used Methods
4	Perceived AI	*Empirical study:* 50 semi-structured interviews
5	Trust in AI-assisted decision making	*Empirical study:* 7 semi-structured user interviews + 7 expert interviews
6	Invisible AI	*Conceptual study:* Sci-fi Film Analysis
7	Perceived Algorithmic Harms	*Empirical study:* 7,043 review analysis + 30 semi-structured interviews
8	Alexa, We Need To Talk	*Empirical study:* 11 semi-structured interviews *Technology Design:* Research through Design *Evaluation:* 11 cognitive walk-through + thinking aloud
9	VA's Accountability	*Empirical study:* Living lab (field work, semi-structured interviews) *Technology Design:* Co-design, Role-Playing *Evaluation:* 17 semi-structured Interviews

responses, facilitating the collection of rich, contextual data. As a result, semi-structured interviews strike a balance between maintaining coherence and allowing the conversation to follow the participant's thoughts and narrative (Honer, 2004).

- **Expert interviews:** Are designed to gather insights from individuals with specialized knowledge or expertise in a specific area, offering valuable perspectives that can inform targeted research questions or areas of study (Busciantella-Ricci and Scataglini, 2024). This method focuses on the interviewee's role as a representative of a particular field or group, such as professionals in the food or health industries, rather than their individual viewpoints (Meuser and Nagel, 2009).

To collect a comprehensive and diverse dataset while mitigating the risks of in-person contact during the COVID-19 pandemic, I transitioned to online interviews and adopted a distributed data collection approach, involving participants in the data-gathering process. This method is supported by various studies that adapted their data collection strategies by engaging community members and locals to gather data, thereby minimizing in-person contact between researchers and participants while ensuring safety and efficiency during the pandemic.

For example, a study in Sudan used a structured questionnaire to collect data from 800 participants, leveraging local community members to gather information on the socioeconomic and social impacts of COVID-19 (Mohammed, Abdallah, et al., 2024). Similarly, a qualitative study in South Africa involved health care workers and community members, both providing and accessing health services, using in-depth interviews and focus group discussions to collect data (Rwafa-Ponela, Price, et al., 2022).

In another instance, researchers in New Zealand adapted their data collection methods by conducting focus groups and online interviews, collaborating with community leaders to recruit participants and using telephone calls and letter writing for those without digital access (Ofe-Grant, 2022).

In Chapter 4, I worked with students to gather data from their immediate contacts. This approach enabled me to collect interview data from diverse cohorts while safely managing the risks posed by COVID-19.

Science Fiction Film Analysis

In HCI, analyzing science fiction films involves exploring how futuristic technologies and user interactions are portrayed to inspire and guide design. This approach uses the imaginative scenarios in these films to examine potential technological advancements and their societal impacts (Mubin, Obaid, et al., 2016).

For example, one study investigated the representation of software development in science fiction films, identifying novel programming methods and envisioning the democratization of software development (Krings, Bohn, et al., 2023). Another study focused on the lessons HCI and user experience design can draw from science fiction, analyzing how human-computer interactions are depicted and what insights can be gained (Marcus, 2023).

In Chapter 6, I used this method to contrast the futuristic portrayal of the Turing test in the sci-fi film *Blade Runner* with emerging AI policing practices on social media platforms. This comparison not only critiques these practices but also offers insights for designing more meaningful user control in the future.

Application Review Analysis

App review analysis is a method used to collect data by examining user feedback on mobile applications. It involves gathering and interpreting re-

views from platforms like the App Store and Google Play to uncover valuable insights regarding user experiences, preferences, and problems. Automated tools, such as sentiment analysis, can classify reviews into positive or negative categories, providing a numerical gauge of user satisfaction (Gao, Zeng, et al., 2018). Additionally, approaches like active learning and keyword-based topic modeling can categorize reviews into groups such as features, bugs, and user experience, minimizing the need for manual data labeling (Uthirapathy and Sandanam, 2024). This method has been widely used in studies to gather contextual data from social platforms, offering insights into user behavior and sentiment (Q. Hou, Han, and Cai, 2020).

During the COVID-19 pandemic, app review analysis became a useful tool, as it allowed for the collection of large-scale rich data without physical contact, maintaining both safety and data quality (N. Jain and Brockova, 2022). In Chapter 7, I used this method to collect contextual data from a wide range of Tinder users. This approach allowed me to gain a broader understanding of the harms and concerns perceived by the app's large user base.

Cognitive Walk-through and Thinking Aloud

The Cognitive Walk-through method is a usability evaluation technique aimed at assessing how easily a product or service can be learned. It involves a small group of product team evaluators who go through a series of tasks, asking questions at each step to assess whether a user would understand what to do and whether they would take the correct action. This method is particularly effective for identifying usability challenges that new users might face (Wilson, 2013). It has been applied in various contexts, such as

evaluating mobile apps for caregivers of children with severe burns, helping to assess user interactions and the app's overall usability (Fitria, 2024).

The Thinking Aloud method requires participants to verbalize their thoughts as they complete a task. This real-time verbalization provides insight into users' cognitive processes and decision-making strategies. With roots in cognitive psychology, it has been widely used in fields such as education and user interface design (Gardin, 2010). For instance, in reading comprehension research, the method has been used to uncover the flexible, goal-driven strategies of expert readers and to support text comprehension among students (Bowen, 1994).

Chapter 8 covers a study where both methods were applied remotely to explore how users make sense of Alexa's data request process, the collected data, and the visualization tool designed to support their sensemaking. While these methods were conducted online due to COVID-19 restrictions, combining them with video calls and screen sharing from participants enabled a comprehensive and detailed data collection process.

Experience Sampling Method

The Experience Sampling Method (ESM) is a research technique used to collect real-time data on participants' experiences and behaviors, typically through prompts delivered via smartphones or other digital devices (F. Fischer and Kleen, 2021). Participants are asked to report their thoughts, feelings, and activities at random intervals throughout the day, providing a detailed and dynamic snapshot of their daily lives (Keedle, Schmied, et al., 2018). By minimizing physical interactions between researchers and participants, ESM

offers significant advantages over traditional fieldwork during the COVID-19 pandemic (E. Kaur, Haghighi, et al., 2020).

In Chapter 9, I discuss using ESM as part of the research process to capture participants' immediate reactions to Alexa's new features. I prompted households via email to document their experiences in text format right after using the feature. This approach enabled me to collect real-time data while minimizing COVID-19 risks. To get a fuller picture of the user experience in specific situations, I combined this method with semi-structured interviews and co-design workshops, as explained earlier.

Role-Playing

Role-playing in HCI is an engaging method for examining and understanding how users interact with intelligent systems. By having participants take on the role of the system, this method offers valuable insights into how users might respond to various system behaviors (Iacucci, Kuutti, and Ranta, 2000). Its flexibility allows the person acting as the system to adjust their responses based on user input, creating a more natural and realistic interaction (Park and Faraway, 1998). This approach enables researchers to explore a wide range of user experiences and system reactions without being restricted by predefined scripts, making it especially useful for prototyping and testing new concepts.

During the COVID-19 pandemic, role-playing proved beneficial as it could be conducted remotely, allowing researchers to gather rich data while following social distancing measures (Barnett, 2020). This method not only helps identify potential usability issues but also promotes creativity and innovation

in the design process by encouraging participants to explore a variety of interaction possibilities.

In Chapter 9, I made this method part of the co-design workshops to keep users more engaged. They stepped into the role of voice assistants, acting out imagined dialogues where the assistant behaved in unexpected ways. Although used remotely, this method allowed me to capture not just verbal responses but also other cues, like users' gestures and facial expressions.

Part II

How Users Make Sense Of AI

Introduction

"*We see the world as 'we' are, not as 'it' is.*"– Anaïs Nin

This part of the dissertation explores how people make sense of AI systems through the lens of **symbolic interactionism**, focusing on the meanings they attribute to AI, its properties, and its capabilities. I take a broad approach here, looking not at specific AI applications but at how users generally conceptualize the term *AI*. To uncover these meanings, participants were given prompts encouraging them to explain AI to a friend—either through definitions or examples—allowing me to explore their meaning-making in two distinct ways:

1. **Intentional definitions** – where participants describe key features of AI (e.g., *"AI is a system that learns"*), use metaphors (e.g., *"AI is like an electric brain"*), or draw analogies (e.g., *"AI is like a tireless assistant"*) (Parry and Hacker, 1991).
2. **Extensional definitions** – where participants provide specific examples they associate with AI, like chatbots, virtual assistants, or self-driving cars (Parry and Hacker, 1991).

I also look at how participants discuss AI within their social circles, asking about conversations they've had with family, friends, and colleagues. These exchanges shed light on how the meaning of AI is socially constructed and shared, reflecting collective understandings and the influences shaping their perceptions.

To explore AI's broader social and global impact, I introduced a thought experiment: imagining a world where AI suddenly disappears. This ex-

ercise revealed just how deeply participants see AI as embedded in—and shaping—various aspects of everyday life.

By examining users' *perceived AI*—and contrasting it with scientific definitions and popular narratives—this study establishes a foundation for the thesis. It illustrates how my research is positioned within the larger conversation about the history and societal understanding of AI.

4

I Don't Know, Is AI Also Used in Airbags?

An Empirical Study of Folk Concepts and People's Expectations of Current and Future Artificial Intelligence

4.1 Abstract

In 1991, researchers at the center for the Learning Sciences of Carnegie Mellon University were confronted with the confusing question of "where is AI?" from users, who were interacting with artificial intelligence (AI) but did not realize it. After three decades of research, we are still facing the same issue with the unclear understanding of AI among people. The lack of mutual understanding and expectations among AI users and designers and the ineffective interactions with AI that result raises the question of "how AI is generally perceived today?" To address this gap, we conducted 50 semi-structured interviews on perception and expectations of AI. Our results revealed that for most, AI is a dazzling concept that ranges from a simple automated device up to a full controlling agent and a self-learning superpower. We explain how these folk concepts shape users' expectations when interacting with AI and envisioning its current and future state.

4.2 Introduction

In 1991, several artificial intelligence (AI)–enabled prototypes of higher educational programs were successfully used at the center for the Learning Sciences of Carnegie Mellon University. Researchers were thus frequently confronted with the puzzling question of "where is AI?" by the confused users, who did not recognize AI in the interaction (Schank, 1987). Schank considered the mismatch between the prototypes and the then-current users' definition of AI to be the main reason for this phenomenon and thus analyzed the different viewpoints on AI, distinguishing among the following four groups (Schank, 1987):

1. AI means magic bullets: A machine is intelligent if it builds unanticipated connections.
2. AI means inference engines: A machine is intelligent if it turns experts' knowledge into rules.
3. AI means exceeding the existing: AI means getting a machine to do something you did not think a machine could do (the "gee whiz" view). A machine is intelligent if it completes a task that no machine has ever accomplished before.
4. AI means machines that can learn: A machine is intelligent if it can learn by itself.

In the past three decades, AI has attracted increasing attention from both researchers and the industry leaders (Gil and Selman, 2019; GoodAI, 2019; Stone, Brooks, et al., 2022; zoovu, 2018). Nevertheless, regarding the public perception of AI, several studies have shown that even today, users do not

realize that they are interacting with an AI-enabled technology (Davies, 2020; Eslami, Rickman, et al., 2015; Eslami, Vaccaro, et al., 2019; Long and Magerko, 2020; Pegasystems, 2019). What makes this the problem worse is that defining AI is confusing, not only for the users with no computational background, but even for the researchers and AI practitioners (Molnar, 2020; Schank, 1987). This is due, among other factors, to the evolution of the term over time and to the fact that it has never represented a single specific technology in a single specific time period (Eslami, Rickman, et al., 2015; GoodAI, 2019). As a result, the current limited perception, misconceptions, and unrealistic expectations of AI (e.g., the superhuman fallacy) (Boden, 2004; Long and Magerko, 2020) not only lead to frustration due to users' unmet desires, but also prevent effective interaction and collaboration with AI (Fast and Horvitz, 2017; Long and Magerko, 2020).

Individuals create informal theories (folk theories) that are not necessarily accurate or realistic in order to be able to perceive how a system works and to interact with it (Eslami, Karahalios, et al., 2016). With the widespread use of recommender systems, collaborative filtering, and personalized services, there is an undeniable need for the active interaction of users with these intelligent agents (Eslami, Karahalios, et al., 2016; Kulesza, Stumpf, et al., 2012). However, the misconceptions and limited folk theories lead to active but not effective interactions and cause frustration among users (Long and Magerko, 2020). A huge body of research has focused on explainable AI (XAI) (Adadi and Berrada, 2018b) to support users' understanding of an AI-based system and combat incomplete mental models (Hoffman, Mueller, et al., 2018; Lim, Dey, and Avrahami, 2009; Ribera and Lapedriza, 2019). However, previous scholars have neglected people's pre-existing folk theories

(Lim, Dey, and Avrahami, 2009), and some have admitted that the results that they reported on users' perceptions were fragmented (B. Zhang and Dafoe, 2019).

To help close this gap, we take a step back and investigate the current public perception of AI. To this end, we consider "users' perceived AI" as an evolving phenomenon and investigate its aspects and dimensions in comparison with the perception of AI held by its designers. By acknowledging the fact that XAI does not shape appropriate mental models, but reshapes those already existing, and by analyzing users' pre-perceptions and folk theories, we will be able to take two steps forward and evaluate the XAI system more effectively.

4.3 Research Background

This study is grounded in three research areas, namely, the previous research on public acceptance and awareness of AI-enabled systems; the design of more explainable AI-systems; and people's folk concepts, theories, and mental models of how a specific AI-enabled system (e.g., intelligent agents or internet of things) works. Therefore, we first outline previous research on the terms related to public understanding of AI, namely, acceptance, trust, and awareness. We then move to clarify the necessity of more explainable systems to support users' interaction with AI, and finally, we argue for the distributed investigation of public informal theories on AI systems within the realm of four AI-enabled systems and services.

4.3.1 Public Acceptance and Awareness of AI

If we take a look at several large-scale reports on public acceptance and trust regarding AI-enabled technologies, we will recognize a trend of increasing positive reactions towards AI. To begin with, in a survey with over 2000 respondents conducted in 2015 by the British Science Association, one-third of the respondents considered the rise of AI to be a danger for mankind (Association, 2016). In 2017, although some researchers indicated that users still felt uncomfortable with AI making decisions on their behalf, the majority of the 27,901 interviewed EU citizens had a generally positive view of robots and AI (Hoffman, Mueller, et al., 2018; Ribera and Lapedriza, 2019). In more recent surveys, we will see even greater support for future AI developments among users (Davies, 2020).

In addition to the terms trust and acceptance, AI awareness has also been discussed in the context of perception of AI. A study from HubSpot of over 1,400 consumers worldwide revealed that 63% claimed that they did not use AI tools while actually using them without being aware of it (Santiago, 2023). Pega-systems confirmed these findings in a study of 5000 consumers, of whom only 34% agreed that they had previously used AI technology but of whom 84% actually had (Pegasystems, 2019). Furthermore, researchers at the Center for the Governance of AI at the University of Oxford found that the majority of users believed that virtual assistants, smart speakers, driverless cars, social robots, and autonomous drones use AI but that Facebook photo tagging, Google Search, Netflix or Amazon recommendations, or Google Translate do not use AI (B. Zhang and Dafoe, 2019). The results from a Northstar survey of 3804 consumers in 2020 also indicate a significant difference in recognition between “invisible AI” (i.e., AI algorithms behind

the scenes) and "visible AI" (i.e., tangible AI devices) (Davies, 2020). Here, 90% of the respondents knew that voice assistants are AI enabled, but almost one-third did not consider social media to be an AI-based technology. Davies argued for the dependency of public awareness of AI on the visibility of its application (Davies, 2020). Nonetheless, their study shows that users do not define AI as humanoid and know the difference between sci-fi and reality (Davies, 2020). By distinguishing between public perception of AI and public acceptance, trust and awareness, we aim to contribute to closing the gap in the literature and to study users' definitions of AI and informal theories on how AI-enabled technologies work.

4.3.2 Designing Explainable Technology

To move towards a more transparent design of technology and address trust concerns (Adadi and Berrada, 2018b), "explainable AI" was introduced as AI systems whose activities can be effortlessly comprehended and examined by humans (Hagras, 2018). Van Lent, Fisher, and Mancuso, 2004 first used the term in 2004 to explain the behavior of the AI-enabled elements in the simulation applications. Over time, the scholarly community and professionals have paid renewed attention to explainable AI (Molnar, 2020).

By refining users' mental models of AI-enabled systems and resolving their misconceptions, XAI promises more effective performance of the users (Ribera and Lapedriza, 2019). However, previous researchers have criticized XAI for resulting in less efficient systems and less flexible and capable output (Molnar, 2020; Ribera and Lapedriza, 2019). To address this issue, XAI researchers have argued that explanations are not always necessary and

instead introduced specific application domains in which they can bring significant benefits (e.g., healthcare and military) (Eslami, Karahalios, et al., 2016). In this regard, Molnar, 2020 introduces human-friendly explanations as selected and focused on the abnormal. In other words, he suggests that people do not expect complete explanations that cover a full list of causes, but rather expect them to explain why a system behaved in a way they did not expect (Molnar, 2020).

We must bear in mind that users were already interacting with AI-enabled systems long before the relatively new trend of XAI was applied to design (Molnar, 2020) and that as the operation of these systems was opaque, they have often developed theories of how they work (folk theories) to plan their interaction with them (DeVito, Gergle, and Birnholtz, 2017; Eslami, Rickman, et al., 2015; Eslami, Vaccaro, et al., 2019). Folk theories function as a frame in which user expectations are formed (Long and Magerko, 2020). Therefore, understanding them is not only helpful in designing effective explanations but is also necessary in order to understand users' assumptions and expectations of the systems and discover more valid explanation use cases.

4.3.3 Making Sense of Technology

To make sense of technology, users often generate folk theories based on direct experiences and social interactions (Eslami, Vaccaro, et al., 2019; Long and Magerko, 2020). These informal theories explain how systems operate and support users in responding to them (DeVito, Gergle, and Birnholtz, 2017). Previous studies have thus argued that folk theories act as a window to capture users' perceptions and assumptions (Eslami, Rickman, et al., 2015;

Eslami, Vaccaro, et al., 2017, 2019). However, the existing studies on mental models and folk theories in the field of AI are highly distributed (Eslami, Rickman, et al., 2015). Hence, by dividing earlier work into different AI domains, we provide an overview of prior research on folk theories and their role in understanding users' perceptions, misconceptions, and expectations.

4.3.4 Content Curation Algorithms and Recommender Systems

Although curating, filtering, and clustering information are processes that are used to shape users' perception of the world, previous studies have suggested that more than half of users are not aware of content curation (Kulesza, Stumpf, et al., 2012). Eslami et al. conducted several studies on users' perceptions and folk theories regarding data curation algorithms (Eslami, Rickman, et al., 2015; Eslami, Vaccaro, et al., 2017, 2019). In a study of 40 Facebook users to understand their perception of the Facebook News Feed curation algorithm, the researchers found that 65% of users were completely unaware of it. They continued their study by investigating the perceptions of Facebook users and discovered very specific folk theories ranging from curation based on users' engagement with other accounts or content to the algorithm balancing their friends or content (Eslami, Vaccaro, et al., 2019). Devito et al. pursued the same aim and analyzed 102,827 tweets from a hashtag related to rumors on algorithmic curation on the Twitter timeline (#RIPTwitter). Their research revealed the abstract and functional folk algorithmic theories that define algorithms as concepts or processes, respectively (DeVito, Gergle, and Birnholtz, 2017).

For their in-the-field learning, personalized agents are also highly dependent on the participation of active users (Kuhl, Lobana, and Meske, 2020; Kulesza, Stumpf, et al., 2012). Kuhl et al. investigated the interplay between users' mental models and users' behavior. Among others, they found that there was no consistent understanding of learning algorithms and users' mental models varied based on their background and experiences (Kuhl, Lobana, and Meske, 2020). In a study of a music recommender system, Kulesza et al. also showed that providing users with structural knowledge on the reasoning of their recommender systems improves the soundness of their mental model and increases their active participation in the interaction in order to receive the desired results (Kulesza, Stumpf, et al., 2012).

4.3.5 Internet of Things (IoT)

Regarding sensor-enabled systems such as the Internet of Things (IoT), users develop folk theories on how this data can be used, rather than on how it is collected in the first place (S. Lee, Kim, and Lee, 2019; Rader and Slaker, 2017). Rader et al. showed that users explain their activity based on their perception of their performed activities and the processed data provided by the interface (Rader and Slaker, 2017). Although users entered data (e.g., age and weight) themselves (so that distance and calories were calculated), they did not have a full understanding of their raw data. Therefore, they could not make informed decisions about the collection of their personal data (S. Lee, Kim, and Lee, 2019).

Misconceptions may lead to unawareness of privacy risks. This is particularly true when information is collected without consent, as in the case of Blue-

tooth Beacon Systems (Yao, Huang, and Wang, 2019). Beacons were part of an invisible IoT infrastructure in which direct user interactions were needed to enter personal data (e.g., location).

Yao et al. used drawing as a methodology to reveal folk theories on how beacon-based systems work (Yao, Huang, and Wang, 2019). Here, folk theories focused more on the reasoning of the visible data collection. For example, the most common misconception was that these sensors collect and store user information, which they actually did not. Furthermore, users assumed that they initially needed to actively consent to receive location-based information. Finally, most of the users thought that the location such as the store or mall that collected the data also owned the data. As a result, users who decided, consciously or unconsciously, to be part of the sensor-enabled system had little understanding of the actual mechanisms of data collection and processing.

4.3.6 Robots

Humans tend to anthropomorphize computational artifacts to rationalize actions and behavior that they cannot reasonably explain. Inaccurate mental models frequently deceive people, who as a consequence credit autonomous system with more capability and knowledge than they actually have (Phillips, Ososky, et al., 2011). Those mental models are influenced by appearance and physical attributes (Kiesler and Goetz, 2002), dialogue, personality traits, language, and origin (Phillips, Ososky, et al., 2011) and lack a clear understanding of mechanical and conceptual functionality. These attributes affect the credibility of robots. Powers and Kiesler showed that the robots'

facial features, such as the dimensions of the forehead and chin, impact the perception of intelligence (A. Powers and Kiesler, 2006).

Hence, people developed high expectations that were not fulfilled and harmed their collaboration and relationship building, which require trust in autonomous systems. Anthropomorphism can support users in their approach and reactions to the robot and prevent initial rejection. However, it might also lead to misconceptions, which have detrimental consequences for the outcome and even endanger human lives, such as in military settings or critical workspaces (Phillips, Ososky, et al., 2011). Therefore, humans need a clear and accurate understanding beyond anthropomorphism of how robots collect and process data and make decisions.

4.3.7 Conversational Agents

Besides anthropomorphism of conversational agents (CA) (Kuzminykh, Sun, et al., 2020), an extensive study in HCI explored voice Interaction and the adoption of domestic CA. In particular, some studies have investigated the perception and understanding of the attributed intelligence (Druga, Williams, et al., 2017; Garg and Sengupta, 2020; Kuzminykh, Sun, et al., 2020; S. Lee, Kim, and Lee, 2019; Y. Xu and Warschauer, 2020). Researchers used methods of drawing (Kuzminykh, Sun, et al., 2020; S. Lee, Kim, and Lee, 2019; Y. Xu and Warschauer, 2020), interactive tasks (Druga, Williams, et al., 2017; Garg and Sengupta, 2020; Kuzminykh, Sun, et al., 2020), and interviews (Druga, Williams, et al., 2017; Garg and Sengupta, 2020) to explore the reasoning and explanations of system behavior that children of different ages create.

Xu and Warschauer showed that children used behavioral references like listening and talking to justify cognitive properties, reciprocity as an indicator of psychological properties, and biological references like mechanical causality or fantasy reasoning to explain behavioral properties (Y. Xu and Warschauer, 2020). They tended to allocate CA rather to a continuum between humans and artifacts than to a distinct category. Especially younger children who were less aware of the underlying concepts attempted to make sense of computational artifacts by personifying them. When children acknowledged the devices to be more intelligent than themselves (Garg and Sengupta, 2020; Y. Xu and Warschauer, 2020), they were more likely to trust and believe the information provided, which contributed to improved learning (Garg and Sengupta, 2020).

To examine the state of intelligence or consciousness, the study of Druga et al. indicated that younger children asked personal questions such as "what is your favorite color?" and older children asked the device to perform actions of which they knew humans were capable (Druga, Williams, et al., 2017). With prior experience with technology or an engineering background, more thoughtful reasoning was applied (S. Lee, Kim, and Lee, 2019). Nonetheless, voice and tone effected the perceived friendliness (Druga, Williams, et al., 2017), and available input modalities communicated the expected level of intelligence (Druga, Williams, et al., 2017; Garg and Sengupta, 2020). To enable meaningful interaction and collaboration, design decisions for the influencing factors (S. Lee, Kim, and Lee, 2019) should aim to meet users' expectations or communicate the actual capabilities of CA transparently.

4.4 Research Setting

In this section, we explain our research method and analysis technique. We start by presenting our research questions and data acquisition approach, whereafter we introduce our research sample and the implementation of the research method in practice. Finally, we explain our data coding and analysis technique.

As stressed by Blumer, "people act towards things (such as physical objects, people as well as abstract ideas) on the basis of meanings they ascribe to them" (Blumer, 1986). Hence, if we want to understand how and why people interact with AI, we should also understand what meanings people ascribe to AI. We call these meanings folk concepts, as this term expresses the folk- or ethno-conceptualization of the subject matter. This concept is closely related to the idea of folk theories (Eslami, Rickman, et al., 2015; Kulesza, Stumpf, et al., 2012), but we try to avoid the latter term to avoid the impression that folk concepts are reflected, systematic, and consistent theories analogous in their logic to scientific theories. Folk concepts serve more as sensitizing concepts used as a resource to frame and make sense when interpreting and interacting with or talking about something. Our study had two primary research questions:

- RQ1: What meanings do people ascribe to AI?
- RQ2: What expectations do people have when they ascribe AI to something?

When it comes to folk theories and mental models, as well as people's assumptions about how a complex system works, in-depth interviews have previously shown a great potential (e.g., Eslami, Karahalios, et al., 2016;

Eslami, Vaccaro, et al., 2019; Ngo, Kunkel, and Ziegler, 2020; Yao, Huang, and Wang, 2019). Hence, with the aim of understanding people's perceptions of AI, we applied a semi-structured interview method to gather a rich dataset and establish a research framework towards user-centered AI. The data corpus was collected within an undergraduate course on explainable AI, in which we asked the students to conduct interviews using a semi-structured guideline that was collectively developed. Along with the demographic information, the guideline included the four following questions:

• A friend of yours asks you what AI is: How would you explain AI to him/her?

Concepts can be described in various ways. A common method is to provide an intentional definition (Parry and Hacker, 1991), in which a concept is approached by listing essential features (e.g., a car has an engine), characterizing it by metaphors (e.g., a car is an iron-horse), or drawing analogies (e.g., the car is like a carriage). The aim of the question above was to elicit such an intentional definition from participants expressing their understandings of AI.

• Can you name three examples of artificial intelligence?

Another way to explain a concept is to provide an extensional definition (Parry and Hacker, 1991), naming examples that represent the concept. The question above aimed to provoke such extensional definitions. In the case of a complex issue, it is sometimes easier for people to explain a concept with the aid of an example. In addition, as normally there is no label such as "this is AI" or "AI inside," the answers to this question help to understand what people perceive as AI from their own perspective.

• Have you ever had conversations with others (family/friends/colleagues) about AI? What did you talk about?

As mentioned by Blumer, the meaning "is derived from, or arises out of, the social interaction that one has with others and the society" (Blumer, 1986). This view is also reflected in Battarbee's notion of *co-experience* (Battarbee, 2003), which "takes place as experiences are created together or shared with others." As they stress, "People find certain experiences worth sharing and 'lift them up' to shared attention" (Battarbee, 2003). With this question, we aimed to shed light on the co-experience of AI, especially what people find so exciting about the idea of AI that makes it worth elevating and sharing with others.

• If we tell you that AI is embedded into these objects, what would you expect from them? a) a door; b) a bank account.

When we interact with something new, we apprehend the new based on our existing idea and apply the new to this idea. Garfinkel used this necessity in his breaching experiment to make ethno-methods and social norms, typically implicit and taken for granted, visible (H. Garfinkel, 2023). In this spirit, we also wanted to elicit and provoke deeper, generative folk concepts of AI by asking the participant what happens if AI were to be embodied in two known things. Although access control (either to abstract data or physical properties) plays a crucial role in both of the examples, by choosing one visible and familiar object and one rather abstract concept, we tried to compare people's expectations based on the visibility of AI and on previous experiences with it. To elaborate, people have already experienced automatic doors and could relate more easily to the concept of a smart home, whereas a smart bank account leaves more space for imagination.

• If we told you that AI will disappear tomorrow, what do you think the consequences would be?

One of the main reasons that people do not recognize AI in technologies is that AI technologies have already ingrained themselves in our everyday life. Therefore, there is a good chance that we are already so integrated into the world of AI that we take it for granted. To grasp people's understanding of the role of AI in everyday technologies, we confronted them with an imaginary scenario, in which AI vanishes from our world. We tried to formulate and ask the question in a neutral matter, so that the participants felt free to take a position on the current and future impact of AI in their life.

To ensure the quality of our outcome, we trained the students to conduct semi-structured, in-depth interviews and explained the idea behind the convenience sampling method (Acharya, Prakash, et al., 2013). They interviewed people in their social network who had experience with "smart" technologies such as voice assistants, fitness trackers, social media, robo-advisors and recommender systems (e.g., in YouTube or Netflix).

Each student conducted three interviews, but we did not consider 25 of them due to the low quality (e.g., strongly noisy recordings, suggestive and closed questions, lack of participants' motivation to elaborate on their ideas, etc.). Overall, the final data corpus consists of in-depth interviews with 50 people (28 M, 22 F). The majority of the participants were between 20 and 30 years old (min 19, max 57, age average 28), had an educational background (min. B.A.), and were familiar with the daily use of technology. The interviews were conducted remotely (either via video or audio call) and took 20–30 minutes. With the consent of the participants, the interviews were audio-recorded and transcribed.

Our analysis was based on the German interviews, and we used MaxQDA as a supportive tool. By taking a thematic analysis approach (J. A. Smith, 2024), each interview was in-vivo coded and analyzed by two different researchers. The codes were then discussed in interpretation sessions using a narrative analysis approach. We translated the statement into English only to quote them in this paper.

4.5 Findings

In this section, we provide the results of our data coding and analysis from four different aspects, namely, AI's conceptualizations, competencies, future and current impacts and expectations related to it.

4.5.1 What is AI: A Thing, Characteristic, or Set of Instructions?

Conversational openings are of special interest, as they provide the context framing for what follows (Flick, 2013; Schegloff, 1968). Metaphorically speaking, the opening prepares the stage on or the guard rails within which the narrative can unfold. Therefore, when asking for respondents' definitions of AI, we particularly noted the initial sequence.

Most of the participants chose a relatively broad and vague framing (such as referring to AI as something, a thing, a system, or an entity), while some were more focused, describing AI as a device, a brain, or a computer. These ambiguities, vagueness, and difficulties in the definition of the term do not

indicate a restricted or defective understanding on the part of the respondents, but are, in our view, well founded. Artificial intelligence is something that only presents itself to the user through its embodiment and its performance; it exists not as its behavior itself, but rather in what that has produced the behavior. Normally, this was described as a property or competency of the entity. However, some mentioned program code, detaching this property from the concrete thing, meaning that the computer itself does not have AI, but that AI manifests only with the correct program, which itself is not AI either. The difficulties people faced locating AI were reflected in their framing of what AI was.

Our brief analysis shows that providing a clear, uniform answer to the question of "What is AI" was tricky for the respondents. Hence, we also asked them to define AI extensionally by giving three examples of it. As expected, participants mainly focused on the thing-ness of AI, naming tangible devices, while the more abstract concept of AI as a specific characteristic or specific instruction were less prominent. For instance, machine learning was rarely listed as an example of AI, whereas Amazon's Alexa, robots, and self-driving cars, followed by the visible but not tangible examples such as Twitter-bots and recommendations in Amazon or Google, were often mentioned.

4.5.2 On the Competencies of AI: From Automation, to Simulation and Superpower

Despite the vagueness of these openings, they provided the respondents with certain frames to which certain competencies could be assigned. Based on

these, we categorized the understanding of AI's capabilities into six groups, shown in Table 4.1.

Tab. 4.1.: Folk concepts of the meaning of AI and the identified categories.

Characteristic	Folk concept	Examples
Automation	AI can support humans by fulfilling a pre-programmed task	"A machine that takes over our routine tasks." (P21)
Agency	AI can work independently and make decisions for humans	"AI can make decisions independently and does not need humans to make decisions for it." (P47)
Like humans	AI can act and react like humans	"A computer that acts human." (P38)
Simulating humans	AI can simulate human competencies	"Simulation of humans' intelligence through machines" (P16)
Superpower	AI can improve human intelligence, knows everything, and make predictions that surpass human knowledge	"AI is the creation of machines that surpass the intelligence of humans." (P37)
Self-learning	AI can learn and develop by itself	"You feed the computer with one million cat pictures among others. Thus, AI learns what a cat looks like." (P27)

- **AI is automated:** Like P21 (see Table 4.1), many participants described AI as a technology that operates autonomously with the purpose of fulfilling the interests and ideas of its developers. This view underlines the superiority of humans by emphasizing the never-ending need for human support for AI's existence and functioning. The view also emphasizes the artificial nature of AI, placing AI in the long history of technical development and understanding it as similar to the other everyday devices with which we interact. Artificial intelligence does not

have its own agency and cannot work independently. "*AI is dependent on the implementation of the programs*" (P25).

- **AI has agency:** The quote from P47 (see 4.1) indicates the understanding of AI as autonomous in a peculiar way. This becomes evident when we split the phrase into two parts. In the first part, P47 describes AI as a system that makes decisions for humans. We observed several variants of this phrase, such as that AI supports humans in everyday life, AI provides services for humans, etc. In the second part, however, P47 stresses that AI does not need humans and is independent of them, which points out the complex relationship between AI and humans. Artificial intelligence is not just independent, but independent of humans. Various participants also mentioned that independency is not a future stage of AI, but its current status.
- **AI is just like humans:** Participants often described AI with reference to humans. To understand the folk concepts about AI, an understanding of the folk concepts of human beings is thus a prerequisite. This is notable as, for instance, common-sense descriptions of natural things (such as trees, mountains, planets, etc.) as well as technical things (such as houses, cars, toasters, etc.) typically do not involve such a reference. Here, human beings are referred to as a model for AI, which is an image of human beings and in order to conceive of which one needs knowledge of what human beings are. From this standpoint, a technology can be seen as AI if it can pass the Turing Test: "*AI has the same behavior as humans and acts and reacts at the same level*" (P46). Some participants characterize AI with regard to specific capabilities or

properties of humans: "*AI means systems that act like humans regarding learning and understanding others*" (P44).

- **AI learns by itself:** Half of the participants addressed the competency of AI to learn and improve by itself. It was remarkable that, although we did not ask interviewees about how AI learns by itself directly, they explored this topic themselves. A closer look allows us to analyze this folk concept along three interconnected issues: How learning works, what the sources for learning are, and what the results of learning are:
  - *Learning source and strategy:* There are thousands of ways to learn; for example, you can read a book, think about a problem, or become wise through experience. As in the phrase of P27 (See Table 4.1), most of the interview partners had a computational understanding, in which AI learns from data. Furthermore, some participants, such as P27, pointed out that AI can process large amounts of data, while some did not mention data itself, but patterns as sources of learning: "*Computers analyze the patterns of human faces in such great detail and in such large quantities*" (P31). In addition to technical vocabulary, many descriptions contained implicit and explicit analogies to human learning, for example, when terms are used, such as learning from experience, learning from mistakes, or learning from the analysis of facts. Such kind of anthropo- or zoo-amorphization can be found in phrases such as "*AI uses previous experiences as a basis for shaping connections, learns through experiences and makes connections (example: needle pricking pain)*" (P7). In this phrase, AI is thought of as a small child or a rat in a Skinner experimental setup that learns by operant

conditioning. In a similar vein, P10, P15, and P17 referred to the common strategy of learning from mistakes: "*The program itself learns from errors and improves itself*" (P10). Sometimes the interviews simply referred to the ability of AI to learn without explaining how this learning takes place: "*AI is the ability to learn, conclude*" (P43).

  - *Learning outcomes*: "*They can recognize faces relatively well and reproduce them digitally*" (P31). Unlike learning strategies and sources, learning outcomes were only mentioned vaguely and could not be easily categorized into groups. We can approximately divide descriptions to those in which the learning goal and scope are quite precisely defined, those in which they are quite broadly described, and those in which they are undefined. The first category includes description in which specific goals such as speech recognition and image recognition are used as paradigmatic examples to explain AI learning. The second group includes descriptions in which AI learning is only described in very general terms, such as "*AI learns through experience and makes connections*" (P7) or "*AI is a group of intensities that can learn from data and produce the appropriate output from input and continue to improve.*" (P1). In such a folk concept, it seems that specific goals or purposes play a small role whether a machine can be considered self-learning. Instead, learning seems to present a general competency of AI that is in principle not limited to a specific area.

- **AI simulates human intelligence:** As the Turing test also suggests, the intent and ultimate goal of AI is to reach a human-level in its

performances. "*AI attempts to transfer human intelligence to machines to promote their learning ability*" (P23). Here, AI is perceived as an infant in its early development, moving towards reaching the human level someday: "*AI is a digital brain in a computer that learns behavior like a toddler*" (P18). Therefore, unlike the responders with the "*just like humans*" point of view, P18 does not believe that AI can act and react like humans or pass the Turing test, but rather is approaching achievement of these goals in the future.

- **AI is a superpower:** Various participants, such as P37 (see Table 4.1), assumed that AI has a super-intelligence that surpasses that of humans. This superintelligence not only knows the answer of any possible inquiry but also can predict the future: "*AI draws conclusions from past events about what will happen in the future and predicts beyond our knowledge*" (P20). One reason that AI is understood as a superpower is its unlimited data processing and storage capacity, seen, for example, in the phrase of P4: "*Artificial intelligence is when algorithms have unlimited knowledge*" (P4).

  Super intelligence is also grounded in the super-learning competency of AI, which was mentioned by the respondents who perceived AI not as a predefined code, but as a system that can generate unexpected results: "*AI means self-learning systems that generate outputs through the corresponding inputs, which are not expected*" (P1). The idea of a superpower places high demands on AI, so that one can ask whether today's systems meet these requirements. P46 posed this question and answered that these capabilities do not exist in the today's technologies:

"*AI? It does not exist yet. We call them smart but they are not really smart*" (P46).

4.5.3 On the Expectations of People: When Things Become AI

The question about what happens when things become AI helps us shed more light into the current beliefs of people and uncover how people apply their often-implicit ideas of AI and what their latent expectations are. We grouped the answers into four categories (see Table 4.2). The analysis reveals that the folk expectations of AI are shaped by the folk concepts about what AI is (see Table 4.1). Yet, how the concepts are enacted and interpreted depends on the concrete scenario and application area.

- **Automation:** Half of the respondents mentioned that they would expect a smart door to automatically open and close (one representative is P4; see Table 4.2), and two believed that this feature alone is an indicator of AI: "*I expect that the door opens and closes automatically*" (P50). Six others added some sensors to their automatic door and introduced the intelligent door in this way: "*It should recognize the movement and closes and opens based on its recognition*" (P20). Therefore, we can say that some users do not see any differences between intelligent and automatic devices. P23 solved this problem by introducing different levels of intelligence: "*The most primitive is the motion sensor, which detects when you approach the door. One step further would be image/face recognition. On the next level, the door could be connected to a smart home system, so the system could detect an intruder in the house*" (P23).

Tab. 4.2.: Folk concepts of an AI-enabled objects and the identified categories.

Expectations	Folk Concept	Smart Door	Smart Bank Account
Automation	AI-enabled means working automatically	"Automatic opening and closing of the door or based on time." (P4)	"It should pay my bills automatically." (P5)
Agency	AI-enabled means having the agency to make decisions independently	"The door should be able to receive parcels." (P11)	"I would expect a stop on spending, so that the artificial intelligence says, now you have spent up to this amount, please don't continue." (P29)
Personalized services	AI-enabled leads to the capability of personalized interaction	"The door should clearly be personalized for me." (P30)	"It can learn my behavior and adjust its services for example by finding the ideal investment." (P19)
Smart security	AI-enabled means protecting users' security in a smart way	"The door should keep me safe from the dangers outside." (P15)	"Security of my money is the most important thing for me. So, the trust in AI must be stablished first." (P34)
Intelligent user recognition	AI-enabled means recognizing people in a smart way	"The door recognizes people and only those who are known are allowed to open it." (P21)	"AI can be embedded, so that I can withdraw money with face recognition." (P3)

Interestingly, in the example of the bank account, only three respondents mentioned the automatic payment of the bills: "*A bank account, that can buy things independently or transfer money*" (P27). This might be due to the higher risk that our participants associate with the monetary services: "*A bank that helps me save and transfers my monthly repeated bills automatically*" (P36).

- **Agency:** Regarding intelligent doors, only two respondents (P47 and P11; see Table 4.2) expect more comprehensive agency beyond a smart

automatic door opening. P47 described how "*The door should perceive its environment and react accordingly; for example, if needed, it should make less noise while opening and closing*" (P47).

Although agency did not play an important role in users' common expectations from a door, it was the most common expectation they held of an intelligent bank account. It is worth mentioning that some users struggled with whether or not the embedded AI in their bank account should be trusted to have its own agency. Therefore, they had different expectations of the levels of control that the bank account would possess.

- *Improved comfort:* AI was seen as smart enough to eliminate the need for manual authentication: "*No bank card is required. For example, with a chip in your hand you can access your account*" (P37).
- *Replace service worker for simple tasks:* The participants said that AI could start to compete with bank employees and take over simple tasks: "*The possibility to open new accounts without having to communicate with the bank employee*" (P43).
- *Replace service worker for complex tasks:* The same was true for complex tasks. Those most commonly mentioned were the bank account serving as personal investor and financial advisor. Here, the smart bank account is able to consult and suggest a financial approach to the user but is not allowed to make decisions on its own: "*The smart bank account should evaluate my expenditure,*

suggest a consultancy and support me with my account management" (P9).

  - *Making decisions automatically for the client:* The fourth level covers the expectations in which the bank account makes decisions on its own. For instance, it prevents users from spending money irrationally by not permitting such transactions (P29; see Table 4.2). On this level, the AI is transparent with the users and lets them know why the decision was made. This decision might not be aligned with the users' present interests but would fulfill users' long-term goals.

  - *Making decisions automatically even against the client:* Users placed AI in a position in which it has full control, making decisions even against their interest. For instance, P13 envisions a scenario in which AI has full agency for better or worse: "*For example if the account often has a negative balance, this could affect the creditworthiness of the users and they cannot receive loans*" (P13).

- **Personalized services**: In the case of the bank account, about one-third (16 of 50) expected personized financial services. For instance, the bank account was expected to play the role of a financial coach by analyzing transactions and payment patterns and providing users with plans for saving or investment: "*A bank account can over time make recommendations for potential savings or investment opportunities*" (P17). Others also mentioned improved feedback about spending patterns: "*It analyses transactions on a monthly/yearly basis and evaluates how much money was spent on what and whether it differs at certain intervals*" (P23). In the case of the intelligent door, this rate was much lower.

Only four participants mentioned to expect personalized services from their intelligent door. P1 and P7 explained these personalized services as recognition of the behavioral pattern/schedules and customization of the door accordingly: "*The door learns habits and can recognize through behavior patterns whether to lock or unlock*" (P1). In addition, P9 also expected an intelligent door to begin communicating with people in a manner similar to that of a friendly porter: "*I would expect a personal greeting*" (P9).

- **Smart Security:** Security was among the most common expectations for our participants. Although it made a greater contribution to the example of the smart bank account, four of the respondents also expected it from the intelligent door: "*Security is the top priority here, the door should not only be protected against robbery. It should also be able to detect if an attempt is made to hack it, because anything that has to do with technology can be hacked and is very vulnerable*" (P33). P15 (see Table 4.2) did not address the danger clearly, but in general. Here again, the expectations from a bank account were higher.

  Two participants even expressed their reluctance to answer the question because of distrust in AI: "*If AI is embedded in a bank account, then I will have problem with trusting it*" (P16) and "*In a bank account, I want AI to be embedded as low as possible*" (P25). P34 (see Table 4.2) addressed this problem by setting a condition for a smart bank account. He believed that before anything else, a smart bank account should establish trust. Some others focused on the new opportunities that a smart bank account can provide to detect phishing attacks and protect against them: "*A smart bank account can keep my bank details safe*

and can detect whether a website is fake or not when I want to make a transaction" (P46).

- **Intelligent user recognition:** The importance of security also covers access control, authentication, and authorization. In the example of the door, more than half of the users mentioned face or finger recognition for access as the expectation. This group includes P21, whose idea can be seen on Table 4.2. For a smart bank account, however, only five users mentioned expecting this capability. This might be because of two reasons. For a door, unlike a bank account, face/voice recognition is not a common feature, as many bank applications already provide this feature. Therefore, one can conclude that users expect a smart device to be capable of more uncommon actions. However, as security played the most important role in users' concepts of a smart bank account, one can think of another possible reason, namely, that the users are still skeptical about this feature and would regard a smart bank account as too intelligent to risk the security of its users in exchange for offering more usability to them.

  As we saw from the results, a higher priority was placed on some attributes such as automation and face/voice recognition for the smart door, while others such as agency and security were of a greater importance for the smart bank account. One reason might be the previous experiences of the particular participant with the object and the possible difficulties encountered. Some of the participants, such as P13, even referred to previous experiences with the object: "*When I think about a door and artificial intelligence, a revolving door immediately comes into my head, as is so often the case in large shopping malls that*

peoples' speed automatically slows down and vice versa" (P13). In general, the expectations from a smart bank account were higher than those of a smart door, which in most of the cases would only automatically open and close. This might be due to the abstract nature of a bank account, to which users can ascribe a human role of coach or adviser more easily than to a physical object such as a door, which is obviously not a human.

4.5.4 On the Impact of AI: If AI Disappears ...

When asked what would happen if AI were to vanish tomorrow, the respondents had to think of the technologies and services in which AI was embedded meaningfully. Our observations showed that it was a complicated matter for the respondents, some of whom, such as P30, even asked the interviewer whether they were right about AI use cases: *"AI makes our everyday lives easier, for example, for driving, we have driver assistants, I don't know, is AI also used in airbags?*" (P30).

Our analysis revealed that this complexity pushed respondents towards envisioning the impact of the loss of AI mainly based on the inputs they had from personal conversations, as well as from mass and social media. We categorized the mentioned impacts of AI into three following groups:

- **Social impact:** *"Disappearance of AI will cause a chaos in the world. There is definitely more AI in everything than we think*" (P39). Several participants thought that AI has spread into all areas of life, so that the loss would lead to severe dislocation. P44, who took a seminar about AI at university and usually talked about the accuracy and simplicity of

the roles in which AI has taken over, stated, "*AI is boundless. I believe a life without AI is without development, it may not be the primitive life, but of course, there will be negative effects on all areas of life like engineering, medicine, education, agriculture, etc.*" (P44).

- **Economic impact:** "*If AI disappears then the economy will get paralyzed and we cannot have a normal life anymore*" (P18). Some participants, such as P18, stress that the disappearance of AI would have serious economic consequences. P1, for instance, stressed that AI is important for the digital revolution in Germany with regard to the economic progress: "*Economic harm because companies like Amazon cannot use their recommender system anymore*" (P1).
- **Personal impact:** "*I cannot imagine a life without AI. I ask myself, how we could drive a long distance without navigation 20 years ago*" (P42). For participants like P42, who cannot even imagine the disappearance of AI, the spectrum of the personal impacts ranges from extreme cuts of personal living to small or even no personal changes: "*There won't be great personal consequences, expect for Siri and Alexa*" (P2). P5 detached his personal impact from the impacts on others, saying, "*There will be no consequences for average people, only those who have smart homes*" (P5).

Do not have AI for better or worse: The respondents typically also expressed a position about the role of AI in today's world and chose whether to perceive it as a positive or a negative one. In this regard, we identified four positions:

– *Deplore life without AI*: "*If AI vanishes that would mean that our work and our life will be more difficult*" (P50). About 60% would deplore a life without AI. A representative of this view is P50, who, believed that AI makes life easier: "*If we want to make things easier, we should find more and more applications for AI in them*" (P50). The extent of the regret also depends on how AI is understood: "*So, if I lose my smart phone today and get back to the old phones without internet connection, then I will lose 20% of my own identity and 80% of my social relations*" (P46).

– *AI is Janus-faced*: "*The fate of humans in the midst of this great upturn in science is terrifying*" (P43). Less than 10% mentioned both negative and positive sides. A representative of this position is P43, who noted, "*There are positive and negative sides. Natural disasters happen all over the world, AI saves humanity by improving the methods and strengthening the means to cope with them. But if AI disappears, then the unemployment rate will decrease*" (P43). This quote highlights the complex relationship between humans and AI: humans as the creators of AI, a superpower that begins to compete with them.

  This view was shared by others, who considered AI to be a sign of a brighter future that comes with a risk. There were also other double evaluations due to the consideration of AI in different areas. For instance, P6 differentiated between the social and the personal areas: "*Without AI, the world would fall several years into the past. But for me personally, I will have more personal interactions and the social relationships won't be digital like now*" (P6).

– *Welcoming life without AI:* More than 10% of the participants were skeptical about AI and had some serious concerns and thus welcomed life without AI. One example is P26, who had several discussions with her husband about AI and the fear of being manipulated by it: "*If AI vanishes, I will feel relived, although many things would be done manually again*" (P26). Similarly, P22 had a conversation with her daughter about the latter's homework on the ethical issues of AI in the area of self-driving cars: "*If there is no AI tomorrow then the ethical issues, we have about the self-driving cars will be solved*" (P22). The extreme case was presented by P47, who referred to AI in absolutely negative terms: "*If AI vanished, the most important consequence would be more jobs for everyone*" (P47).

– *Being incurious about AI:* There was also a group (about 20%) of the participants who were incurious, as they found AI to be not essential. They either believed that the disappearance of AI would bring about no specific changes or that the changes would be exceedingly small. A representative of this group was P23, who mentioned that many find AI not to be a relevant subject and avoid discuss it: "*It would be a loss if we don't have this technology anymore, but it's not essential. Because until now, there aren't any huge dependencies to AI. The economy can grow further with or without AI*" (P23).

4.6 Discussion

In this segment, we discuss the essence of our results with regards to their potential to support AI practitioners and XAI researchers in achieving more effective and meaningful interactions of the users with AI. We thus outline users' perceptions and expectations of AI through the identified folk concepts and the set of user demands. Based on our case study, we also discuss the possible reasons that AI has remained imperceptible to its users after three decades.

4.6.1 Meaning Ascribed to AI: The Interconnected Folk Concepts

Regarding our first research question, examining what meaning people ascribe to AI, our study updates the 30-year-old study of Schank (Schank, 1987). Most of the definitions of AI noted by Schank were also among the folk concepts that we identified. For instance, users' definition of AI as magic bullets is closely connected to the idea of AI being a superpower. In addition, the view of AI in terms of learning machines was highly prominent in our study. Complementing the work of Schank, our research shows that people have different views regarding the sources, strategies, and outcomes of the self-learning AI systems. This ranges from a highly technical understanding of machine learning to metaphorical notions and analogies with human learning. In contrast to Schank's findings, the understanding of AI as an inference engine to turn experts' knowledge into rules played only a subordinate role in our findings. This might reflect the prevailing concerns of the time when

Schank wrote versus today, for while expert systems were an important topic 30 years ago, they are rarely present in today's AI discourse.

When asked for three examples of AI, our respondents referred to the visible and tangible artefacts (such as Alexa, robots, and self-driving cars), which confirms the results from the research done by Davies, 2020 and shows that people recognize visible artefacts as AI more easily. However, this result should be handled carefully, because here, the thing-ness of the conceptions of AI seems to be shaped by the nature of the question. In other words, when the question was formulated as "how do you explain AI?" the participants did not provide an extensional definition, but instead more abstract and conceptional answers (such as AI having agency, simulating human competencies, or being a set of instructions).

Despite having a relatively homogenous sample group (on average, representatives of a young generation of educated German users), our respondents had different understandings of AI and its capabilities. In this regard, our study confirms the previous work done on specific AI systems (such as content curation algorithms, recommender systems, and conversational agents) and the variety of folk theories and misconceptions reported by them (Eslami, Karahalios, et al., 2016; Eslami, Rickman, et al., 2015; Kulesza, Stumpf, et al., 2012; M. K. Lee, 2018; Y. Xu and Warschauer, 2020; Yao, Huang, and Wang, 2019). Our study indicates that mental models about specific AI systems are not isolated but shaped by common folk concepts and the grand narratives of AI that machines become superhuman. Moreover, we argue, that folk concepts are interconnected and build upon each other. To elaborate further, people's perception of AI and AI-enabled systems influences

their understanding of its meaningful and essential use cases that helps them evaluate AI's current and future impacts in their lives.

An example of this interconnectivity can be seen in characterization of the folk concept by the complicated relationship between man and AI. Almost all explanations defined AI with a reference to humans (such as like humans, simulates humans, is independent of humans, is made by humans, etc.). Metaphorically, AI was frequently spoken of in terms of adolescent development, strongly influenced by parents and oriented towards them, while also turning away from them and claiming independence. This complicated relationship leads to a complex cooperation between the two, in which the role each plays is not clear for the users. This complex relationship also pervades the evaluation of AI as a technology that is part of our everyday lives.

4.6.2 Expectation Ascribed to AI: Design for the Unrealizable?

Design for user expectations presents an important guiding principle in usability engineering (ISO 9241-11:2018(en), 2018). However, our study shows that people have quite diverse and high expectations of the AI-enabled artefacts such as "*AI has the same behavior as humans and acts and reacts at the same level*" and ambitious expectations such as "*a door that recognizes me and communicates with me based on its knowledge about me*" that surpass existing technological capabilities. This phenomenon suggests that people do not have appropriate knowledge when it comes to implementation, which can lead to unrealizable expectations of technology and disappointment

when these expectations are not met. This may explain the phenomenon described in the literature of why users are dissatisfied with voice assistance systems such as Alexa, Google Home, and Apple Home pod (Bentley, Luvogt, et al., 2018; Luger and Sellen, 2016; Porcheron, Fischer, et al., 2018; Sciuto, Saini, et al., 2018): although these systems are useful for simple tasks such as turning on the light and playing a music track and work reliably, they cannot meet the high expectations of users.

Therefore, in our next study, we aim to identify this gap between expectations and possibility by using more profound research techniques to grasp folk concepts, theories and mental models of AI. Moreover, based on our study, we recommend that explainable AI should be transparent not only in how the system works, but also in what a user can expect from it. We should also take seriously the insight from research that user experience is shaped not only by the actual interaction, but also occurs before, during, and after use (Law, Roto, et al., 2009). From this perspective, realistic expectation must be made a part of the advertising of AI-enabled products and services, as well as of the AI-related narratives in science fiction literature and movies.

One other ambiguity for the respondents regarding the current state of AI concerns its users. Is AI only used by big high-tech companies and infrastructure? By tech savvies who live in smart homes and possess self-driving cars? Is it used by the average consumer? Or can anyone be considered an AI user? To address their lack of knowledge in this area, the respondents referred to whatever input they had gathered from others. This might also be the reason for the widely diverging evaluations of AI's impact and current meaningful use cases among people, which ranged from wider areas such as

the economy, medicine, politics, transportation, and so on to narrower use cases, such as in voice assistants or gaming.

4.7 Limitations and Future Work

Our work has certain limitations that should be considered and that can motivate future work. One major limitation regards the representativeness of our sample group and the known problems of self-selection bias. Hence, our findings especially express the folk concepts of a young, well-educated, urban generation of digital natives in western society, in Germany. Less can be said about the folk concepts of elderly people, as well as folk concepts in other societies. This can be improved by using a larger sample with more diversity in terms of technical background and age group, as well as by conducting a similar study in other countries. Deeper and more detailed direct questions could also be used to identify the barrier between automation and simulation or self-learning sources and strategies as well as the extent to which users' engagement in social conversations impacts their perception and evaluations of the AI's meaningful use cases.

Therefore, in future, we aim to investigate these folk concepts and theories along with their interconnections and impact, as well as other influencing factors (e.g., sci-fi and social media trends) in greater depth. Our study also encourages the implementation of a more user-friendly XAI by focusing on clear communication of what users should expect from AI-enabled systems. Such transparent AI systems can help us analyze how users' expectations can be influenced and modified for more satisfying interactions.

4.8 Conclusion

In this work, we emphasized the importance of understanding AI as it is perceived by people in designing efficient and productive explainable AI systems. By outlining the fragmented previous work on people's mental models and folk theories and analyzing the results of our in-depth interviews with 50 participants, we presented a thematic analysis that systemized the various perspectives and expectations of people regarding AI and AI-enabled systems. Our results, based on a relatively homogenous sample group of mainly young and well-educated people, suggest that users' perceptions of the current and future state of AI are not purely dependent on their computational background, but are shaped by various interconnected folk concepts of AI's definition, capabilities, and meaningful use cases.

Like previous studies (e.g., Ribera and Lapedriza, 2019), we also argue that explanation should not only address users' needs, but should be adjusted to their understanding. Hence, the design for XAI cannot be done without knowing how people perceive AI and should make what they can expect from it transparent. We also argue that without modifying users' expectations of the AI-enabled systems, designing for users' satisfaction will be a losing battle, and encourage XAI practicians to aim for clearer communication in this regard.

Part III

AI-Assisted Decision Making

Introduction

> "*Trust is a perception, not a fact. It differs from person to person and moment to moment.*" – David Amerland

In this part of my research, I focus on **AI-assisted decision making systems** in high-risk scenarios—like healthcare, job applications, and loan approvals. These are everyday encounters where trusting the system becomes especially important because the decisions can have serious consequences, adding to users' sense of uncertainty and vulnerability.

I specifically explore **trust** as a key characteristic of these systems, using Lee and See's definition of trust as *"an attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability."* (J. Lee and See, 2004). This definition highlights three important aspects of trust:

1. **Vulnerability**—the risks people face when relying on AI,
2. **Positive expectations**—believing the system will work as intended,
3. **Trust as an attitude**—a mindset, not just a behavior.

To better understand how trust takes shape in these systems, I focus on two key groups: **decision subjects** (the individuals affected by AI decisions) and **AI practitioners** (those designing and implementing the systems). This focus on trust, as experienced and understood by these two groups, highlights the importance of context. It shows that trust in AI systems isn't universal—it emerges in specific situations and is shaped by the roles people play within the system. By taking this multi-stakeholder approach, I explore how users

and practitioners make sense of trust when the stakes are high, and decisions matter most.

.

5

Trust in AI-assisted Decision Making

Perspectives from Those Behind the System and Those for Whom the Decision is Made

5.1 abstract

Trust between humans and AI in the context of decision-making has acquired an important role in public policy, research and industry. In this context, Human-AI Trust has often been tackled from the lens of cognitive science and psychology, but lacks insights from the stakeholders involved. In this paper, we conducted semi-structured interviews with 7 AI practitioners and 7 decision subjects from various decision domains. We found that 1) interviewees identified the prerequisites for the existence of trust and distinguish trust from trustworthiness, reliance, and compliance; 2) trust in AI-integrated systems is strongly influenced by other human actors, more than the system's features; 3) the role of Human-AI trust factors is stakeholder-dependent. These results provide clues for the design of Human-AI interactions in which trust plays a major role, as well as outline new research directions in Human-AI Trust.

5.2 Introduction

Decision making assisted by artificial intelligence (AI) has become more widespread in high-stakes domains, where decisions have real impacts on people's lives such as public safety (Kalhan, 2013), hiring (Ajunwa, Friedler, et al., 2016) or loan approval (O'Dwyer, 2018). Typically, the AI-based systems considered are based on automated processes (such as data-driven machine learning techniques) that provide assistance to human decision makers in a form of recommendations. Because Human-AI trust plays an important role in the adoption of these technologies (Hoff and Bashir, 2015) and the improvement of decision making (Bansal, Nushi, et al., 2019), it has become a priority for their design and development, as well deployment and regulation (on AI (AI HLEG), 2019). To understand how to achieve appropriate levels of human trust in these systems, more research at the intersection of Human-Computer Interaction and social study of AI is needed.

Trust is a complex and multifaceted concept (Lewicki and Brinsfield, 2011; Lyon, Möllering, and Saunders, 2015) and several studies have focused on a better understanding of the factors that can affect Human-AI trust (e.g., Ma, Lei, et al., 2023; Ooge, Kato, and Verbert, 2022; X. Wang and Yin, 2023; Yin, Wortman Vaughan, and Wallach, 2019; Y. Zhang, Liao, and Bellamy, 2020). In these studies, trust is predominantly investigated through the lens of users, who are the persons interacting with the AI-assisted decision making system and its recommendations in order to deliver their decision (Lai, Chen, et al., 2023). Less is known about the perspectives from which other stakeholders involved in, and impacted by the design, deployment and use of these systems, view the notion of trust in AI, while

this outlook on Human-AI trust can be shaped by their role. Jakesch, Buçinca, et al. (2022) demonstrate that the ethical values embedded in AI-assisted decision making systems can hold varied significance and interpretations for different groups. For example, people working on AI, on average, considered responsible AI values less important than general public and crowdworkers that contributed to the training of AI models. Such differences might also be reflected in the understanding of and opinions on Human-AI trust of the various stakeholders. For example, Lockey, Gillespie, et al. (2021) identify that different types of users do not encounter the same issues related to Human-AI trust: trust in AI of domain experts, e.g., doctors in medical decision-making, might be particularly affected by the factors that challenge their professional knowledge, skills, identity, and reputation. In contrast, fairness-related factors might impact general users' and society's trust in AI. Therefore, examining how stakeholders other than AI users view the definitions and factors of Human-AI is essential to advance the understanding of how trust is accounted for in the development and design of AI-embedded systems assisting decision-making and whether the existing approaches match the varying needs of different stakeholders.

In this article, we investigate how two groups of stakeholders - AI practitioners and decision subjects - understand Human-AI trust. Exploring the views of these groups on Human-AI trust definitions and factors allows to understand to which extent they prioritize and value the same aspects of Human-AI trust.

AI practitioners are involved in system design and deployment (from AI developers to project managers). Given that they make decisions that influence the shape of human-AI interaction, impacting trust in AI-based technology

and its acceptance for decision-making, understanding AI practitioners' views on trust can shed light on the factors they prioritize to build trust in AI among different stakeholders. Therefore, we explore the following first two research questions: RQ1a) *According to AI practitioners, what are the critical elements of human-AI trust in decision-making?*; RQ1b) *What do AI practitioners think influences the trust of various stakeholders in AI in the context of decision-making?*

The second group is **decision subjects**, i.e., people who do not interact directly with the systems incorporating AI but are affected by the decisions made by users based on the recommendations of these systems. For example, doctors are users, and patients are decision subjects in the medical context. Although decision subjects do not generally interact with AI-based systems the same way as users, they may nevertheless want to decide whether or not they wish to be impacted by the system (Gemalmaz and Yin, 2022). For example, if a patient decides that the doctor's AI-based recommendation is not fair or trustworthy, they may want to change doctors or clinics. Therefore, we are investigating the following research questions: RQ2a) *According to decision subjects, what are the critical elements of trust between humans and AI in decision-making?*; RQ2b) *What factors influence decision subjects' trust in AI?*

We thus conducted semi-structured interviews with 7 AI practitioners related to AI-assisted decision making and 7 decision subjects from various risk-sensitive contexts (finance, law, management, medicine). The questions revolved around defining trust and trustworthiness when related to AI, and what they think can affect Human-AI trust. Using thematic analysis (Braun and Clarke, 2012a; Clarke and Braun, 2013), we established three themes:

1) definition of trust through three prerequisite elements and differentiation from other related concepts. The interviewees define Human-AI trust similarly to the literature with vulnerability and positive expectations, and additionally propose task complexity as a trust prerequisite. Moreover, AI practitioners distinguish trust from trustworthiness and trust-related behaviors such as reliance and compliance; 2) the effect of relationships between various stakeholders on Human-AI trust. We found that the extent to which decision subjects, AI practitioners, and users trust each other has an impact on their trust towards AI and can moderate the effect of some factors on Human-AI trust; 3) stakeholder-dependency of the role and effects of some factors on Human-AI trust. We found that AI transparency, AI literacy, and interactivity of the system affect Human-AI trust differently for different stakeholders. Based on our findings, we provide a set of implications for academic researchers in HCI and AI practitioners. In particular, we recommend investigating the breaking and calibration points of trust between humans and AI beyond direct interaction with the system, and re-examining the techno-centric trust factors between humans and AI from a socio-technical point of view, as well as from the point of view of stakeholders other than users.

5.3 Related Work

In this section, we provide a brief overview of the methods to study Human-AI trust in assisted decision making and the different stakeholders at play with such systems.

5.3.1 Background on AI-embedded Systems Assisting Decision Making

While there is no universally accepted definition of AI (Duan, Edwards, and Dwivedi, 2019), in this paper, we follow the definition provided by the European Commission: AI is a system capable of "*perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal*" (Samoili, Lopez, et al., 2020). Therefore, what we refer to as "AI-embedded systems assisting decision making" are the systems that analyse data to derive information used to facilitate human decision making (Castelluccia and Le Métaye, 2019; Schmude, Koesten, et al., 2023). Usually, such systems provide assistance to human decision makers in a form of one or multiple recommendations, and when the system is not fully automated, it is the human who has the last word while making decisions. If the AI's recommendation differs from the decision maker's initial opinion, the decision maker finds themselves in conflict between her initial opinion and the new information received, which means that she has to choose between their opinion and the recommendation in order to make a better decision (Yaniv and Kleinberger, 2000).

Making a better decision based on a recommendation means to be able to interpret the quality of the recommendation. However, it can sometimes be difficult to understand how a system arrived to a certain conclusion due to their "black box" nature (Adadi and Berrada, 2018a; Lagioia and Contissa, 2020). This, in turn, obfuscates understanding why a certain AI recommendation was produced, anticipating potential biases in decision

Icon	Acronym	Stakeholder	Definition
	U	Users	Individuals directly interacting with the system
	P	AI practitioners	Individuals who design, develop and deploy AI-based solutions
	DS	Decision subjects	Individuals affected by an AI-assisted decision-making system

Tab. 5.1.: The different stakeholders related to the AI assisted decision making systems. This article focuses on AI practitioners and Decision subjects, two stakeholders who received less attention in the Human-AI trust literature.

making, and identifying the reasons for wrong predictions (Scherer, 2015; Yu and Kohane, 2019). When one is uncertain about how to correctly assess the quality of a recommendation (Sniezek and Van Swol, 2001), one can rely on their level of **trust** towards it to decide whether to stick to one's own opinion or to follow the system (Van Swol and Sniezek, 2005; X. Wang and Du, 2018). As AI-embedded systems are becoming more widespread for assisting in making decisions have real impacts on people's lives, such as public safety (Kalhan, 2013), hiring (Ajunwa, Friedler, et al., 2016) or loan approval (O'Dwyer, 2018), to name a few, the need for considering what contributes to human trust in the design of AI has arisen (Board, 2019; Commission, 2020; Fund, 2019; G20, 2019; KPMG, 2019; of Science and Policy, 2020; Services, 2019; Taskforce, 2019; Villani, Bonnet, Rondepierre, et al., 2018).

5.3.2 Human-AI Trust

Human-AI trust literature has two major themes of interest: defining what trust is and what factors affect it. The first line of research builds primarily on theoretical works, (e.g., Jacovi, Marasović, et al., 2021), taking a top-down approach to understanding Human-AI trust. A systematic literature review on Human-AI trust in the decision-making (Vereschak, Bailly, and Carami-

aux, 2021) defines Human-AI trust through three prerequisite elements, all encompassed in the trust **definition** by (J. Lee and See, 2004): *An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability*. These three prerequisites are: vulnerability (or risk) of humans to the actions of the AI-based system, positive expectations of humans with respect to the AI-based system outcomes, and attitude as opposed to a behavior. Some scholars further define more granular facets of trust such as affective and cognitive trust (Lewis and Weigert, 1985; McAllister, 1995), weak and strong trust (Baratella, Amaral, et al., forthcoming), warranted and unwarranted trust (Jacovi, Marasović, et al., 2021) or differentiate between trust in a particular AI tool, in people who built this tool and in AI in general (Ryan, 2020).

The second major theme investigates Human-AI trust through a bottom-up approach, empirically studying what **factors** can affect users' trust. Glikson and Woolley (2020) in a literature review of studies empirically investigating Human-AI trust factors identify that the main ones for trust in AI are: tangibility, transparency, performance (reliability), task characteristics, anthropomorphism, and socially-oriented behaviors of the system. While they did not propose any classification of the factors, almost all of them belong to a category related to the system characteristics (a category present in trust frameworks from the fields other than Human-AI interaction (B. D. Adams, Bruyn, et al., 2003; Bindewald, Rusnock, and Miller, 2018; P. A. Hancock, Billings, et al., 2011; Hoff and Bashir, 2015; Schaefer, Billings, et al., 2014; Schaefer, Chen, et al., 2016).

Type of task is the only trust factor that is related to the context of interaction, rather than the interaction with the system itself. Another framework

on Human-AI trust factors in the medical context Browne et al. calls for expanding the current literature's focus on the contexts other than users' interaction with the system (Browne, Bakker, et al., 2022). The authors argue that considering trust factors in the contexts beyond use reflects better the entire clinical AI deployment process in the real settings and, thus, opens up the floor to new trust calibration points (Browne, Bakker, et al., 2022).

As most of the work on Human-AI trust targets a single type of stakeholder - direct users of the systems, we expand the analysis of Human-AI trust definition and factors to the stakeholders other than users that are related to the Human-AI decision making systems.

5.3.3 Human-AI Trust and Stakeholders Other Than Users

In this article, we focused on stakeholders that are the most linked to the development or the use of AI-assisted decision making systems: **AI practitioners**, people who develop the systems; **users**, people who use these systems to make decisions; and **decision subjects**, people who are affected by those decisions (see Table 5.1). Additional stakeholders, however, exist, such as regulators and policy makers, whose contributions, although interesting, are out of the scope of this paper (the reader can refer to different taxonomies (e.g., Ayling and Chapman, 2022; Deshpande and Sharp, 2022; Güngör, 2020; Jakesch, Buçinca, et al., 2022; Scott, Carter, and Coiera, 2021; Yurrita, Murray-Rust, et al., 2022) for more information).

The stakeholders that have received the most attention in the literature on Human-AI trust are the **users** of the systems (Lai, Chen, et al., 2023). Researchers have repeatedly pointed to the need to explore and assess users' trust in these systems to facilitate their adoption (see, for instance, (Bisantz and Seong, 2001; Schöffer, Machowski, and Kühl, 2021; Seong, Bisantz, and Bisantz, 2002). It is not surprising that the literature focuses on system users, as understanding what affects their trust in the AI algorithms embedded in these systems can inform the development of interfaces and interactions that would facilitate the emergence of trust. However, different stakeholders may have different needs, expectations or roles when it comes to trust between humans and AI, and may also have an implicit impact on users' trust in systems. The research on AI with human-centered values has investigated the perspectives and needs stakeholders other than users, notably AI practitioners (e.g., Ashtari, Mullins, et al., 2023; Deng, Yildirim, et al., 2023; J. Y. Kim, Boag, et al., 2023; Varanasi and Goyal, 2023; Veale, Van Kleek, and Binns, 2018; Yildirim, Pushkarna, et al., 2023) and decision subjects (e.g., Gemalmaz and Yin, 2022; Lima, Grgić-Hlača, et al., 2022; Lyons, Miller, and Velloso, 2023; Lyons, Wijenayake, et al., 2022; Marian, 2023; Yurrita, Draws, et al., 2023).

Here we first present a set of previous works that have explicitly demonstrated differences between these stakeholders when it comes to concepts such as AI ethics, explanations, or fairness. While these results are not directly about Human-AI trust, they are related to our domain and motivate our approach.

Regarding responsible AI, which aims to deploy AI-based systems in line with ethical and legal frameworks, previous work shows the importance

Id	Role	Background	Organization	Type of AI	AI Application
P1	XAI R&D	CS and Maths	Large	CNNs	Transport, paleontology
P2	XAI R&D	Eng. and Maths	Small	OR	Task planning
P3	CEO	Maths	Small	Supervised ML	Evaluation of law cases
P4	Research mgr.	HCI	Large	OR, supervised and unsupervised ML	Project-based
P5	Research mgr.	Human Factors	Large	Not specified	Project-based
P6	CPO	Engineering	Small	ML (not specified)	Finance and business
P7	CEO	Bio. Eng. & Research	Small	Deep learning	Medical

Tab. 5.2.: Characterization of AI practitioners, their companies, and AI they work with as reported by the interviewees themselves. "Small" refers to the companies with less than 20 employees, "Large" - with over 1000 employees. Explanation for abbreviations: *XAI* - explainable AI, *R&D* - research and development, *mgr.* - manager, *CEO* - chief executive officer, *CPO* - chief product officer, *CS* - computer science, *eng.* - engineering, *CNNs* - convolutional neural networks, *OR* - operations research, *ML* - machine learning.

of including the AI practitioners' perspectives to ensure that the system is designed to meet the actual needs of business and industry (Holstein, Wortman Vaughan, et al., 2019). Typically, *AI practitioners* are pushed to quickly develop a service or a product that one can sell, which sometimes conflicts with ethical practices valued by the *users* (Ali, Christin, et al., 2023; Madaio, Egede, et al., 2022; Rakova, Yang, et al., 2021; Widder, Zhen, et al., 2023). Regarding Explainable AI (XAI), which aims to propose means to explain AI-based predictions and help their interpretation, the usefulness of the explanation of a recommendation given by AI can vary depending on who sees it (Ferrario and Loi, 2022; Schmude, Koesten, et al., 2023). A *user* might want to learn to which extent a recommendation can help them save money for instance (Marabelli and Newell, 2019), while *decision subjects* might want to know to which extent this recommendation is biased against a certain population in which they may belong (A. Brown, Chouldechova, et al., 2019; Woodruff, Fox, et al., 2018). Regarding fairness, (J. J. Smith, Buhayh, et al., 2023) take the case of microlending and show that depending on the different strategies to achieve fairness, stemming from its different definitions,

Human-AI decisions favor *decision subjects*, *direct users* or *the organization behind the system*. Finally, regarding power relations in interaction, *users* and *AI practitioners* might see AI recommendations as tools, assistants or servants (T. Kim, Molina, et al., 2023), while *decision subjects* might see the same AI recommendations as coming from someone in a more powerful position than they are. Such difference in perceived hierarchical roles between different stakeholders and AI can influence their attitude towards the system and interaction with it (Cabitza, Campagner, et al., 2023; Y. T.-Y. Hou, Lee, and Jung, 2023; Saxena, Moon, et al., 2023).

Previous work thus demonstrates the importance to study different stakeholders in the context of Human-AI interaction. In the context of Human-AI trust, Passi and Jackson (2018) investigate how AI practitioners establish trust among themselves while working with data. Ammitzbøll Flügge, Hildebrandt, and Møller (2021) and Okolo, Kamath, et al. (2021) emphasize the importance of trust between users and decision subjects. Ferrario and Loi (2022) analyze the importance of XAI for decision subjects' trust in AI. Lastly, Ramesh, Kameswaran, et al. (2022) show that decision subjects overtrust AI due to seeing it as a higher authority for financial decisions. These works tend to focus on a small set of factors influencing Human-AI trust. A more global perspective of how AI practitioners and decision subjects build and perceive Human-AI trust is yet to be explored.

5.4 Methodology

We adopted an interview-based qualitative methodology to answer our research questions about what trust is and what that trust depends on in the

context of AI-assisted decision-making from the perspective of the real-word stakeholders. The project started in 2021.

5.4.1 Participants

We recruited participants through a convenience sampling technique combined with snowballing among colleagues and friends, and through announcements at events and on the project's social media channels. We had two selection criteria to find interview participants: 1) they either work (as practitioners) on AI-embedded systems that support risk-sensitive decision making (e.g., in health, law, finance)[4] or they have been affected by their decisions (as decision subjects), 2) the system is used in the real world. We did not focus on any particular corporate position nor on any specific AI application in order to obtain a diversity of perspectives among interviewees. In total, we conducted 14 semi-structured interviews (7 with AI practitioners[5], 7 with AI decision subjects).

Id	Background	Decision Context
DS1	Software developer	Job application
DS2	Medical student	Access to services
DS3	Mechanical engineer	Job application
DS4	Business economics researcher	Loan application
DS5	Mechanical engineer	Job application
DS6	Accounting and project management	Job application
DS7	Computer engineer	Job application

Tab. 5.3.: Characterization of decision subjects, notably their background and in what context they received a Human-AI decision.

[4]Risk in risk-sensitive applications is understood as defined by the European Union (EU) regulatory framework proposal on AI: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[5]We initially contacted 14 AI practitioners, 5 of them did not reply, and 2 did not have availability for an interview

	AI practitioners	Decision Subjects
Context	*How would you describe your role? What is the main objective of your system?*	*Could you please tell me about your experience with Human-AI decision making?*
Trust Definition	*How would you define Human-AI trust in your own words? How would you define Trustworthy AI in your own words?*	
Trust Factors	*What is your strategy to establish trust of various stakeholders in your AI?*	*Have you ever trusted AI too much / too little?*
Trust Evaluation	*How would you know if someone trusts your AI?*	*Do you think AI developers consider human trust?*

Tab. 5.4.: Structure and examples of questions per each group of participants. Data analysis of this paper mostly relies on answers around the definitions and factors of Human-AI trust. A full list of questions is in Appendices A and B.

The participation in the study was on a voluntary basis. The AI practitioners are based in Europe and Oceania, and each worked for a different company. Table 5.2 provides an overview of the AI practitioners' backgrounds, their roles in the company, and the application areas of AI. Three participants work on XAI (two are responsible for implementation and research, and another is the company's chief executive officer - CEO). Three other participants are senior project and product managers.

The AI decision subjects are all based in Europe and had been affected by AI decision making in three different risk-sensitive areas: job application, access to services, loan application. The decision subjects we interviewed were not the people affected by the AI tools developed by the AI practitioners who participated in our study. Table 5.3 provides an overview of the decision subjects' backgrounds and in what context they received a Human-AI decision. Although 4 out of the 7 interviewed decision subjects have a background in computer science and engineering, we did not explicitly evaluate their prior experience with AI or their level of expertise in the field.

5.4.2 Interview Protocol

We conducted semi-structured interviews (Magaldi and Berler, 2020) of the recruited participants. The questions were compiled by the two first authors. They were independently reviewed by the other two authors and approved by the ethics committee of the research institution. In addition, we conducted a mock interview with an AI practitioner and a decision subject and adjusted the wording of the questions to improve their understanding. These data were not used for analysis. The questions were designed in English and translated to French and German for those participants preferring one of these languages. Interviews took place either by telephone or videoconference, whichever participants preferred. Participants could choose to allow us to record the interviews for note-taking purposes. All 14 participants agreed to do so. A total of 685 minutes were recorded, and each interview lasted an average of 50 minutes. Participants had access to our written notes before we used them in the article to ensure that their anonymity was maintained. All participants allowed us to quote them in the study.

The interview protocol consisted of four parts (Table 5.4) evolving around: the context with respect to their interaction with AI, Human-AI trust definitions, trust factors, and trust evaluation. In this article, we focused on the data regarding definitions and factors in the analysis. Where possible, we kept the formulation of questions identical (see Trust Definition in Table 5.4) for both groups of the participants. We adjusted the formulation of the questions related to the personal experiences to reflect the role of each group (example in Trust Factors, Table 5.4). We asked the questions around the Human-AI and trustworthiness definitions to understand what participants consider to be prerequisites of trust, that is in what contexts it

is appropriate to consider Human-AI trust. We asked AI practitioners about their strategies to establish trust in their AI tool to understand what factors AI practitioners think influence trust of other stakeholders and which factors and stakeholders they prioritize. We did not explicitly refer to any group of stakeholders in our questions to let the AI practitioners spontaneously name the stakeholders relevant to the discussions around Human-AI trust. We asked decision subjects to share their experiences with receiving Human-AI decisions and, notably, what made them trust these decisions to identify the factors that influence their trust in AI. We also wanted to know whether decision subjects thought they trusted these decisions or AI in general too much or too little to gain more insights about what factors they prioritized to calibrate their trust towards more appropriate levels.

There were 8 questions in total as approximate guidance for the interviewers (Appendices A and B). When needed, we deepened the topic with follow-up questions about all the stakeholders involved in an anecdote, clarifying theoretical terminology, possible solutions to a described challenge, and whether a proposed factor always has effect on Human-AI trust.

5.4.3 Analysis of the Interviews

The first and second authors transcribed all interviews, removed all personal information (name of team, company, city, etc.) from the text, and assigned a code name to each interviewee, **P** for AI practitioners and **DS** for decision subjects. After transcription, the researchers deleted the audio files and allowed participants to review the interview text if they wished. The two researchers also translated the French and German texts to English and

validated the translation with native speakers of the respective languages. Subsequently, the two researchers independently read all interviews at least twice, first without taking any notes and the second time highlighting the phrases or words related to people's experiences and needs with AI, to get familiarized with the data.

The further data analysis was based on the inductive thematic analysis (Braun and Clarke, 2012a; Clarke and Braun, 2013), that is a bottom-up approach to coding and analysis driven by the data itself. The two authors independently assigned to each highlighted phrase a code that encapsulates the best its main message, focusing on the semantic content of the data. They then compared the list of highlighted phrases and their codes, discussed whether to include or not the phrases highlighted only by one of the researchers, and fine-tuned the wording of the codes for the finalized list of the selected phrases. After three iterations, the first author organized the codes in a series of sub-themes. They were further reformulated or merged with the consensus of all four authors in the process of writing the paper, and organized, under three main themes: one on the definition of trust, one on the role of interpersonal relations, and one on the divergent opinions between AI practitioners and decision subjects on the factors affecting trust (further described in the next section).

5.5 Findings

The thematic analysis yielded three main themes discussed in this section. We first explore the definitions of Human-AI trust from the perspectives of AI practitioners and decision subjects. Secondly, we find that both groups

of respondents attribute significant importance to trust in interpersonal relationships, rather than in system characteristics. We conclude the results section by emphasizing some differences in opinions between the groups regarding the impact of AI transparency, AI literacy, and interactivity on Human-AI trust.

5.5.1 On the Definition of Trust

When prompted to define Human-AI trust in decision making, the interviewees identified three prerequisites for trust: positive expectations that AI will be beneficial in achieving the goals, the perceived risk associated with a decision, and the complexity of the task at hand. Importantly, the interviewees differentiated between trust, trust-related behavior, and trustworthiness.

Positive expectations and perceived risk are prerequisites for the emergence of trust, but the nature of risk is debated.

The interviewees state that for trust to emerge, people must have **positive expectations** that AI will help them achieve their goal and is aligned with their interest. They defined goal as *"the best answer in the shortest time"* (DS5, DS7). P6 also highlights that AI recommendations must be aligned with the goal of people interacting with or affected by the system as opposed to the technology owner's interest: *"It is important that the owner [of an AI-embedded system] does not recommend something in the company's interest"* (P6).

Moreover, the interviewees refer to the **perceived risk associated with a decision** as another prerequisite for the emergence of trust: *"When my*

physical integrity or money is at risk, trust becomes a consideration, especially when something important is at stake for me" (P4)[6]. Several participants associate risk with health (DS2, DS4, DS5, DS6) or financial stability (DS4, DS5, DS6). P4 refers to risks related to economic loss or threats to life and health as universal, stating, *"... a foundation [for defining risk] would be the physical needs and individual and social integrity from the Maslow's Hierarchy."* However, some, like P5 and P2, broaden the concept of risk to include *"vulnerability"* (P5) or *"responsibility"* (P2), showing that risk extends beyond just financial or health concerns. P4 notes that what is considered risky varies from person to person, as *"not everyone has the same priorities"*. For example, DS4 found even Tinder recommendations could induce vulnerability, recounting moments when *"the algorithm says that I am ugly, something about myself that I do not want to accept"* (DS4). Therefore, DS4's experience of feeling vulnerable when their appearance was judged by AI indicates that the associated risk goes beyond monetary losses or health hazards and is closely related to one's personal vulnerabilities and priorities. This points to the situatedness of the risks involved and suggests that the mere application domain of the AI-assisted decision is not enough to indicate the level of risk associated; rather, it is the perceived risk based on individual vulnerabilities and priorities that matters.

[6]In this quote, AI practitioner P4 refers to their general reflection about what could trigger one's trust to emerge, not taking a particular perspective as an AI practitioner nor a decision subject

Task complexity as a new prerequisite for the emergence of Human-AI trust.

Besides positive expectations and perceived risk as prerequisites for human trust in AI to emerge, some interviewees (P2, P4-P6, DS5) also mention **task complexity**. P2 and P6 describe "complex task" as a situation when a person cannot determine the quality of AI recommendation and, as a result, has many doubts around the final decision. DS5 agrees with P2 and P6, citing data analysis as an example of a task that is complex because: *"it is very difficult for a human to perform calculations and test the system."* A task is also perceived as more complex if the decision to make is a long-term one (P4). P5 suggests that when users face a complex task, trust emerges as a tool to mitigate the complexity: *"Sometimes you can't evaluate everything, you sort of use that quick «I just trust you, I just trust you to do the right thing»."* Interestingly, while the interviewees reported task complexity as one prerequisite for trust, it is not present in the usual definitions of trust Jacovi, Marasović, et al., 2021; Vereschak, Bailly, and Caramiaux, 2021, which typically considers two prerequisites: "positive expectations" and "vulnerability".

Trust is differentiated from trust-related behaviors and trustworthiness.

Some interviewees differentiate between trust (which is defined as an attitude Jacovi, Marasović, et al., 2021) and trust-related behaviors. For instance P4, P5, and P6 postulate that inferring users' level of trust in AI from simply observing their behaviors could be misleading. Because users *"can have a complex and elaborate way of thinking [about AI-embedded systems and recommendations]"* (P4). P3 indicates that a user might follow AI recom-

mendations not out of trust, but because they *"have no other solutions"* (P3). The interviewees thus clarify that it is trust-related behaviors, not trust itself, that are in action. But trust-related behaviors are useful as they can serve as *"indicators"* of trust. As P2 notes, *"as long as there aren't too many complaints, no negative comments, [...] and the user uses the solutions, we can consider that trust is not broken"* (P2).

Additionally, four interviewees (P2, P4, P5, P7) explicitly differentiate trust in AI from AI trustworthiness. Contrary to trust, which is seen as *"human reaction"* (P5), trustworthiness relates to features of the system (P2, P5), e.g., *"whether the job has been well done"* in designing and developing the system (P7). Such distinction further supports the stance that it is important to focus not only on what makes AI trustworthy, but also on what makes people trust AI Q. Liao and Sundar, 2022. Interestingly, two interviewees associated trustworthiness with AI governance, i.e., the institution or organization behind the AI. P4 states: *"For me, it [trustworthiness] is not so much a question of AI, it's more between the individual and the entity or the organization that makes the system."*

5.5.2 The Role of Inter-personal Relations on Trust

We discovered that trust between humans and AI is influenced by the trust among various stakeholders involved in the creation, use, and evaluation of the AI-based decision support system. Furthermore, AI certification, as a potential solution, is also contingent on trust within an interpersonal relational network.

The team behind AI plays an important role in (Human-AI) trust

The interviewees indicated that an individual's trust in AI is closely linked to the level of trust they have in other stakeholders (Human-Human trust) within the socio-technical ecosystem. This ecosystem includes interactions between various stakeholders (such as AI practitioners, users, and decision subjects) and even the technical characteristics of the system. We found four cases illustrated in Figure 5.1. AI practitioners (P2, P4, P7) particularly emphasize Case 1: trust between the **users and the AI team**, where the AI team is the group of people behind the creation of the system. If users trust the AI team, their trust in AI "*[...] is established before the system exists. [...] Trust is very strong in the co-design phase [between users and the AI team]*" (P4). Interestingly, previous work on Human-AI trust does not generally consider trust between users and the AI team as a factor in Human-AI trust, even though it is likely to occur in real-world scenarios.

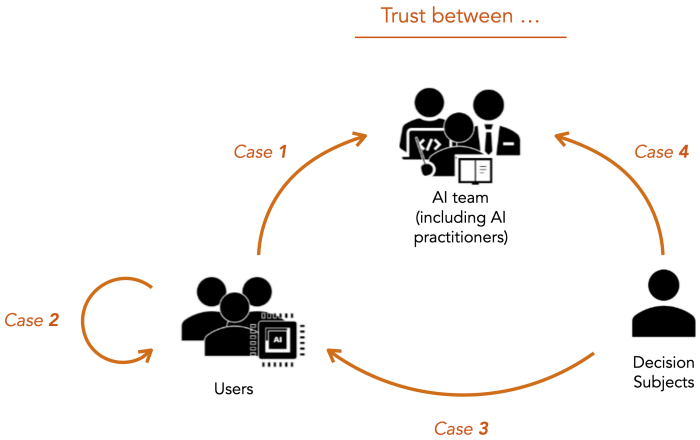


Fig. 5.1.: Schematic representation of the extent to which trust between different stakeholders groups discussed in relation to how it can affect Human-AI trust in the context of decision making.

AI practitioners (P3 and P6) also talk about the trust between the **users and other users** of the same system (Case 2). They claim that previous experiences of other users influence users' trust in AI: *"We have 10,000 users, and 90% of them say «the feedback from the AI was very interesting», now [knowing this, current users] will tend to trust the AI"* (P6). This trust in AI is further strengthened if *"a domain expert confirms what the AI recommends"* (P6).

Case 3 examines the trust that **decision subjects** place in the **users** of the system, concerning their usage and purpose. For instance, DS3 noted that trust in the system and its users are intertwined: *"There is trust in the system and trust in those who use the system [...]. They [the users] should at least tell you they are using such a system [embedding AI] so you don't lose your chance, just because you don't know how it works [...]"*. DS7 highlighted the complexity of this trust dynamic involving both humans (users) and machines from the perspective of the person being impacted by the decision: *"I don't trust mixing humans and machines. Either the decision should be entirely made by a machine or a human. If you have only one machine, then you know what to expect. But if you have a machine and a human, then it would be very unfair because the users' roles are not defined, and the priority is not clear.*"

Finally, one **decision subject** also talked about the role of **trust in AI team** for decision subjects (Case 4). DS7 cites the example of Elon Musk and Tesla (at the time of the interviews), explaining that the trust of decision subjects in the company's high-level management influence their perceptions of and trust in the AI systems they develop: *"[he] is building trust with people through his own presence in the media [...]. People trust him and love his personality, so they trust his product even if it does not benefit them in the end."*

The effect of AI certification on Human-AI trust depends on who is behind it.

The interviewees (P1, P4, P6, DS1-DS3, DS5, DS7) share the view that knowing that an AI system has been certified is a factor that influences trust in that system, because *"certification has always been a way to gain confidence in technological tools, whether they are AI [or not]"* (P6). This is especially true for critical systems: "*The objective is clear - we [AI team] want certification*" (P1). P4 says that "*the certification alone should be enough [for Human-AI trust] if it is done well*." However, some interviewees highlight the importance of who is behind the certification, rather than the sole fact of AI having been certified (DS1-DS3, DS5, DS7): *"AI certificates are very important [for Human-AI trust] if there are organizations [that issue them] that people can trust"* (DS2).

Finally, P5 and DS7 are more suspicious about certification in general because they think there is not yet enough scientific evidence that *"certification will build trust [in AI], I am not quite convinced of that yet"* (P5) or because a certification does not warrant that everything will be alright *"if there is a hack or a problem"* (DS7).

5.5.3 Diverging Opinions on Three Factors impacting Human-AI trust

Three factors playing a role for Human-AI trust were considered differently by AI practitioners and decision subjects, and are important to be highlighted. These factors are: AI transparency, AI literacy, and the increase of interactivity

on the system. We decided to focus on them, rather the factors that both groups of stakeholders agreed on (e.g. AI performance and errors, marketing of the system, expectations about the system), as we believe that the identified diverging opinions provide interesting insights and implications for the research community.

AI practitioners and decision subjects do not share the same view on the role of AI transparency on trust.

AI transparency is one of the most discussed Human-AI trust factors in the interviews (P1-P7, DS3, DS4, DS7). The interviewees define two levels of transparency: a) explaining why a specific **AI recommendation** was shown and its quality, and b) explaining the **working processes** of AI development team.

The opinions about the effect of **explanations of AI recommendations** (a) on trust diverges not only between decision subjects and AI practitioners, but also among AI practitioners themselves. Some AI practitioners believe that explaining why a specific recommendation was shown can affect Human-AI trust, because it provides better understanding of how the recommendation was derived and, thus, lets estimate recommendation's quality (P2, P3, P6). At the same time, P4 strongly questions the necessity of understating for trust: *"One has to stop wondering how one can make tools that are more explainable, interpretable, or whatever, because sometimes there are tools that are not explainable in which we trust, a plane or a car, we don't know how it works inside, and yet we use them [...]"* (P4). Additionally, P1 and P7 raise concerns about the extent to which explanations can contribute to one's understanding of an AI recommendation: *"All the latest methods [of explainability] that have been developed are often so complex that humans [laypeople] do not*

understand them, so the methods do not help them at all" (P1). Decision subjects further disagree with the AI practitioners supporting usefulness of AI explanations (P2, P3, P6). They state that besides AI explanations being complex (DS3, DS4), they have limited contribution to understanding of AI recommendations and, consequently, trust, because of the real world constraints: *"If people had the time to go through the explanations and review them in practice, they would have made the decision themselves in the first place"* (DS7).

Additionally, the AI practitioners (P3, P4, P7) seem to put considerable importance on transparency around the **working processes** (b) of AI development team, while decision subjects did not mention this aspect of transparency at all in connection to trust. The AI practitioners believe that the working process is the most actionable means of AI transparency for their clients, i.e. users that request development of a specific AI algorithm either for their own business or for a third party.

The examples of explaining the working processes could be explaining the data, e.g., *"you have to be very, very transparent about how you prepared the data, because any AI is biased just because of the quality of the data (and also the quantity)"* (P7) and explaining the specifics of the system and AI in general, e.g. *"when we [...] try to be as transparent as possible on how it [the AI-embedded system] works, we try to explain it to them [clients], because it can be sometimes quite technical, even mathematical, and then there are no more problems, no problem of trust..."* (P4).

AI literacy: decision subjects perceived AI literacy as more specific and operational than AI practitioners There have been diverging opinions about the role of AI literacy for Human-AI trust between AI practitioners and de-

cision subjects: while AI practitioners emphasize the need for raising general public awareness around AI, decision subjects believe in system-related literacy, i.e. more specific and operational knowledge about AI. P5 believes public education on the general understanding of AI could be beneficial for calibrating human trust in AI, *"because people will say «I do not trust AI», without really understanding what AI is"* (P5). Similarly, P7 believes that users should understand the boundaries in AI performance - what AI can do and cannot do.

However, for decision subjects, it is not enough to raise public awareness about how AI works in general, because it is not specific enough, e.g. *"educational events [about AI] do not really make sense to me, because often nobody knows how the system really works"* (DS4), or not actionable enough, e.g. *"the educational sessions [about AI] do not make sense to me, how can they help?.."* (DS6). Therefore, affecting Human-AI trust through AI literacy seems to be possible by accounting for needs of a specific stakeholder group. For example, for decision subjects to understand how a Human-AI decision is made to be able to act upon it, P7 provides training tailored for their decision subjects: *"[we] create materials, [...] flyers, [...] content for patients so that they are informed, that they are not afraid of this new technology"* (P7).

Interactivity: exploration tool for AI practitioners, means to be included in the loop for decision subjects

AI practitioners and decision subjects agree that interactivity is another factor impacting Human-AI trust in the context of decision making (P1-P4, DS2, DS3, DS4, DS6). They also agree that interactivity is often limited. For instance, *"I give you [AI] input data - you [AI] send me back the solution, and I have no other contextual elements, elements of interaction with you*" (P2),

"I would like to have the opportunity to negotiate and influence the [the AI's] decision and say, «Hey, but look at this and that»" (DS4) or "*these [AI] systems should be more tolerant to human error. Right now, it's so strict*" (DS6). However, their opinions differ when considering the consequences of this limited interactivity. For AI practitioners, it hampers one's ability to explore the system, *"asking for more explanations"* (P3) and establish *"a dialogue"* (P4) or *"cooperation"* (P1) between users and AI. For decision subjects, the limited interactivity leads to more serious consequences. It provides a feeling of being excluded from the loop. Decision subjects feel they lose their sense of agency. They see themselves as "*statistics*" (DS2) or simply "*filtered out*" by AI (DS3) because the system is not "flexible" (DS3) or does not allow "to negociate" (DS4).

5.6 Discussion

In this paper, we investigated Human-AI trust from two perspectives - what AI practitioners think is important for trust in AI of other stakeholders and what decision subjects think is important for their trust in AI. Combining these perspectives allows for understanding similarities and differences in how these different stakeholders define Human-AI trust in the context of decision making and what factors affecting Human-AI trust they prioritize. In this section, we discuss what our results mean for 1) re-envisioning what factors affect Human-AI trust in the socio-technical ecosystem; 2) defining Human-AI trust and its key prerequisite elements for its existence; and (3) in terms of stakeholders' agency over the system. Finally, we present some

limitations of our study and propose future research directions that could address these limitations.

5.6.1 On the Important Role of Inter-personal Relations on Trust Within the Socio-technical System

Our results revealed the important role of interpersonal relationships on trust. In other words, AI practitioners and decision subjects stressed the importance of trust links with other stakeholders involved in the system: its design, development, deployment or use in real applications. Moreover, this importance seems to take precedence over the technical characteristics of the system. These results complement recent findings on the under-explored concept of social transparency for AI-assisted decision-making (Ehsan, Liao, et al., 2021a). Through highlighting the history of other users' interactions with AI recommendations rather than the inner workings of AI, social transparency embraces the interviewees' emphasis on trust factors related to social interactions, information actionability, and expectations as a part of the system's design. In this sense, our findings about the importance of interpersonal relationships also support recent approaches arguing for trust calibration beyond direct interaction of people with AI (Browne, Bakker, et al., 2022).

From the different cases of trust links between stakeholders elicited in the findings, trust in the AI team (cases 1 and 4 in Figure 5.1) is generally absent in the literature, while respondents believe that this plays an important role in the trust between humans and AI. So far, the literature suggests that the reputation of the organization that develops AI plays a role for doctors' trust in AI recommendations (Burgess, Jankovic, et al., 2023; Sivaraman,

Bukowski, et al., 2023), and our study confirms this for the domains beyond medical decision making. Another difference is that users' trust in other users (case 2) is more emphasised in the academic literature than in the interviews (A. Brown, Chouldechova, et al., 2019; Ehsan, Liao, et al., 2021a; Jacobs, He, et al., 2021; Okolo, Kamath, et al., 2021; Saxena, Badillo-Urquiola, et al., 2021a). Research shows that observing other users (especially colleagues) trusting the recommendations of the system can increase one's own trust in AI (Ehsan, Liao, et al., 2021a; Jacobs, He, et al., 2021).

However, from the interviews, AI practitioners often serve as intermediaries between users and convey feedback as product reviews. Finally, the relationship of trust between decision subjects an other users (case 4) is barely present in the interviews. The academic literature shows that if decision subjects (e.g., a patient) trust the direct user (e.g., a clinician) and the direct user trusts the AI recommendations, then they would also trust the AI recommendations (Okolo, Kamath, et al., 2021) and vice versa (A. Brown, Chouldechova, et al., 2019; Das Swain, Gao, et al., 2023).

Our results suggest that these bonds of trust are either transversal (e.g. users to users) or upstream (e.g. users to AI team). We believe that trust, in this case, relates to the people who have either more expertise on the domain and technology or means of actions over the technology (such as the AI team). Therefore, these trust links might take an even more important role for decision subjects than other stakeholders. In fact, we saw this in the perception of the role of AI transparency on trust. Contrary to AI practitioners, decision subjects do not see how transparency can affect their trust in AI since explanations might be difficult to understand and the additional information about AI or a specific system is usually not actionable. Specifically, neither the

interviewed AI practitioners, nor the literature provide ample reflections for the role of transparency for trust in AI of decision subjects. Transparency is, hence, viewed as a factor affecting primarily users' trust in AI, targeting their needs for quality evaluation of an AI recommendation and for refining their mental model about AI, which does not necessary encompass actionability and contestability - the needs of decision subjects (Marian, 2023; Yurrita, Draws, et al., 2023).

Research implications

1. **Investigating the points of Human-AI trust breakdowns and calibrations beyond direct interaction with the system**. To this end, research needs to involve more fieldwork with the various stakeholders to understand how trust in AI is shaped and influenced within the complex web of relationships among AI practitioners, decision subjects, users, and other stakeholders, and identify key patterns and dynamics of trust flow among these stakeholders.
2. **Re-examining the techno-centric trust factors between humans and AI with a social lens**. Following the example of Ehsan, Liao, et al. (2021a), who proposed the term of social transparency, moving away from providing more information about how AI works to more information about how other users make decisions with the system, we envision other Human-AI trust factors can be relooked in the same manner. For instance, in addition to reporting AI accuracy, one can inform users about how AI recommendations affected the performance of other users.

5.6.2 On the Prerequisites for the Existence of Trust

In order to understand what AI practitioners and decision subjects expect from a system they trust, we analysed how these stakeholders understand trust, i.e. what essential elements, or prerequisites, they associate with this notion. Both groups elicited the need for positive expectations and a situation of vulnerability. These two prerequisites are how theoretical work in the literature defines trust. This was unexpected, because trust is a complex and abstract theoretical concept that leads to frequent theoretical confusions (Jacovi, Marasović, et al., 2021; Q. Liao and Sundar, 2022; Vereschak, Bailly, and Caramiaux, 2021). It remains that we found a more nuanced outlook on the key elements of trust in comparison with the academic literature. The interviewees' discussions highlight that vulnerability and positive expectations cannot be boiled down to monetary losses and high levels of accuracy as they are often presented in the empirical studies (Vereschak, Bailly, and Caramiaux, 2021). Vulnerability denotes a state in which someone feels the possibility of being emotionally attacked, and therefore finds themselves in a position of weakness. In our results, we had the example of a judgement based on physical appearance. So these prerequisites for the existence of trust depend on the individual or the community with which the individual identifies. Recent examples of the behaviour of algorithms that discriminate against a certain population, such as black women (Buolamwini and Gebru, 2018), place them in a vulnerable position more than other individuals. Additionally, decision subjects report to feel vulnerable, because they have no control over how the data they share about themselves for Human-AI decision making get interpreted by the users in charge of these decisions (Das Swain, Gao, et al., 2023). Sometimes, in order to appear cooperative,

they provide more data about themselves than needed, which puts them at risk of "algorithmic stigmatization" (Andalibi, Pyle, et al., 2023b; Saxena, Moon, et al., 2023) - wrongfully assigned a certain label "at risk", e.g., risk of recidivism, child maltreatment, suicidal tendencies, based on the an algorithmic assemblage.

Our results also highlighted a new prerequisite for the existence of trust, namely the complexity of the task. Behind this prerequisite is the idea that if the task is simple, it can be easily solved by the person using the system or receiving a decision from it. Thus, if one knows the right answer, evaluating the quality of AI recommendation is straightforward, the conflict of between one's own opinion and the AI recommendation does not emerge, and consequently, neither does the state of trust. However, there is an ambiguity about the definition of complexity. Typically, we could envisage two scenarios. Firstly, complexity can arise from the impossibility for a human to process a large amount of information (for example, a large amount of data in a database) in order to produce a decision. Secondly, complexity can arise from a lack of expertise, either related to the decision domain, or related to the underlying AI techniques. If task complexity is a prerequisite for the emergence of trust, along with vulnerability and positive expectations, this implies that future research should study it and include it in the way experimental tasks are designed to focus on trust, rather than confidence (Vereschak, Bailly, and Caramiaux, 2021).

Research implications

1. **Understanding the role of vulnerability in AI-based decision-making systems**. As our findings indicate that feeling vulnerability to AI-based decision-making systems can go beyond monetary gains and losses,

especially in the case of decision subjects, qualitative studies, such as interviews and case studies, could be utilized to gain deeper insights into individual and community experiences of vulnerability towards AI in order to inform further research on trust.

2. **Incorporating complexity into experimental studies of trust**. Given that task complexity could be a key element in trust formation, it should be accounted for in designing experiments that study trust in AI to distinguish between trust and confidence in the system's recommendations. Future research should investigate what aspects of task should be considered to vary the task complexity as well as to which extent it contributes to the formation of trust as a function of different levels of task complexity.

5.6.3 On the Notion of Agency over the System

AI practitioners and decision subjects both stress the importance of AI interactivity for trust, but their views on the purpose of interactivity differ. For AI practitioners, interactivity is a means to explore AI recommendations. From this point of view, they agree with what previous work has shown about the fact that interactivity contributes to explore to which extent nuances are accounted for AI recommendations (Saxena, Badillo-Urquiola, et al., 2021a). Other works have shown, in addition, that interactivity also contributes to the refinement of the mental model about AI (Cai, Reif, et al., 2019) and gives a sense of striving to improve decision making (Okolo, Kamath, et al., 2021). Decision subjects, on the other hand, see interactivity as a way of getting involved in the decision-making loop. In other words, they see inter-

activity first as a way of being represented in the decision-making process, before being able to formalise what this representation could bring in terms of understanding the system's mechanisms and creating a mental model of its behaviour. We therefore see, in these different opinions between the stakeholders, a difference in power relationships. AI practitioners have the means to act on the system, and are therefore in a position to imagine what these means can bring them.

This interpretation suggests that interactivity is related to the notion of agency. In fact, decision subjects discuss the sense of agency and its relationship to Human-AI trust more than AI practitioners, which is expected considering the mentioned frustrations about their lack of actionability and power over the systems. This means that decision subjects value more the factors of trust linked to their inclusion in the decision-making loop in comparison with AI practitioners. It is an empowerment over Human-AI decisions so as not to feel solely "*part of the statistics*", as put by DS2. These findings align with the prior work (Jakesch, Buçinca, et al., 2022) showing that different groups of stakeholders prioritize ethical values differently. Our findings extend this line of research by demonstrating this for trust and underlines the importance of undertaking a multi-stakeholder approach (Yurrita, Murray-Rust, et al., 2022) for Human-AI trust.

That being said, although the academic literature on Human-AI trust examining the interactivity of AI recommendations have led to certain results, as those mentioned above, this research remains scarce (Bridgwater, Giuliani, et al., 2020; Cai, Reif, et al., 2019; Gupta, Basu, et al., 2022; Okolo, Kamath, et al., 2021; Saxena, Badillo-Urquiola, et al., 2021a). Moreover, these studies are primarily about users rather than decision subjects. Similarly, while previ-

ous work has investigated the relationship between agency and trust in AI, it focuses exclusively on the agency of direct users (e.g., Buçinca, Malaya, and Gajos, 2021; Fogliato, Chouldechova, and Lipton, 2021; Kapania, Siy, et al., 2022; Schaffer, O'Donovan, et al., 2019; Sutherland, Harteveld, and Young, 2015; Tolmeijer, Christen, et al., 2022; Xiao, Stasko, and Catrambone, 2007). Additionally, in all these articles, participants are fully aware to which extent they have control over AI recommendations, and their level of agency remains unchanged throughout the experiment. Hence, the issue of varying levels of control over AI is not largely studied in the Human-AI trust literature in the context of decision making. Moreover, in the interviews, agency is mostly referred to as ability to contest a Human-AI decision, while in the literature, it is mainly represented as control over seeing an AI recommendation: full - AI recommendations are optional and appear on demand (Buçinca, Malaya, and Gajos, 2021; Kraus, Wagner, and Minker, 2020; Schaffer, O'Donovan, et al., 2019; Sutherland, Harteveld, and Young, 2015; van Maanen, Wisse, et al., 2011), limited - mandatory AI recommendations that appear immediately (Buçinca, Malaya, and Gajos, 2021; Fogliato, Chouldechova, and Lipton, 2021; Kraus, Wagner, and Minker, 2020; Rajaonah, Anceaux, et al., 2006; Schaffer, O'Donovan, et al., 2019; Sutherland, Harteveld, and Young, 2015; Tolmeijer, Christen, et al., 2022; van Maanen, Wisse, et al., 2011) or only after users' initial decision (Buçinca, Malaya, and Gajos, 2021; Fogliato, Chouldechova, and Lipton, 2021), and none - AI recommendations executed autonomously (Kraus, Wagner, and Minker, 2020; Maurer, Erbach, et al., 2018; Rajaonah, Anceaux, et al., 2006; Tolmeijer, Christen, et al., 2022; van Maanen, Wisse, et al., 2011). Therefore, it remains unclear to which extent the solution of "introducing four levels [of AI recommendations] instead of

the binary [...]" proposed by P7 to increase the sense of agency for decision subjects would work.

Research implications

1. **Investigating the role of different mechanisms of interactivity for trust in AI of various stakeholders.** Our findings indicate that interactivity plays a different role for decision subjects than for AI users, and thus might affect their trust in AI not through the same mechanisms. HCI researchers could conduct in-depth studies to examine how different interactive features (e.g., feedback loops, adjustable parameters) empower decision subjects or change trust links between them and practitioners or users.
2. **Investigating agency as human capability instead of a feature of the system.** Our results have shown the importance of human agency, particularly for decision subjects. While agency tends to be seen as a feature of the system (e.g., providing means to act on system behavior), it is also related to people's perception of actions on the system and their representation by the system. In the same way as trust, this concept needs to be better understood from a human-centric point of view in the context of interactions with AI-based decision-making systems.

5.6.4 Future Work Directions

In this article, we interviewed representatives from a varied panel of decision domains (e.g. medicine, finance, recruitment). Although our main objective

was to study the factors of trust between humans and artificial intelligence for risk-sensitive applications, each domain may nuance the effects on trust due to the diversity of decision-making flows, types of stakeholders involved, etc., which is one of the limitations in the interpretation of the study's results. Considering that type of task and level of risk also have impact on Human-AI trust, it could be interesting to conduct a cross-domain comparison to see to which extent they put importance on the same Human-AI trust factors. Understanding the differences and similarities between various task domains can inform researchers and policy makers on higher level classification of domains (Lai, Chen, et al., 2023). Additionally, we did not account for individual differences such as gender, age, and explicitly assess prior experience with AI, and other demographic information in our analysis while these factors can further influence how certain ethical values are prioritized (Jakesch, Buçinca, et al., 2022).

Secondly, we considered two types of stakeholders that are not users - AI practitioners and decision subjects. While there is no widely established categorization, some researchers propose a set of 11 stakeholders' groups (Ayling and Chapman, 2022) that are connected to the AI ecosystem, spanning from policy makers that work on high level strategies to hiring managers that recruit AI developers. An interesting research direction would be to extend the presented research to these stakeholders and inspect differences and commonalities in findings.

Lastly, we took an organization-focused approach to studying Human-AI trust when talking to AI practitioners. In other words, the AI-embedded systems that they are responsible for are developed, trained, designed, deployed, and monitored by the same company. However, nowadays AI technologies are

often a product of "algorithmic supply chains" (Cobbe, Veale, and Singh, 2023), that is multiple independent actors are responsible for commissioning different phases of production and deployment. As these actors have distributed responsibility over the outcomes of Human-AI decisions with imperfect control over how their work is used further down in an algorithmic supply chain, this can raise additional concerns over whether an AI recommendation produced by "many hands" can be trusted. Further investigating implications for Human-AI trust resulting from such a production set-up can shed light on new nuances about interpersonal dynamics between different stakeholders and identify new potential Human-AI trust breakdown points and factors that affect it.

Part IV

AI-Mediated Social Platforms

Introduction

> "*In the face of uncertainty, we find meaning through the stories we create, the questions we ask, and the actions we take*" — Unknown

This section focuses on the next AI application domain in my dissertation: **AI-mediated social platforms**.

Chapter 6 examines invisible AI policing practices, specifically access control systems. Although I refer to this chapter as AI mediation on social media, the examples of access control discussed here—such as CAPTCHA and re-CAPTCHA tests—extend beyond social media and apply to a broader range of AI systems. However, social media users often experience the consequences of these systems' errors more severely. For example, being falsely identified as a bot can result in losing access to family and friends or harming their online businesses (Myers West, 2018). To explore this issue, I focus on social media users and draw on how science fiction has envisioned the future of AI policing. This allows me to propose implications for designing meaningful control mechanisms when users' behavior and interactions are evaluated by invisible AI systems.

Chapter 7 shifts the focus to dating apps and explores how users make sense of negative experiences with algorithmic mediation and perceived harms on these platforms. The chapter examines the folk theories users develop to explain these occurrences and the coping strategies they use to adapt, navigate, or work around the algorithm. Additionally, it highlights users' needs for empowerment across different levels: feeling, knowing, and doing.

Together, these chapters paint a clearer picture of why user control is so important, especially in the face of perceived algorithmic harms or biases. They emphasize the need for systems that empower users and provide them with a greater sense of agency when engaging with AI-mediated platforms.

6

Social Media

Does Anyone Dream of Invisible A.I.? A Critique of the Making Invisible of A.I. Policing

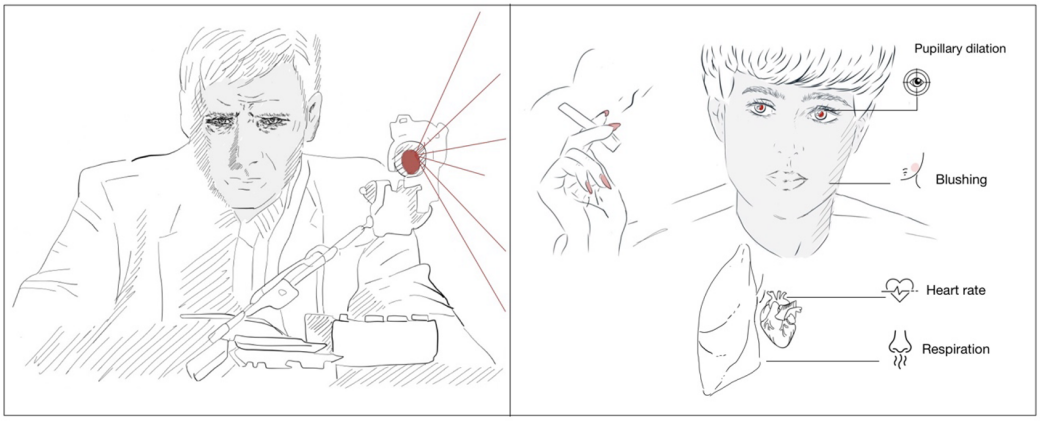


Fig. 6.1.: A sketch of the Voight Kampff (V-K) test in the original Blade Runner movie, focusing on the biometric information collected by the V-K machine during the test

6.1 Abstract

For most people, using their body to authenticate their identity is an integral part of daily life. From our fingerprints to our facial features, our physical characteristics store the information that identifies us as "us." This biometric information is becoming increasingly vital to the way we access and use technology. As more and more platform operators struggle with traffic from malicious bots on their servers, the burden of proof is on users, only this

time they have to prove their very humanity and there is no court or jury to judge, but an invisible algorithmic system. In this paper, we critique the invisibilization of artificial intelligence policing. We argue that this practice obfuscates the underlying process of biometric verification. As a result, the new "invisible" tests leave no room for the user to question whether the process of questioning is even fair or ethical. We challenge this thesis by offering a juxtaposition with the science fiction imagining of the Turing test in *Blade Runner* to reevaluate the ethical grounds for reverse Turing tests, and we urge the research community to pursue alternative routes of bot identification that are more transparent and responsive.

6.2 Introduction

It appears that the internet public is constantly facing the imminent risk of a bot attack. This is the claim that legitimizes the precipitous rise of a new generation of reverse Turing tests, in other words, tests that control whether a user is actually human. Pioneering this effort, The Google No CAPTCHA reCAPTCHA (reverse Completely Automated Public Turing test to tell Computers and Humans Apart) promises to be “tough on bots” and “easy on humans” (Google, 2024).

In practice, this means that the company employs advanced risk analysis techniques that quantify the user’s entire engagement with the CAPTCHA and evaluate a series of cues to come to a verdict about whether a user is (non)human. However, the user does not witness these processes. These algorithms run in the background, escaping the user’s attention. All the user sees is a little rectangle with a ticking box, asking them to tick it to verify

that they are, indeed, human. And yet, as stated above, this test hides a sophisticated risk assessment software that "actively considers a user's entire engagement with the CAPTCHA—before, during, and after—to determine whether that user is a human" (Agarwal, 2023).

Even though Google argues that this test is more usable because it entails less user interaction (Google, 2024), that must not be misunderstood because No CAPTCHA reCAPTCHA tests still require a plethora of user information. In other words, if we take a different perspective to that of Google, the new CAPTCHA tests do not test our performance, but rather they exhibit a familiar "desire to read identity off the body" (Magnet, 2020). To recap, we have noted two key characteristics of the new generation of biometrics-based user authentication tests: 1) the actual testing process is invisible for the users 2) the tests do not require interaction from the user but they still necessitate biometric information.

In this paper, we shall critique this new generation of reverse Turing tests by juxtaposing them with the science fiction imagining of the Turing test in the film Blade Runner. Building our critique on this juxtaposition, we argue that there are serious societal implications as a result of the making invisible of reverse Turing tests, and further, that the widespread fascination with the quantification of the body and the use of biometric measurements as an indicator of human nature jeopardize both the accuracy of such tests, and also, crucially, our understanding of human nature itself. This critique addresses a vital aspect of Human Computer Interaction (HCI) that is currently not addressed within the literature: the agency of the user during their interface with CAPTCHA technologies, and their ability to exert meaningful control over these events.

We see meaningful control as follows: 1) the user understands that they are being tested, and how the test is performed, 2) the user can communicate their frustrations and contest the result of CAPTCHA tests, 3) the user has the right to information regarding the handling of any data collected through biometric tests.

We expound on these arguments in four sections. First, we provide a short content analysis of the film paying special attention to the Voight Kampff (V-K) test, a fictive version of the Turing test. Then, in the first part of our critique, we refer to academic discourse on the topic of transparency of algorithmic systems and the importance of interface design.

In the second part of the critique, we bring up the scholarly discourse around biometrics and we point out the flaws in the logic of biometric CAPTCHA tests. Then, we present the limitations of this project; we underline the importance of critiquing biometric CAPTCHA tests in the field of HCI, while also outlining possible future trajectories for this research. Finally, we summaries our critique and offer some concluding results.

6.3 Blade Runner: A Closer Look at the Voight Kampff Test

The quantification of human behavior as a means to determine one's nature as human or otherwise is not a new phenomenon. In fact, it's a science fiction trope that has existed in the genre since the 20th century. Its most prevalent representative is arguably the V-K test, first encountered as the 'Voigt-Kampff' test in Philip K. *Dick's Do Androids Dream of Electric Sheep,* and

popularized by the 1982 film *Blade Runner*, directed by Ridley Scott (Wheale, 1991). In the film, the police have a problem. Humanity needs to be saved from wayward androids that act so much like humans that they can barely tell them apart. To solve this problem, they turn to a futuristic version of the ancient lie detector and conduct the V-K test. The test measures bodily functions such as heart rate, respiration, blushing in the cheeks and pupillary dilation in response to what are considered to be 'emotionally provocative' questions. It takes twenty to thirty cross-referenced questions to detect a bioengineered humanoid, otherwise known as a replicant. Those who fail the test are destroyed immediately. To lay the groundwork for our critique, we first describe the two main scenes that visualize the V-K test in the film:

We first see the V-K machine in the interrogation scene of Leon (a replicant) (Identities, 2024). As Leon enters the section, we see a device the size of a thick computer lying flat on the table. We are faced with a dark, geometric object, at the center of the screen, its edges pronounced by the direct light falling on it. Once the test subject walks in and the machine is on it takes over the room with continuous mechanical sounds as we observe it flex its parts, unraveling a camera that directs a red light directly on the test subject and us, the audience, who share their perspective in that moment. Holden (the blade runner in charge of the interrogation) immediately turns the device on and a small adjustable armature rises from the machine, at the end of which is a complicated device with a powerful camera glows red. Leon is dull-witted, inquisitive, uncomfortable, and watches the test device with wary suspicion.

When Holden points the camera at one of Leon's eyes and we see a magnified image of Leon's eye on one of the mini-displays (there are three of them),

Leon asks if it is okay for him to speak. As he speaks, we see a series of lights that reflect the volume of Leon's speech. Leon continues by mentioning that he is nervous about the test. We see Holden press the left of the three buttons to the right of the display and the iris of the monitor becomes brighter (Noessel, 2019). He then interrupts Leon's remark by saying, "Just please don't move." Small bellows on the machine rise and fall as though the machine is continuously inhaling and exhaling, as if it were breathing - just like Leon (Žižek, 2021).

- "You're in a desert," says Holden, "walking through the sand, when all of a sudden you look down ..."
- "What one?" Leon interrupts him. Holden replies that it does not matter, it is completely hypothetical. But Leon insists and asks how he got there.
- "Maybe you're fed up," Holden says, adding an emotional note. "Maybe you want to be by yourself. Who knows?" Then he continues with the script. "You look down and see a tortoise, Leon. It's crawling toward you ..."
- "Tortoise? What's that?" Holden asks him if he has seen a turtle and says it is the same. But when Leon says he has never seen a turtle either, Holden gets irritated. Leon picks it up and reassures Holden, saying:
- "But I understand what you mean." Holden continues, "You reach down and you flip the tortoise over on its back, Leon." But Leon is having trouble concentrating.
- "Do you make up these questions, Mr. Holden? Or do they write them down for you?"

We see the V-K machine again later, as Deckard (a Blade Runner) tests Rachel (a replicant), from a different angle but with the same focus on the sensor arm, bellows, and three mini-CRT displays that magnify Rachel's iris. Deckard begins the test with a brief introduction:

- "Alright I'm gonna ask you a series of questions, just relax and answer them as simple as you can."

  During this test, 100 questions are being asked, but the only questions we get to hear in the film are those related to animals. Except for one question:

- "You're reading a magazine, you come across a full-page nude photo of a girl..." Rachel interrupts the question by: "Is this testing whether I'm a replicant or a lesbian, Mr. Deckard?"
- "Just answer the question please", says Deckard, "You show it to your husband, he likes it so much, he hangs it to your bedroom wall."
- "I wouldn't let him", answers Rachel.

What we want to highlight here is not the relevance of the generally accepted moral standards in the V-K test, but the fact that Rachel interrupts the test and rightly criticizes the relevance of the question as soon as she perceives an anomaly in the question. Strictly speaking, it is not really even clear to us audience whether this question is actually one of the questions in the test script or whether Deckard inserted it himself to catch the subject off guard. Therefore, Rachel's criticism seems not only fair, but in some ways "very human".

Overall, *Blade Runner* touches on an important question that goes beyond the futuristic moral dilemma of destroying human-like artificial beings to an important challenge of today's digital world. At issue is whether we can judge personhood by measuring external physical reactions and behaviors, or whether such a test is doomed to failure. In the next session, we draw comparisons between *Blade Runner*'s vision of future tests of humanness with current biometric CAPTCHA tests and invisible AI tracking to critique such practices.

6.4 Critique

Although Blade Runner gets almost nothing right in terms of the technology of robots and androids in 2019, its approach to distinguishing between human and non-human beings is similar to toady's reCAPTCHA tests. Both the V-K test and modern reCAPTCHA tests are based on the original Turing test. In Turing's "imitation game," a human evaluator is asked to judge natural language conversations between a human and a machine that is supposed to produce human-like responses. The machine passes the test if the human evaluator is convinced it is human (Turing, 2009).

To be clear, Turing is not trying to answer the original question of whether machines are actually capable of thinking. Instead, he asks: can machines do what we (as thinking entities) can do? Turing thus distinguishes between a human's physical and intellectual capabilities. Yet the reverse form of his test (practiced in reCAPTCHA today) makes a person's physical behavior the ultimate indicator of their human capabilities. That is, if the tables are turned and computers are tasked with distinguishing between human

and non-human behavior, humanity itself, in all its diversity and complexity, should be part of the equation.

In what follows, we take a critical look at emerging practices for distinguishing nonhuman actors and examine the implications and risks they pose for transparency, privacy, and fairness, as well as for the perception of the human body.

6.4.1 Invisible AI: privacy, transparency, and fairness

Blade Runner paints a dystopian picture of a future in which groundbreaking V-K machines can elicit needed knowledge (in this case, identity) from test subjects even without their own will (much like a polygraph, only with the promise of high accuracy). Despite this overwhelming ability to detect what is invisible to the human eye, the V-K machine must still be manually controlled by an investigator who decides how to conduct the tests and interpret the results. This design of the machine begs the question of why, even in our dystopian science fiction vision of the future, there are no machines that can perform such tests and deliver the results fully automatically.

In this sense, either our imagination has failed us, or we have overestimated the agency and control of the future man. Perhaps it is a consolation that the V-K test needs to be turned on, that the process can be intervened in, and that the results should be interpreted by the human mind. Can the same be said of biometric CAPTCHA and reCAPTCHA tests? A look back at *Blade Runner* suggests that we may already be living in a next-level dystopia where a growing class of invisible AI-controlled autonomous systems makes decisions about (non)human behavior (Alm, Alvarez, et al., 2020).

The dystopian view of *Blade Runner* comes up short not only with respect to the V-K machine itself. But also, in terms of the procedure and the nature of the test. In the V-K test, subjects are aware of the test and its purpose. This may seem inevitable due to the structure of the device (the need to fix the machine on the subjects' eyes to track their movements and the inclusion of various modalities such as sound and light). But the film does not stop there. The flow of the test is highly interactive. Not only do subjects receive an appropriate introduction at the beginning of the test and learn the exact starting point, but they are also allowed to share their feelings and ideas about the questions with the investigator during the procedure. The test flow can also be interrupted if the subject needs more information about a question (in which case they are provided with more information). This is despite the fact that, according to the company, the V-K test is 100% accurate and makes no mistakes.

In many ways, this is not the case with today's invisible AI decision making systems. To begin with, several researchers have shown that users are increasingly influenced by pervasive AI through various forms of existing and emerging human interactions, often even without being aware of the AI system "under the bonnet" (e.g., Bucher, 2019; Eslami, Rickman, et al., 2015; Eslami, Vaccaro, et al., 2017; Pasquale, 2015). In addition, while such systems are not immune to error, they provide no help or support to users who are not even aware of the test, let alone able to understand and judge the fairness of final decision. Researchers have shown that users need multiple attempts to pass a CAPTCHA test, and that a false negative in the bot detection algorithm is inevitable (meaning that a human user will be incorrectly identified as a bot and blocked or banned from the platform)

(Waters, 2021). Those users who still struggle with the reCAPTCHA tests report being annoyed by the incomprehensible images and frustrated by not receiving satisfactory feedback on the "help" pages (Greenberg, 2014). All of this is presented in classic Internet irony in a Reddit group called /r/captcha, where users find comedy in faulty tests. To address this issue, Google has updated this method and made it completely invisible. The new "Invisible reCAPTCHA" mechanism uses the same technology as "No CAPTCHA reCAPTCHA", but without the "I'm not a robot" checkbox. Google, on the other hand, "invisibly analyzes the way users navigate through a website and assigns them a risk score based on how malicious their behavior is" (Schwab, 2019).

Judging by the statements of the major platform operators, the new generation of and authentication mechanisms and invisible reCAPTCHAs does not affect the usability of the platforms. Google touts its latest product, reCAPTCHA Enterprise, as a "frictionless user experience, where fraud protection is easily extended across websites" (Fortunati, Manganelli, et al., 2019). Consequently, more and more organizations are deploying invisible audits of their platform activities (Alizadeh, Mniestri, and Stevens, 2022b). On Instagram, the battle begins as soon as a user loses access to their account (Alizadeh and Stevens, 2020). Once the algorithm flags their content, users, especially those who are entrepreneurial through the platform in the form of promotions and sponsorship deals, are targeted by dealers. These dealers offer creators the opportunity to recover their accounts for a fee. While some of them are simply scammers, others inexplicably have enough access to Instagram's moderation infrastructure to restore the account and "right" all the wrongs decided by the algorithm (Alizadeh, Mniestri, and Stevens,

2022b; Waters, 2021). From a privacy, transparency, and fairness perspective, the new invisible authentication and tracking AI technologies are far more dystopian than the dreadful but visible and manually controllable V-K machine. In particular, because they collect a large amount of consumers' personal data without adequate warning or consent.

6.4.2 The metric society

Our second point of criticism centers around the body, the human body, the body of the machine, and the fixation on the body as a metric. We have established that, in *Blade Runner*, the V-K test measures bodily functions in order to determine whether the subject tested is a human or a replicant. It measures respiration, heart rate, blushing and even pupillary dilation in response to what are presented as emotionally stirring questions (see 6.1). However, none of this is disclosed when we first encounter the machine on our screens, neither to the test participant, nor to us as the audience. We are under surveillance. We are not as human as we once were, we are a collection of parts and we are ready to be broken down into biometric measures that are quantified so that they can be perceived by another human through the non-human. In this section, we will discuss how the V-K test unveils crucial insights into how the use of biometrics in reverse Turing tests does not help identify the human, but rather the quantified human, the human that gains subjectivity not in their own right but rather through their interactions with a machine.

Biometric technologies broadly refer to the use of technology to identify an individual based on one or more aspects of their embodiment (A. K. Jain and

Kumar, 2012). Within the discourse of reCAPTCHAs, biometric technologies are considered to have substantial potential to minimize the effort needed to secure personal activities online (Jang-Jaccard and Nepal, 2014), thus minimizing the time 'wasted' by users on completing these activities. De Marsico, Marchionni, et al., 2017, argue that biometric CAPTCHAs maintain the key feature of previous CAPTCHAs by making tests easy for humans and hard for machines. The test they propose, FATCHA, promises to deliver accurate readings by asking the user to produce simple gestures, selected randomly from a wide set of possible directions. The monitor webcam records the user's gestures and then the software analyzes whether the captured frames feature the right gestures in the correct order. In this method, the user produces something instead of analyzing an image/text/sound. In other words, as the authors note, "The user is the CAPTCHA." This observation from the authors is crucial for our analysis. This statement is characterized by an essential flaw, that the biometric technologies used in the testing process do not manipulate the quantified image of a person but rather represent it accurately.

As Pugliese so eloquently demarkates, "biometric technologies are inscribed with infrastructural relations of disciplinary power underpinned by normative categories of race, gender, (dis-)ability, class and age" (Pugliese, 2012). Therefore, to think that the 'user is the CAPTCHA' simply perpetuates the facade that technology has a neutral role in these processes, otherwise seen as non-discriminatory capability. This is in conflict with what we know about biometric technologies; after all, we cannot dismiss cases such as that of Joy Buolamwini, a Ghanain-American MIT grad student, who discovered first-hand racial discrimination in the facial-analysis software with which she was

working. Simply put, the software failed to "see" the student's dark-skinned face even though it had no problem 'seeing' her lighter skinned collaborators. When Buolamwini put on a white mask on her face, the software recognized the mask's 'facial features' without an issue.

Examples such as this should remind us of Kate Crawford's statement that the greatest threat from AI systems is not that they will outsmart humans, but rather that they will "hard-code sexism, racism, and other forms of discrimination into the digital infrastructure of our societies" (D'ignazio and Klein, 2023). So, once again, we stand for meaningful control and communication with CAPTCHA software in the form of self-advocacy. In view of the above, we deem it necessary to include means for users to dispute a positive test result - meaning to dispute the claim that they are a bot as opposed to a human - and to be able to convey their positive or negative experiences with these software so that their perspective can be valued equally to the verdict of the software.

In another biometric CAPTCHA proposal, Acien et al. offer a novel bot detection approach based on modelling user behaviour during smartphone interactions by using multiple built-in sensors, otherwise referred to as BeCAPTCHA (Acien, Morales, et al., 2020). In other words, it's a CAPTCHA system based on swipe gestures. In their methods, the authors write that they have implemented two testing methods with the "aim to emulate human behaviors." While data from real users was used to train the models they used for the testing, we, as readers, do not have information with regards to the types of users whose data were used. Therefore, through these papers there is the construct of the human user as standardized measurements. Then, these measurements are used to test the BeCAPTCHA to see if it can

recognize the bots. However, there is no mention of the different types of users who may not pass the test? To put this differently, the system is used to test for false positives, bots recognized as users, but not for false negatives, human users who are recognized as bots and lose access to internet services and products. The V-K test presents us with a similar dilemma. Wheale draws attention to the irony in the test noting that the V-K test presents questions that are indubitably supposed to cause moral panic in humans, yet the questions refer to "cruelties which we presently accept as routine" such as animal cruelty akin to boiling lobsters alive or bull-fighting (Wheale, 1991). From this, it follows that no universal moral standard can be assumed and, thus, not all humans can get a perfect or even a passing score for this test.

Both in modern CAPTCHA tests and in the V-K test, it is the responsibility of the test subject to prove their humanity. Fortunati et al. have successfully demonstrated that users who execute CAPTCHA tests and are immediately asked for proof of not being a robot, felt "diminished in their humanity." This is a serious consideration that we must hold onto in the design of future CAPTCHA tests (Fortunati, Manganelli, et al., 2019).

The sense that they are dehumanised is further aggravated by the insecurity born of the knowledge that one's biometric data are being collected and stored by CAPTCHA software. Krohl et al. discovered that, in actuality, users feel exposed by CAPTCHA when they have to give up personal data (Krol, Parkin, and Sasse, 2016b). The fact that it is biometric data that they must submit does not affect their tribulations, regardless of what Google and other CAPTCHA providers might indicate on their websites. In order to restore the balance between the software system and the users, and to embolden the users with a new sense of agency, we propose that they should have a right

to the data that they have provided to the system. Then, it will not only be the system that can scrutinize users, but the users will be equally equipped to criticize the system.

By encouraging this participative model of CAPTCHA design and underlining existing problematic protocols in the new generation of CAPTCHA design, we hope to encourage more transparency but crucially we want HCI researchers and other practitioners to begin considering the quality of their involvement of humans in these tests. It's not merely about adding more facets of testing to our proverbial V-K machines, the machine can contain cameras and sensors of all sorts - but if we do not consider the quality of human experiences in these tests, then we are bound to repeat the problems faced by previous generations of software and the mistakes are bound to affect real people.

6.5 Limitations and Implications for HCI

Before discussing the implications of this work for the HCI community, we would like to address some of the limitations of our work. First, our work is not a critique of any direct system or technology, but rather a critique of the current trend of using reverse Turing tests and biometric measurements of the body as a means of figurative authentication. Moreover, this work should be understood as a hypothetical theoretical investigation that uses *Blade Runner* and the narrative and notion of CAPTCHA used therein to endure problems with the logic of the CAPTCHA test from different aspects. Therefore, it cannot be generalized or applied to other settings.

Further, it is worth mentioning that the original *Blade Runner* movie was released in a different historical episode and context. Therefore, we could

not draw on or consider public understanding of the film or reactions to it. We also want to take into account the fact that the social understanding of biometrics might have been different at that time. Another limitation is that we only analyzed the construct of the V-K test as it is presented in the film. We do not claim that this corresponds to the intended meaning or role of the V-K test as envisioned by the director. We also acknowledge that perhaps some decisions about the structure of the V-K test and the role of actors were influenced by the technical constraints of the time and cinematic appeal.

That said, we believe that *Blade Runner* provides a good grounding for discussing meaningful user participation in verification systems. Based on our critique, we propose several implications for the design of such systems in HCI that are not only more transparent and respectful of users' privacy, but also give them more agency and control in the testing process. After all, if the question raised is how AI can be up to the task of distinguishing human activities from non-human activities on the Internet in a user-friendly manner, then the main idea of invisible AI systems as a solution is to eliminate the question rather than provide an effective answer.

First, our critique has shown the importance of informing subjects about their participation in the test. Not only so that they can make an informed decision about their participation, but also to allow them to formally reflect on or question the logic of the procedure or the results of such tests. This conclusion is also supported by the results of previous studies showing that users have concerns about the privacy and security of their biomaterial facial data due to a lack of knowledge about CAPTCHA testing (e.g., Krol, Parkin, and Sasse, 2016a, 2016b).

Second, our critique highlights the lack of two-way communication during the verification process and provides parallels from the V-K test to argue the importance of such communication between users and the verification system to safeguard users' rights under the GDPR and avoid potential misunderstandings between users and the automated system. Krohl et al. also advocated for more communication between the user and the verification system, claiming that people should not be viewed as CAPTCHA-solving machines, but as individuals who view the security effort in terms of their goals and expectations (Krol, Parkin, and Sasse, 2016b).

Finally, the V-K test is a good example of participant engagement in the testing process. This is partly due to its interactivity, but also to the structure of the test, the selection and wording of the questions, and the appropriate briefing of participants. Krol et al. also showed that users are less frustrated when interacting with a gamified verification mechanism than with reCAPTCHA and biometric verification mechanisms (Krol, Parkin, and Sasse, 2016b). We therefore propose to look at the interactivity of the V-K test as inspiration to allow users a participatory role in this cycle and to create new forms of testing that are as diverse as they are potentially fun and engaging.

6.6 Conclusion

In conclusion, our content analysis of the V-K test in the original *Blade Runner* film, has led us to two decisive factors that currently impede the genuine involvement of human users in the process of CAPTCHA testing. For one, CAPTCHA tests are progressively being made invisible and that usability and invisibility are considered one and the same. For another, CAPTCHA

tests purport to protect the privacy of users by examining biometrics as opposed to personal information. By exploring the relations with the V-K test within the film and discussing our observations in combination with findings within academic literature, we argue in favor of meaningful human control in CAPTCHA testing. In a technological landscape where the humans bear the onus of proving their own humanity, they also need to understand the process they are taking part in and feel involved, empowered and emboldened in that participation.

We suggest that this can be accomplished in three ways. By granting users the right to transparency with regard to being tested, meaning they understand when they are being tested and how. Then, users need to be able to answer back when they receive a false positive reading - if they are mistakenly recognized as a bot, and in so, they should be able to communicate their frustration with the system in order to more reliably understand how we can ameliorate CAPTCHA tests in the future. Last but not least, we take to heart the users' concerns regarding security and the handling of the data collected during biometric CAPTCHA testing. We have voiced our disagreement with the claim that users are more secure as a result of biometric testing and we ask for more accountability from CAPTCHA providers. In other words, users should have a right to the data collected during the testing process and access to information with regards to whether/how the data are being preserved or further utilized.

7

Online Dating

When the "Matchmaker" Does Not Have Your Interest at Heart: Perceived Algorithmic Harms, Folk Theories, and Users' Counter-Strategies on Tinder

7.1 Abstract

On online platforms, algorithms help us build and manage our relationships. However, their invisible interventions might also pose harm to these connections. Dating platforms offer a prime example where, despite extensive research on human-inflicted harm, the potential harm from the algorithms themselves, and user strategies for mitigating them, remains largely unexplored. In our analysis of 7,043 reviews and interviews with 30 Tinder users, we unveiled how users perceive algorithmic harm as damaging self-esteem, sabotaging potential relationships, encouraging antisocial behavior, and misrepresenting or marginalizing certain identities. We introduce a new algorithmic folk theory, the "conflict of interest" theory, perceived to perpetuate these harms. This theory encapsulates users' sense of a contradiction between the dating platform's promise of finding the perfect partner (leading to discontinued use of Tinder) and its commercial interest in retaining users to increase revenue. Users suspected various algorithmic processes pursuant

to this theory, such as (a) throttling profile visibility, (b) manipulating users' matches, and (c) recommending large quantities of profiles that will not lead to matches. They also described various strategies in resistance or defense of these suspected algorithmic processes, such as engaging in counter-intuitive behavior to disrupt the unfavorable algorithmic processes or leveraging location based filtering for match variety and safety. We conclude by discussing how the perceived algorithmic harms can inform the development of new algorithmic implementations that balance both user and company interests.

7.2 Introduction

Algorithms can shape and mediate relationships on a variety of online platforms, such as social media and messaging apps (Eslami, Rickman, et al., 2015; J. T. Hancock, Naaman, and Levy, 2020; Wu and Kelly, 2020), online marketplaces (Jhaver, Karpfen, and Antin, 2018), and dating platforms (Courtois and Timmermans, 2018). While such algorithms offer opportunities to maintain old relationships and build new ones, they can also introduce harm to people's social relationships (Eslami, Rickman, et al., 2015; J. T. Hancock, Naaman, and Levy, 2020; Häubl and Trifts, 2000).

A context in which algorithmic harms to individuals and their social relationships remains under-explored is online dating - barring early reflections on racial and sexual biases amplified by matching algorithms (Hutson, Taft, et al., 2018; Notopoulos, 2016). This is a conspicuous gap given the extensive use of algorithms in dating apps for user recommendation, and extensive research into *human-inflicted* harms through online dating such as sexual violence (Agency, 2022; Mladenović, Ošmjanski, and Stanković, 2021; Valen-

tine, Miles, et al., n.d.) and harassment (L. Blackwell, Dimond, et al., 2017; Doerfler, Forte, et al., 2021; Freeman, Zamanifard, et al., 2022; Jhaver, Ghoshal, et al., 2018; Rubin, Blackwell, and Conley, 2020).

In this paper, we use the lens of algorithmic "folk theories," i.e., the theories that users of algorithmic systems develop to make sense of the system's operation (DeVito, Gergle, and Birnholtz, 2017; Eslami, Karahalios, et al., 2016), to understand users' perceptions of algorithmic harms and their strategies for mitigating associated algorithmic processes in the context of Tinder, the most popular dating platform (Tinder, 2023). In recent years, research has explored how users of dating apps develop algorithmic folk theories to make sense of what they experience on the platform (Nader and Lee, 2022), to explain how matching algorithms work and increase their attractiveness rating (S. A. Huang, Hancock, and Tong, 2022; Nader and Lee, 2022), and even to manage the subsequent stages of relationship formation (Sharabi, 2021; Tong, Hancock, and Slatcher, 2016). Yet users' perceptions of the harms that algorithms cause on dating platforms has not been directly studied. Such an exploration is essential for a more complete understanding of the harms that users perceive themselves to be at risk of in online dating - whether from humans or machines - and for informing future designs to mitigate threats from diverse sources. Therefore, we ask:

- **RQ1.** What potential harms do users think Tinder dating algorithms might cause to themselves and their relationships?
- **RQ2.** What are users' perceptions, beliefs, and folk theories about how algorithms instigate such harms?

- **RQ3.** How do users alter their usage strategies of dating apps to reduce the perceived harms of Tinder dating algorithms and regain control of their dating experiences?

To answer these questions we conducted a two-stage study. In the first stage, we analyzed 7,043 Tinder reviews to get an overview of the harms that users perceive Tinder dating algorithms to cause to their relationship building process.

In the second stage, we conducted semi-structured interviews with 30 Tinder users to understand their folk theories and sense-making of such harms and the nuances of their app-usage strategies to mitigate undesirable algorithmic processes and regain a sense of agency and control.

Our findings reveal several algorithmic harms perceived by users such as damaging self-esteem, sabotaging potential relationships, encouraging anti-social behavior, and misrepresenting/ marginalizing certain identities. Participants collectively suggested a new overarching folk theory of dating algorithms that perpetuate these harms, which we call the "*conflict-of-interest*" theory: the perceived contradiction between the dating platform's main promise of finding the perfect partner (which would result in discontinued use of Tinder), and the platform's commercial interest in keeping users on the platform and increasing premium membership subscriptions. Users suspected various algorithmic processes pursuant to this theory, such as (a) throttling profile visibility, (b) manipulating users' matches, and (c) recommending large quantities of profiles that will not lead to matches. Users described various strategies in resistance or defense of these suspected algorithmic processes, some of which would appear directly counterproductive such as not "liking" profiles that one actually finds attractive. The findings also show that users

perceive algorithmic processes to *enable* human-inflicted harm, such as the belief that accounts are automatically banned when one's profile is reported to the app, which can be abused by a user in retaliation for being romantically rejected. This suggests that algorithmic and human-inflicted harms may be intertwined, necessitating that future research on computer-mediated harm mitigation take both human and machine entities into account.

7.3 Related Work

7.3.1 Interpersonal Harm through Online Dating

Interpersonal harms are well studied in HCI, including those that manifest online (e.g., harassment L. Blackwell, Dimond, et al., 2017; Doerfler, Forte, et al., 2021; Freeman, Zamanifard, et al., 2022; Jhaver, Ghoshal, et al., 2018; Rubin, Blackwell, and Conley, 2020, cyberbullying S. Kim, Razi, et al., 2021) and in the physical world (e.g., cyber-grooming Lorenzo-Dus, Kinzel, and Cristofaro, 2020; Miah, Yearwood, and Kulkarni, 2011) to coerce victims into sexual abuse (Mladenović, Ošmjanski, and Stanković, 2021) or sex trafficking (Badillo-Urquiola, Razi, et al., 2020). A context through which computer-mediated interpersonal harm often occurs - and is often studied - is online dating because such apps are designed for users to traverse their interactions into the physical world, thus putting them at risk of harm across virtual and physical modalities. Harassment through messaging interactions in dating app interfaces is well documented (Albury, Byron, et al., 2019; Anderson, Vogels, and Turner, 2020; Shaw, 2016), and dating apps are also a well known facilitator physical sexual harm (Agency, 2022; Valentine,

Miles, et al., n.d.). For example, research has connected dating app-use with increased likelihood of sexual violence victimization (Gilbert, Sarvet, et al., 2019), and other work has found dating apps to be a common method of introduction between perpetrator and victim of sexual assault - accounting for 10% of overall rape cases in some samples (Powell and Henry, 2019; "Swipe right: the emergence of dating-app facilitated sexual assault. A descriptive retrospective audit of forensic examination caseload in an Australian metropolitan service", 2020).

The literature has elucidated various ways in which interpersonal harm manifests in online dating, including both intentional acts and those that are unintentional, without a deliberate intent to cause harm. For example, research has explored intentional harm through deceptive practices like the creation of fake profiles for scams (Whitty, 2015), as well as harassment stemming from romantic rejection (Shaw, 2016). There is also extensive research on targeted harm towards users from marginalized groups, such as online bullying and discrimination (Duguay, 2017; Duguay, Dietzel, and Myles, 2022), as well as 'identity-based harms' (Duguay, Burgess, and Suzor, 2020). Identity-based harms refer to harms that specifically target users based on aspects of their identity, such as their sexual orientation, gender identity, race, or religious beliefs. For instance, Duguay et al. have documented how queer women on dating platforms frequently encounter sexually aggressive "homophobic" users, including instances of receiving unsolicited explicit images or being pressured into sexting (Duguay, Burgess, and Suzor, 2020). Several studies have also explored the outing of gay users in culturally conservative areas, revealing the complex and often dangerous challenges they face (Birnholtz, Rawat, et al., 2020; C. Blackwell, Birnholtz, and Abbott,

2014; Corriero and Tong, 2016). Interpersonal harms in online dating can occur unintentionally as well, often influenced by sexual scripts (Comunello, Parisi, and Ieracitano, 2020; Race, 2015) and consent practices (Zytko, Furlo, et al., 2021) that lead to misinterpretations of consent through dating app interfaces.

Technical solutions developed in response to these risks include improved interaction interfaces for better impression formation (Frost, Chance, et al., 2008; Zytko, Regalado, et al., 2020), AI-based detection of harassment and unwanted images (Simon, 2021; Thompson, 2022), and concepts for mediating consent (Zytko and Furlo, 2023), addressing the challenges in practicing harm-mitigative consent (Zytko, Furlo, et al., 2021).

Online daters have also developed their own safety strategies while using dating apps. Users, especially women and LGBTQ+ individuals who are at higher risk of dating-related harm (Albury, Byron, et al., 2019; Aljasim and Zytko, 2022; Fernandez and Birnholtz, 2019), often manage safety by selectively disclosing personal details and engaging in impression management (Ellison, Heino, and Gibbs, 2006; Zytko, Grandhi, and Jones, 2014). This includes strategies like masking casual sexual interest to avoid "slut shaming" (C. Blackwell, Birnholtz, and Abbott, 2014; Zytko, Grandhi, and Jones, 2015) and carefully evaluating potential partners through their profiles and messaging (Gibbs, Ellison, and Lai, 2011) to reduce the risk of physical harm (Zytko, Regalado, et al., 2020). Woman-identifying users also use safety check-ins with friends and family during face-to-face dates (Aljasim and Zytko, 2022).

7.3.2 Algorithmic Harm through Social Platforms and Users' Folk Theories

It has become clear through the literature that the way users perceive risk of interpersonal harm significantly influences how dating apps are used, and knowledge of that use has in turn influenced how they are designed. Yet there is relatively limited knowledge of how users perceive harm inflicted or facilitated by dating app *algorithms*, and the strategies employed for mitigating these perceived risks.

The term *algorithmic harm* has been used in a variety of ways, dealing inter alia with aspects of racial, gender, and identity bias (e.g., Buolamwini, 2022; Gerrard and Thornham, 2020; Hamilton, 2019; Noble, 2018), discrimination (Citron and Pasquale, 2014; Kantor, 2022), privacy violations, surveillance, news filtering, and electoral impacts (Tufekci, 2015), and so on. Attempts have been made to categorise these numerous and diverse forms of harm. These include the distinctions made by Barocas and Crawford between two specific forms of algorithmic harm (Barocas, Crawford, et al., 2017): allocational and representational. In short, allocational harm occurs when resources or opportunities are unfairly distributed or withheld. Representational harms, on the other hand, pertain to the misrepresentation or marginalization of social groups, as highlighted in studies by (Andalibi, Pyle, et al., 2023a; Karizat, Delmonaco, et al., 2021; Katzman, Wang, et al., 2023). Further classification is provided by Saurwein and Spencer-Smith, who defined roles for algorithms in harm on social media platforms (Saurwein and Spencer-Smith, 2021), comprising (1) deficient tools that lead to errors, (2) instruments that serve manipulation, (3) amplifiers of problematic content,

(4) enabling structures for problematic behaviour, and (5) instruments of platform power. The AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) maintains a repository of over 1000 algorithmic harms and has produced a consequent taxonomy of harms (AIAAIC, 2023).

Social media platforms have been a popular context for studying the adverse impacts of algorithms due to their proliferation for curating and moderating user-generated content. Researchers have shown that sometimes users are not consciously aware of algorithmic processes on social platforms, which can have negative impacts on human relationships (e.g., Eslami, Karahalios, et al., 2016; Eslami, Vaccaro, et al., 2019; Jhaver, Karpfen, and Antin, 2018; Raval and Dourish, 2016). For instance, studies on Facebook's news feed curation algorithm have revealed that users frequently do not realize the algorithmic influence at play. This lack of awareness leads them to mistakenly blame their friends and family for the disappearance of posts, rather than recognizing it as a result of algorithmic selection (Eslami, Karahalios, et al., 2016).

Conversely, in some cases users do perceive algorithmic harm and construct *folk theories* to explain how algorithms work (e.g., Alizadeh, Stevens, and Esau, 2021; DeVito, Birnholtz, et al., 2018; DeVito, Gergle, and Birnholtz, 2017; Karizat, Delmonaco, et al., 2021; T. Liao and Tyson, 2021; Siles, Segura-Castillo, et al., 2020; Toff and Nielsen, 2018). These folk theories, which are intuitive, informal explanations that individuals develop to make sense of technological systems, have been found to guide user behavior and resistance to algorithmic changes (DeVito, Gergle, and Birnholtz, 2017). Devito et al. conducted several studies on the interplay of folk theories and self-representation practices on social media to explore how folk theories

help guide users' behavior in managing impressions to to others (DeVito, Birnholtz, et al., 2018; DeVito, Gergle, and Birnholtz, 2017; DeVito, 2021). Karizat et. al. applied Identity Strainer Theory to refer to users' beliefs that "*an algorithm filters out and suppresses certain social identities*" (e.g., LGBTQ+ identity) (Karizat, Delmonaco, et al., 2021). The study also illustrated that users changed their behavior to resist the perceived algorithmic suppression and associated harm. An analysis of 102,827 tweets of a hashtag (i.e., #RIPTwitter) in a Twitter timeline revealed that users' folk theories influence their resistance to algorithmic change (DeVito, Gergle, and Birnholtz, 2017). Overall, the body of literature on algorithmic harm and folk theories - particularly in regards to social platforms - makes clear that the user experience is significantly impacted by users' perceptions of adverse impacts of algorithms because they employ strategies to resist, mitigate, or adapt to algorithmic processes.

7.3.3 Online Dating Algorithms and their Impact on User Experience

Dating apps are a conspicuously understudied context for how users perceive and attempt to mitigate algorithmic harm given 1) the extensive body of knowledge into user strategies for perceiving and mitigating risk of *interpersonal* harm and 2) the extensive use of algorithms in dating apps. Today's leading dating apps employ algorithms for processes ranging from bot detection (Acker and Beaton, 2016), user verification (Duguay, Dietzel, and Myles, 2022), detection of interpersonal harm (Simon, 2021; Thompson, 2022), and most importantly: user recommendation (also called matchmaking).

Extensive attention has been given to Tinder's matchmaking algorithm in particular (e.g., Bartlett, 2020; Courtois and Timmermans, 2018; Dent, 2019; Junkie, 2020; Leal, 2016; Lopes and Vogel, 2018; Mikonoss, 2019; Nader and Lee, 2022) given the general popularity of the app and the centrality of its user recommendation/matchmaking algorithm to the user experience.

Tinder users view profiles one at a time in a order determined by the recommendation algorithm (Abel, Pei, et al., 2022; VIDAselect, 2022). The ability to message another user depends on both partners establishing a "match" by "liking" each other's profile when it appears in the recommendation interface (Orosz, Tóth-Király, et al., 2016). This makes one's visibility to other users through the recommendation algorithm a crucial factor to their dating success - Sharabi's work demonstrated that simply believing in the effectiveness of matching algorithms impacts real-life first-date experiences (Sharabi, 2021). Users reported better first dates to the extent that they believed in the algorithmic matching process.

While there is no work to our knowledge that has studied online daters' perceptions of algorithmic harms and associated mitigation strategies, attention has been give to folk theories of how Tinder' matching algorithm operates and how users adapt their usage strategies to maximize dating success. Tinder has long been speculated to base its matching algorithm on "Elo scores" (Junkie, 2020; K. L. Smith, 2020) - a measure of one's attractiveness based on the number of "likes" they receive in conjunction with how selective they are in giving out "likes" themselves (the most attractive users would therefore receive the most likes while giving out the least likes). In March 2019, the company announced they had abandoned the Elo score; however, users are still in the dark about how the algorithm works (Tinderpress, 2022a). Cour-

tois and Timmermans used experience sampling method to measure users' swiping and liking ratio, along with resultant number of matches, to infer the inner workings of the matching algorithm (Courtois and Timmermans, 2018). The study concludes that Tinder "prohibits its major assets of attractive profiles and liked profiles to run out too soon" in order to convert non-paying users to premium subscription plans (p. 13). Nader and Lee observed that users' belief in a desirability score (similar to the Elo score speculation) led to various strategies to increase one's score, such as creating new profiles or adjusting search filters (Nader and Lee, 2022). Abel et al. built upon this by conducting 22 in-depth user interviews (Abel, Pei, et al., 2022). Their findings revealed diverse theories among users about how Tinder's algorithms operate. They attributed Tinder's profile curation not only to desirability scores but also to factors like user feedback, romantic compatibility, and the profitability of choices (Abel, Pei, et al., 2022). Similarly, in their interview study, Huang et al. identified folk theories surrounding the Tinder matching algorithm (S. A. Huang, Hancock, and Tong, 2022). They noted perceptions of the algorithm as facilitating a form of 'shopping', being subject to chance and randomness, and functioning as a bracket system for scoring users.

Despite these insights, user perceptions of algorithmic harms in dating apps and associated coping strategies have not been an *explicit* research focus. This paper aims to fill this gap by delving into the nature of algorithmic harm on Tinder, users' theories, and mitigation strategies for such harm. We extend our focus beyond Tinder's matchmaking algorithms to encompass all types of algorithms used on Tinder that can mediate or influence user relationships. For the sake of simplicity, we will refer to these collectively as the ***Tinder dating algorithms***.

7.4 Method

This study was conducted in two stages to explore algorithmic harms in relationship building on the Tinder platform and how users perceive and cope with these harms.

7.4.1 Stage 1: Analyzing Tinder App Reviews

Considering that application reviews contain rich, contextual information about how users perceive an app (Voskobojnikov, Wiese, et al., 2021), we first conducted a qualitative analysis of 7,043 Tinder app reviews from the Google Play Store, focusing on the types of algorithmic harms experienced during interactions with the platform.

Data collection

We collected the data using a Python script based on BeautifulSoup (Documentation, 2022), a package for web scraping and parsing HTML files. The dataset consisted of 7,043 user reviews from the Google Play Store on the "Tinder" app (S. A. Huang, Hancock, and Tong, 2022). Each user review included a rating indicated by the number of stars (0–5), a timestamp, the review text, and the number of likes. The average review was 23.5 words long. Reviews were collected over a year, from March 27, 2020, to March 27, 2021, a period marked by a surge in Tinder's user base and its highest interaction rate to date (Staff, 2021). This time frame offered a comprehensive dataset for our analysis. In compliance with ethical research practices

(Fiesler and Proferes, 2018; Pater, Fiesler, and Zimmer, 2022), usernames were anonymized despite the public availability of this data.

Data analysis

Our data analysis of the Tinder reviews, conducted through reflexive thematic analysis (Braun and Clarke, 2012b), involved the following steps to ensure a thorough understanding:

1. Creating Initial Themes through Thematic Analysis: Two coders started by conducting a line-by-line open coding of 900 reviews using reflexive thematic analysis. This initial phase involved each coder independently analyzing a different set of 200 random reviews to identify potential themes. The authors then discussed the identified categories and subcategories of themes, culminating in a preliminary codebook based on 400 reviews. To validate the comprehensiveness of these themes, both coders independently coded another set of 500 reviews. This process resulted in a codebook encompassing five categories and eleven subcategories of themes.

2. Filtering Irrelevant Reviews Using Keywords: To refine the focus on reviews that discussed algorithmic harms to relationship building, two coders extracted 90 keywords indicative of such harms (e.g., 'self-worth', 'confidence', 'ugly', 'fake', 'unreal', 'ban', etc.). These keywords were manually chosen based on their occurrence frequency in the manually coded reviews, in a manner akin to feature selection in (Voskobojnikov, Wiese, et al., 2021). We then utilized MAXQDA´s automatic coding function to filter the dataset, retaining only the reviews containing

these keywords. This step narrowed down the review corpus to 1,658 relevant entries.

3. Expanding and Revising Themes: With the refined dataset, we further applied thematic analysis using the automatic coding function to assign and expand themes across the remaining reviews. We rigorously reviewed the codebook for relevance and redundancy, focusing exclusively on problem statements that directly referenced 'algorithm', 'system', or 'automatic' issues. Reviews unrelated to the dating algorithm, such as those mentioning technical difficulties (e.g., installation or verification) or general complaints (e.g., about age restrictions, expensive subscriptions, or advertising), were excluded. This resulted in a structured thematic schema with three primary categories of harm and their respective subcategories.

7.4.2 Stage 2: In-Depth Interviews

Although the reviews provided us with a general understanding of the ways users perceive Tinder dating algorithms to damage users' intrapersonal and interpersonal relationships, in many cases these reviews were brief and did not offer a nuanced understanding of users' folk theories and mitigation strategies. To deepen our understanding from the reviews, we conducted 30 in-depth interviews focusing on users' interpretations of algorithmic harms, their mitigation strategies, and overall behavior on Tinder. This approach provided a richer and more balanced view of users' experiences, compared to the typically negative skew of app reviews (Voskobojnikov, Wiese, et al., 2021).

Sampling

When examining users' algorithmic folk theories and their motivations for using a dating application, users' technical and cultural backgrounds, as well as their ethnicity, are of critical importance (Eslami, Karahalios, et al., 2016). Therefore, we recruited our participants via multiple channels (i.e., 40% through Instagram ads targeting users in different countries, 30% via Telegram channels with international members, 10% through Tinder app ads, and 20% through university mailing list announcements across various faculties). This allowed us to target users with different backgrounds and to assemble a relatively balanced group of participants from Western and non-Western cultures (e.g., German, Albanian, Nigerian, Indian, Pakistani, Iranian). Since Tinder is mainly used by younger adults (Sumter and Vandenbosch, 2019), our sample also consisted of users between the ages of 23 and 44 year (average age 30). Our participant gender distribution included 13 cisgender women, 14 cisgender men, and 3 non-binary individuals. Regarding sexual orientation, 24 identified as heterosexual, while 6 were members of the LGBTQ+ community, comprising 2 gay, 2 bisexual, 1 pansexual, and 1 queer participant. The range of Tinder usage varied from 3 months to 5 years among participants, averaging approximately 1 year and 9 months. All participants had actively used the app in the weeks leading up to the study and had been on the platform for a minimum of two months. They were all located in the rural and urban areas of Germany. For a detailed breakdown of participant demographics, refer to Table **??** in the Appendix.

Procedure and Analysis

Employing a semi-structured interview approach (Drever, 1995), we developed a guide covering: (a) users' understanding and theories about Tinder's dating algorithms and how they generate and prioritize recommendations, (b) potential challenges associated with these algorithms, (c) their impact on users' online and offline relationships, and (d) strategies users employ to overcome related challenges. Although our research primarily focused on potential algorithmic harm rather than real-life physical harm, we took several precautions to protect participants from potential stress or discomfort. We gave participants the option to select between a video conference or an audio call for the interview, and informed them that they could choose to end or pause the interview at any point. Conducted online via Zoom and WhatsApp, these sessions were audio-recorded with consent and lasted 40 to 50 minutes. Participation was voluntary without compensation.

In the analysis phase, we began by transcribing the interviews, anonymizing the data by removing personal details, and assigning each participant an index (P#). Applying inductive thematic analysis (Braun and Clarke, 2012b), two researchers independently coded each interview in Nvivo. We focused on primary categories including nuances of perceived algorithmic harms and the conflict of interest theory alongside users' countermeasures. This iterative analysis allowed us to identify and define subcategories within these main themes. For instance, under the "conflict of interest theory," we cataloged specific descriptions of the theory and gathered supporting evidence. After completing this initial phase, we achieved a consensus and laid the groundwork for structuring, discussing, and interpreting the data. In the next phase, we utilized MAXQDA Analyzer to methodically

map all statements from the interviews onto our established categories and subcategories. The final phase involved a series of online meetings where we systematically refined our understanding of the relationships between the themes we had identified.

7.5 Findings

Our analysis from both stages yields insights into the types of algorithmic harms perceived by Tinder users, along with their folk theories and response strategies. The reviews enabled us to outline three main categories of algorithmic harms, addressing RQ1. The interviews not only provided depth and nuance to our understanding of these harms but also helped us to associate these harms with users' folk theories about the algorithms' underlying motivations. Furthermore, the interviews revealed users' strategies for mitigating harm and interacting with the Tinder dating algorithms. Figure 7.1 summarizes our findings, showcasing the types of perceived algorithmic harms (section 7.5.1), the algorithmic processes discussed in reviews and interviews and users' folk theories (Section 7.5.2), and their mitigation strategies (section 7.5.3). The following section will present our findings in detail, reporting the themes extracted from both reviews and interviews, with N_r indicating the number of statements related to a theme from reviews and N_p for those from interview participants.

Perceived algorithmic harm	Example
Damaging Self-esteem	The algorithm is perceived to damage users' self-esteem by limiting the number of matches and likes: "*I didn't have many matches or likes. You might think okay, that's not the algorithm's fault that people don't like me. But what if it is? What is this going to do to my self-worth?* " (P19).
Sabotaging Potential Relationships	The algorithm is perceived to sabotage the potential relationships by suddenly removing the matches from the matchlist: "*I was texting with a guy for some time now and we were planning to go on a date, but Tinder removed all my matches and he was removed too* "(P29).
Encouraging Anti-Social Behaviour	The algorithm is perceived to urge the users to block and report other users by repeatedly recommending the rejected profiles: "*Tinder kept suggesting a friend of mine and I had to block the profile so I wouldn't see her ...* " (P13).
Identity-based Harm	The algorithm is perceived to misrepresent certain gender identities: "*The binary nature of Tinder algorithms, representing users exclusively as either male or female, [...] imposes unnecessary pressure on us*" (P12).
Perceived algorithmic processess	**Example**
Throttling Profile Visibility	The algorithm is perceived as intentionally not recommending a profile to others:"*No matches or likes after a while, not even after a premium membership. The algorithm is crap*" (R4511).
Manipulating Users' Matches	The algorithm suddenly deletes the matches from the matchlist: "*Matches disappear, and it's not because people unmatch you. Do you know why? Because they reappear after some time*" (P17).
Recommending Large Quantities of Profiles that Will not Lead to Matches	The algorithm is perceived as recommending overly attractive profiles rather than those that fit: "*Tinder isn't where everyone finds their soulmate. It seems more like a showcase of attractive people [...] not those who might be perfect for each other...*" (P1).

The "Conflict-of-interest" folk theory: There is a contradiction between the main promise of dating apps to help users find the perfect partner, and these apps' commercial interests in retaining users on the platform:
"*The Tinder algorithm doesn't genuinely aim to connect people, as it contradicts their business model*" (P23).

Counteracts	Example
Optimizing Profile/Search Strategy	Users change the distance range to get better recommendations: "*I changed the filters and chose a broader range to have more recommendations* " (P26).
Knowing Your Competitors	Users observe and analyze the competitors by setting up a fake: "*I created a fake male account to see my competitors in the neighborhood and to evaluate myself* "(P10).
Engaging in Counter-intuitive Behaviors to Disrupt Unfavorable Algorithmic Processes.	Users "like" the repeatedly recommended profiles they do not genuinely like to let the person know they are not interested themselves: "*There was a guy who kept being suggested to me no matter how many times I swiped left [...] To get rid of him, I liked the man's profile and sent him a message myself*" (P3).
Leveraging Location-based Filtering for Match Variety and Safety	Users use VPN to manipulate their location based filtering: *"many start using VPNs to change their location. Sometimes, you match with them only to discover they're actually somewhere like Izmir"* (P5).
Bypassing the Algorithm by Moving to Other Platforms	Users move to other social media platforms to check the authenticity of a recommended profile: *"I just ask for the Instagram and Whatsapp account at the very beginning to see if the person is real"* (P20).

Fig. 7.1.: An overview of the perceived algorithmic harm, perceived algorithmic processes, users' folk theory, and users' counteracts.

7.5.1 The Perceived Algorithmic Harms to Intra- and Inter-Personal Relationships (RQ1)

Our analysis of online reviews uncovered various types of relationship harms, both at the intrapersonal level, a) damaging users' self-esteem and at the interpersonal level, b) sabotagi potential relationships, and c) encouraging antisocial behavior.

The findings from our interviews not only confirmed and enriched the insights from the reviews about the types of perceived algorithmic harm but also broadened the spectrum of harm. In particular, the interviews revealed d) algorithmic identity-based harm, specifically pertinent to our LGBTQ+ participants. This aspect was particularly valuable, given the lack of demographic information about the reviewers in Stage 1.

Damaging Self-esteem:

A common form of perceived intrapersonal harm that emerged from the reviews was self-esteem damage, primarily attributed to a lack of matches or likes (N_r=51). One review encapsulated this by lamenting, "*Tinder algorithm leads to competitive pressure, lack of self-confidence, frustration, and thereby destroys the dating world*" (R3962).

While the low number of likes and matches could be attributed to a profile's lack of appeal to other users, many reviewers pointed the finger at Tinder's dating algorithms, suggesting that the system intentionally did not recommend their account to other users sufficiently. Review 2809 provided a succinct example of this sentiment, stating, "*When I newly registered a while ago, there were likes in the first few days, but nothing has happened for*

weeks. Tinder can't convince me that everyone here finds me so ugly" (R2809). These abrupt shifts in match rates led many reviewers to the conclusion that Tinder's algorithms were designed to incentivize them to upgrade to a premium account. Review 6396 put it starkly, questioning, "*Can it be that you have massively limited the matches in the free version lately, or did I turn ugly in one fell swoop? I used to have regular matches; now, none at all.*"

Reviewers suspected that Tinder's algorithms were burying their profiles, making it hard for others to see them, and resulting in fewer matches. Review 5688 shared these frustrations: "*I'm getting zero matches after multiple days of swiping on 200 people a day!... I don't know what the heck is wrong with the algorithm, but I suspect my profile just got buried in oblivion, so no one ever sees me.*" Review 1843 criticized Tinder for exploiting users' insecurities to increase their revenue, expressing, "*Please, don't let anyone tell you you're unsuitable! This app thrives on your self-doubt so that you sign up for expensive subscriptions and Tinder makes money! Then I'd rather be single, but rich and happy*!" Reviews often interpreted this decreased visibility and subsequent decrease in matches as a hit to their self-esteem: "*If the algorithm is working great, then I'm too ugly*" (R2859).

In the interviews, as with the reviews, the harm to self-esteem due to a limiting number of recommendations and matches on the platform was a recurring theme (N_p=12). P8 experienced such significant frustration that they chose to leave Tinder entirely, feeling it was damaging their self-esteem. He shared, "*I compared my recommendations to my roommate's. Her suggestions were endless, and she would get a multitude of matches. I couldn't understand what was happening. It felt like it was undermining my self-esteem. So, I stopped using Tinder*" (P8).

Participant 19 also voiced a similar sentiment, "*I didn't have many matches or likes. You might think okay, that's not the algorithm's fault that people don't like me. But what if it is? Nothing hurts me more than when the algorithm pairs me with someone who has explicitly stated 'No BLACK PEOPLE' in their bio. What is this going to do to my self-worth?*". This sentiment suggests that due to a clear mismatch between the preferences of the recommended profiles and P19's profile, they concluded that the algorithm may not be presenting their profile to the right audience, leading to a low number of likes and matches.

Sabotaging Potential Relationships:

This type of perceived algorithmic harm emerges once a match is established, and users are poised to initiate a relationship. Many reviewers (N_r=229) have voiced their frustrations, pointing to a common issue - matches mysteriously disappearing from their list of messaging interactions. R431 lamented, "*Another match disappeared (so far #8) in the middle of a really nice get-to-know-you [messaging interaction] and supposedly no chance to recover it... Honestly, I am starting to feel like Tinder is doing this on purpose to keep me as a paying customer.*" This concern highlights a problematic scenario where users lose the opportunity to connect with potential partners due to communication channel failures. R513 refers to these algorithmic processes thwarting their chance of building a potentially successful relationship as anti-social, stating, "*Oddly enough, every time I get matched with a very pretty girl and she even texts me first, which has happened more than 10 times now, Tinder algorithm says no, that cannot be. The match disappears and so does the message. No matter how fast I reply, really antisocial.*"

Although a match's disappearance from the list can occur when another user unmatches themselves, reviews presented various reasons to believe this was a result of algorithmic processes, not human actions. Many cited the instantaneous nature of the disappearance, which happened too quickly for a human to react. R1528 noted, "*Many matches disappear within seconds. The other person couldn't have possibly unmatched so quickly.*" Some even observed the simultaneous vanishing of all matches, with R3210 expressing frustration: "*I just opened the app, and ALL my messages and matches are just gone... Are you kidding me? I haven't exchanged numbers with anyone yet, and now I'm sure I can't find them again.*"

Most significantly, the sporadic reappearance of matches led users to believe that Tinder's algorithms were sabotaging their potential relationships. R2565 remarked, "*Unfortunately, Tinder controls your matches because matches disappear and reappear.*" R400 shared a peculiar incident: "*I was texting a person, but suddenly they were no longer on my match list. Later, I received a message from that same person on Instagram asking why I had unmatched them, which I hadn't. How would you interpret that? Neither of us ended the match, so was it the algorithm? Why did you do that to us?*"

Sabotaging potential relationships was also underscored by interview participants (N_p=9). They offered more examples of perceived algorithmic harms and gave insights into why they believed these problems originate from the dating algorithms. For instance, P17 noted a peculiar pattern of disappearing and reappearing matches, stating, "*Matches disappear, and it's not because people unmatch you. Do you know why? Because they reappear after some time... Tinder wants to keep you engaged, but you have no control over the*

*situation [...] It happened to me several times. The worst part is, they are going to think you were a b**ch and just unmatched them*" (P17).

Expanding on this theme, P30 referred to another facet of this perceived sabotage and added a layer of complexity, pointing out that the harm created might be reciprocal and detrimental to other users' self-esteem, depending on the circumstances of the conversation. They said, "*I was texting with a guy for some time now and we were planning to go on a date, but Tinder removed all my matches and he was removed from my match list too. I feel sorry because he was so insecure and also younger than me, I don't want him to think he has been rejected*" (P29). Further voicing distrust, P24 added, "*I don't trust the algorithm there, it's so arbitrary. I'd rather stick to other communication platforms where at least you have the person's phone number and won't lose touch completely if something goes awry. With this, you're just left hanging with nothing.*"

These accounts suggest that users perceive the algorithmic processes as manipulative, jeopardizing their prospects of forming meaningful relationships. Notably, as emphasized by (P29), the perceived impact extends beyond individual users to potentially affect their conversation partners on the platform, underscoring a mutual potential for harm.

Encouraging Anti-Social Behaviour:

Reviews have brought to light how Tinder's algorithms might inadvertently encourage antisocial behavior by repeatedly recommending the same unwanted profiles as well as enabling abuse of reporting mechanism. The former is represented by the repetitious recommendation of profiles that users have already swiped left on or rejected. This may leave users with

little choice but to block or report others on the platform as a way to avoid seeing their profile yet again. For instance, R6938 stated: "*The same people are always shown, whether you swipe left or right! The user is urged to block the account so that it is no longer displayed*".

In addition, a significant number of reviews (N_r=205) have raised concerns about the potential misuse of the reporting feature on Tinder, and the perceived lack of enough algorithmic intervention to validate and mitigate baseless reports. Some reviewers believed that the platform automatically bans users based on unverified reporting, which could easily be exploited. R1846 shared a personal account of being reported and subsequently banned for merely refusing an in-person meeting, asserting that the developers do not authenticate the reports: "*I was just chatting with a guy, and I guess he got upset because I wouldn't meet him, and reported me. The developers don't even check if there is a valid reason. It feels like I'm forced to meet anyone who asks, or else they could report me...*" (R1846). This sentiment resonates with R6007, who suspects they were reported by a match they had rejected, resulting in an immediate ban without any verification: "*I was instantly blocked because I turned down a guy. Nothing was checked, and I simply couldn't log into my account afterwards.*"

Similarly, R3032 believes the report feature was abused by an ex-partner for personal vendetta: "*After my recent breakup, I reactivated my Tinder account and it was banned after a short time. I know that my ex-girlfriend and her friends reported me for personal reasons. Unfortunately, support is not responding and will not tell me how I am supposed to have violated the terms of use!*" (R3032). Moreover, R5922 highlighted the potential misuse of the automatic banning process, stating that it could be exploited by anyone

bearing a grudge: "*Theoretically, anyone who doesn't like you can automatically ban you; even if you had done nothing wrong to them, they will find your profile and just report you along with their friends*".

The interviews further elucidated the nuances of the algorithmic harm related to encouraging anti-social behavior, such as blocking other users on the platform (N_p=10). P9 elaborated on this effect: "*The only option to avoid being recommended to unwanted people is by either blocking or reporting them. So, people often resort to reporting others as soon as they find something they don't like.*" (P9). The interviewees also suggested that this anti-social behavior extends beyond the digital confines of the platform, leading to real-world bullying and harassment. This harm can occur simply from being recommended to nearby profiles that users wish to avoid. P25, for instance, recounted an experience where his classmates bullied their teacher after her profile appeared as a recommendation on their Tinder feed: "*She (the teacher) was suggested to one of my classmates; she wasn't unattractive, but my classmates circulated her pictures in public social networks and started bullying her*" (P25). P28 shared a personal ordeal related to workplace discomfort due to her Tinder profile still being active and suggested to others, even though she was banned on the platform: "*Unexpectedly, I found myself logged out and unable to log back in; I was blocked, yet my profile was still visible and being recommended to others. I discovered this when some colleagues started making me uncomfortable at work. They had seen my profile on Tinder and thought I was available, but now I'm in a relationship, so this gave them a misleading impression*" (P28). P29 resonated with P28's account, adding: "*It's really annoying when your profile is recommended to men in your social circle and they assume you're desperate for a relationship and are making yourself*

available for them. They approach you in real life and harass you. I've already had two such cases."

These narratives emphasize that some users interpret algorithmic recommendations on Tinder as an indication of the profile owners' availability and may approach them in real-life based on this presumption, potentially leading to harassment.

Identity-based harm:

Insights gathered from our interviews underscored a distinct form of harm from the ones mentioned in the online reviews: This harm, labeled as identity-based harm, was particularly salient to the LGBTQ+ participants, representing the damage resulting from misrepresentation or misclassification of certain identities such as gender and sexual orientation.

All of our LGBTQ+ participants mentioned feelings of being marginalized and excluded on the platform due to this problem. For example, P2 shared their frustration, saying, "*I feel marginalized on Tinder because it's literally like 'you're a man' or 'you're a woman.' There is no in-between. There is no space for others to express themselves if they don't fit into this binary*". P19 echoed this sentiment, stating, "*In society, we often feel like we don't belong, and I believe Tinder plays a nasty role in perpetuating these biases. I have a mustache, wear very masculine clothing, and am typically perceived as a man. However, in intimate situations like dating, I don't want that stereotype. Yet, Tinder presents me as male to others*" (P19).

P12 succinctly encapsulated the harm caused by misrepresentation of identity on Tinder: "*As a member of the LGBTQ+ community, our self-assurance can be*

fragile, often disrupted by seemingly minor incidents. Personally, being perceived strictly as a man or woman triggers feelings of gender dysphoria, an experience that generates insecurity, creates anxiety, and impairs my sexual arousal. The binary nature of Tinder's algorithm, representing users exclusively as either male or female, contributes to this issue by inaccurately representing our identities and imposing unnecessary pressure on us" (P12). P22 also expressed discomfort with the platform's lack of nuance: "*I always have to double-check everything with others to avoid misunderstandings: 'are you really gay? are you really into me?*" (P22).

Our queer participants have reported incidents where such harm transcends virtual interactions. P2 explained, "*Heterosexual men can make queer men feel very threatened, you know, we've been usually bullied by them as kids in school because we are a little effeminate or whatever. They are the ones who normally create this power imbalance and we feel uncomfortable in their presence. And Tinder algorithm is making this worse. Because we can't avoid heterosexuals seeing our profiles. I have friends who even got physically hurt by straight people*" (P2). P5 also shared, "*This is happening in India and also my friends from Pakistan have told me similar stories about religious folks targeting gay guys with fake profiles. They bait them with promises of a party, but the reality is way different—they either harm them or publicly out them at local religious places. With no way to verify profile authenticity, the whole thing becomes a big mess, leading to serious issues and potential harm for us*"(P5).

As these experiences illustrate, the perceived harm of the algorithm on dating platforms does not stop at misrepresentation; it potentially perpetuates harm inflicted by others through recommendation of profiles made for nefarious

purposes, leading to serious physical and emotional consequences for this already vulnerable group.

7.5.2 The “Conflict-of-Interest Theory”: Competing Interests and Clashing Values Between the Dating Algorithms and Users (RQ2)

Throughout the interview process, a theory repeatedly emerged among participants (N_p=16) to explain the root of the perceived harms by Tinder’s dating algorithms. We termed this the "conflict-of-interest" theory, which underscores a contradiction between the perceived "*promise*" of Tinder’s algorithms to help users find the right partner, and the app’s "*business strategy*" of retaining users on the platform. This theory carried the assumption that Tinder’s matching algorithm knows the "*right*" matches, yet intentionally sabotages them in order to retain users on the platform and spur them into paid memberships. For instance, P14 said: “*I don’t think the algorithm wants to show the user the right matches. They want us to stay on the platform by deceiving us with so many matches, none of which truly match. This encourages users to continue swiping*” (P14).

Participants across the interviews and reviews considered several perceived algorithmic processes as tangible evidence in support of their conflict-of-interest theory, which we describe in the following subsections (also listed at Table 2).

Throttling Profile Visibility

Some users believed that Tinder's algorithm intentionally does not show one's profile to nearby users after a certain period of time in order to persuade users into paid memberships to sustain their prior level of visibility (N_p=10). Some participants claimed to have proved this algorithmic process by creating new accounts and comparing recommendations with their older account. Per P14: "*It's clear that after a while, they don't show your profile to others anymore. You can create a new account and check. The number of recommendations changes a lot. Maybe that's a payment strategy.*" P20 echoed a similar sentiment:"*They don't show your profile to others after a while. I know because I create new profiles and then I have better recommendations. People start seeing my profile again. I guess it's to push you to pay.*" Some participants did end up paying for a premium membership, yet remarked that the quality of their profile recommendations did not improve, furthering the perception that the algorithm intentionally avoids matching compatible users. This thinking was also exemplified in the online reviews. Reviewer R4511 wrote, "*Stay away from Premium! Pure rip-off. No matches or likes after a while, not even after a premium membership. The algorithm is simply unsatisfactory; its main goal is to keep you engaged without delivering on its promise*".

Manipulating Users' Matches

Some users alleged manipulation and sabotage of their match lists by Tinder's algorithms (N_p=7), with the suspected intent of either coaxing users back onto the app with non-existent matches or deleting already-established matches that could have manifested into real relationships and eventual

account deletion. For example, P7 stated: "*Matches are added to your match list that you don't remember swiping right on, and then some just disappear and pop up again. Tinder just wants you to be on the app. Nothing else matters to them.*" P23 echoed a similar sentiment: "*I lost some of my matches. You just open your match list and they are not there. Maybe it seems like a punishment for not having a premium account.*" As P23 alludes to, these supposed algorithmic manipulations of match lists were interpreted by some users as subtle prodding by the platform to purchase a premium membership so their matches are not further tinkered with. Per R1371, "*The algorithm is so annoying. I get notifications that I have matches but nothing is there. Or they have vanished in thin air. The app wants to tell me that I need a subscription... A huge disappointment; they only want your money*". Yet other users believed match list manipulation still occurs with paid memberships to ensure that users do not successfully start a relationship and leave the platform: "*Tinder intentionally removes the matches to keep you paying... Rip-off appears to be part of their business model*" (R6332).

Recommending Large Quantities of Profiles That Will Not Lead to Matches

Users across interviews and reviews spoke to a suspicion that the matching algorithm prioritizes profile recommendations that are unlikely to result in a romantic connection so that users are kept active on the platform. There were three related algorithmic processes that users identified: recommending profiles that are too attractive for the target user, repeatedly recommending the same users - including fake profiles, and ignoring recommendation preferences. Many found the profiles they were presented with to be "*too*

good to be true" (P3) or "*overly attractive*" (P29), thus making it unlikely that recommended users would reciprocate their interest and establish a match (N_p=15). P1 encapsulated this feeling: "*Tinder isn't where everyone finds their soulmate. It seems more like a showcase of attractive people with long waiting lists, not those who might be perfect for each other. It's their business strategy to keep you on the app*"(P1). Similar sentiments were reflected in reviews, with one user questioning: "*For a while now I seem to only get profiles from the absolute most attractive users who probably have hundred matches anyway. Is this a new marketing strategy?*" (R2757).

Some users described repeated recommendations of the same profiles, seemingly as a way to pad the quantity of profiles shown without bringing the user closer to establishing a legimate match. As P10 described, "*It's always the same profiles they show you over and over, that's their business model—to show you some recommendations rather than nothing*" (P10). This issue was particularly frustrating when users were shown profiles they had previously rejected. Relatedly, participants (N_p=7) expressed concerns about the prevalence of fake profiles, with some asserting that Tinder intentionally does not want to remove these profiles so that it can inflate the number of recommendations to a user. They identified profiles as potentially fake based on suspicions of photoshopped pictures, blank profile photos, and inauthentic behavior. P21 observed: "*Some profiles are obviously faked. For example, with a photoshopped or blank profile picture. But Tinder doesn't remove them. They seem to want these profiles to exist to increase the number of recommendations*" (P21). This perception of fake profiles contributing to user retention was echoed in online reviews as well: "*Constant matches... but you get NO ANSWER from the profiles!!! Or just one reply, then none for days... Either they are fakes or*

they're simply incapable of writing. I feel rather fooled... They only want you to keep paying them" (R1340).

Others suspected that the rash of unfit profiles they were being shown was due to the algorithm purposely disregarding explicit preferences for profile recommendations (N_p=22).

For example, P18 described their age preferences being ignored: "*They regularly suggest women outside of my specified age range. It seems they prioritize keeping the number of profile recommendations high, rather than ensuring they meet my criteria*" (P18). Users were also frustrated when recommendations included profiles far outside their specified distance range, even from other countries or continents. P7 added: "*Sometimes, the recommended profiles were from China or Russia. It's clear that I won't travel around the world to meet someone*". Queer participants, in particular, faced unique challenges and frustrations in setting profile recommendation preferences because several important settings do not exist on the platform. Some users believed such preferences could be inferred by the algorithm through user behavior, yet were being intentionally ignored. P22 expressed such frustration: "*For the LGBTQ+ community, there are specific characteristics that matter a lot, such as whether someone is a top or bottom. But you can't filter for that on Tinder. I get a lot of matches, but they don't align with my preferences. I thought maybe the algorithm would learn from my interactions, but it doesn't. After some swiping, it even starts to suggest female users to me, even though that's not what I want. It's like Tinder has its own agenda to keep me on the platform*" (P22).

7.5.3 Working With, Through, and Around the Tinder Dating Algorithms to Regain Agency (RQ3)

The perceived harms caused by the Tinder dating algorithms, combined with doubts regarding the alignment between the algorithms' "*promise*" of romantic matches and its financial incentives, led users to engage in a struggle to regain agency and control over their dating experiences. This included working "*with*" the Tinder dating algorithms to increase the number of suitable matches, working "*through*" the algorithms to counteract the unfavorable algorithmic actions, and working "*around*" the algorithms by bypassing their influence altogether. Below, we explain these strategies and how they were applied by users to mitigate, reduce or prevent perceived algorithmic harms.

Resetting the algorithm with user account changes

Users tried to manipulate their profile (e.g., by adding more photos) or their search strategies (e.g., by expanding the filters) in an attempt to regain control of their profile visibility and possibly get more recommendations. For example, some participants (N_p=13) believed that they could improve the quantity of profiles recommendations by selecting a broader distance and age range: "*I changed the filters and chose a broader range to have more recommendations*" (P26). Others (N_p=7) believed that Tinder's algorithms would recommend "*more attractive*" profiles to users who uploaded more pictures of themselves: "*When I uploaded more photos of myself, I felt that the men who were suggested to me were also more attractive*" (P7). Logging in and

out of the platform or creating a new account (N_p=4) were also mentioned as methods to "*reset the algorithm*" and "*obtain better matches*" (P1).

Knowing Your Competitors

One strategy for mitigating algorithmic harm to self-esteem involved being aware of "competitors" in the area - referring to other users in their geographic area who would be interested in the same pool of available daters. Many participants (N_p=15) used the platform to observe and analyze competing users in the neighborhood, and some optimized their account accordingly. While Tinder's matching algorithm is designed to only allow recommendation of profiles that match one's preferences, users circumvented this restriction by creating a fake profile of the opposite sex, changing their sexual interests, or using another user's account to view profiles of nearby users. For example, P10 explained, "*I created a fake male account to see my competitors [other women] in the neighborhood and to evaluate myself, but others' [profiles] were not so good. So I felt reassured. I thought my chances were high*" (P10). P15 mentioned that he created a fake female profile with some of his friends "*to get a better idea of other male profiles*" and even "*texted with them to find out what online dating strategies they had and feel more confident.*"

Engaging in Counter-Intuitive Behaviors to Disrupt Unfavorable Algorithmic Processes

When participants encountered unfavorable algorithmic acts, some engaged in counter-intuitive behaviors to counteract them such as disliking profiles that they find attractive, liking profiles without genuine interest, and blocking users who have not done anything wrong. For instance, P16 mentioned that

he never swipes right on profiles that are "*too attractive*" for him, believing that they are unlikely to reciprocate his interest and that his "*ranking will decrease by the algorithm*" as a result.

When the algorithm persistently recommended profiles they had previously disliked, some participants decided to like these profiles and handle the situation themselves by conveying lack of interest via messaging. P3 is a representative example of this group: "*There was a guy who kept being suggested to me no matter how many times I swiped left on him, so I figured he must have a paid account. I took a screenshot of his profile, showed it to my roommate, and asked her if he kept getting suggested to her too, and she said yes. So, to get rid of him, I liked the man's profile and sent him a message myself to let him know I was not interested*" (P3).

Other users decided to block these persistent profiles despite the profile owners having done nothing wrong, as a last resort to removing the profile from being continually recommended. As P13 explains: "*Tinder kept suggesting a friend of mine and I had to block the profile so I wouldn't see her and she wouldn't see me. But I'll never know if she saw that I blocked her. Hopefully not, because she was my friend*" (P13).

While P3 and P13 both engaged in counter-intuitive behaviours on the platform, P13's action can be perceived as antisocial behaviour, which we previously discussed as a type of harm to other users.

Leveraging Location-based Filtering for Match Variety and Safety

Some participants contended that Tinder's dating algorithms create socioeconomic bubbles. To circumvent these algorithmically constructed bubbles,

many participants (N_p=17) opted to use Tinder in different neighborhoods or larger cities. For instance, P11 shared, "*I visited the luxury district with my friends and used Tinder there. It's a fantastic way to meet guys from those neighborhoods*" (P11). P4, a resident of a rural area, also mentioned that he hacks Tinder's dating algorithms' location-based filtering by using the app whenever he is "*in big cities nearby*" in order to "*meet girls from big cities*" (P4).

For queer participants, the harm caused by location filtering could be significantly greater if they found themselves in an environment or country that is not "*queer-friendly.*" P2 and P9 mentioned refraining from using Tinder in certain places or during specific periods if they were in a "*toxic environment*" due to the fear of harm or bullying. As P2 explained, "*Being part of the LGBTQ+ community, I've had my fair share of bullying, so I'm really careful with Tinder to protect myself. I was once in a really toxic workplace where I just didn't vibe with the people around me. The thing is, Tinder's algorithm doesn't let you exclude specific people, so I stopped using the app when I was in that situation. I didn't want to give them another reason to target me*" (P2). P5 revealed that it is a common practice for them to use a VPN to circumvent the location filtering of the algorithm, "*The location-based filtering in the algorithm can pose a real issue for the queer community. Things can get risky in places like India, Turkey, or Pakistan, where using Tinder can seriously jeopardize people's safety, families, and jobs. That's why many start using VPNs to change their location. Sometimes, you match with them only to discover they're actually somewhere like Izmir. It's fascinating to see how individuals navigate the system for their safety*" (P5).

Bypassing the Algorithm by Moving to other Platforms

To prevent perceived sabotage to their potential relationships, participants attempted to bypass Tinder dating algorithms by switching communication to other social media platforms or exchanging phone numbers. Although this might happen naturally as users develop relationships with one another, we found some evidence that participants (N_p=13) might do this very early in order to verify an account's authenticity (in part due to distrust that Tinder's algorithms want to remove fake profiles) or to prevent Tinder's algorithms from deleting matches. For example, P20 preferred to manually conduct account verification by finding the respective person on other social media platforms: "*I don't ask for other photos, I don't even care if the account is verified or not; that doesn't help, I just ask for the Instagram and Whatsapp account at the very beginning to see if the person is real*" (P20).

P17 referred to authenticity concerns too, but also the fear of Tinder's algorithms sabotaging interactions that occur on the dating app: "*The first thing you do on Tinder is find the person on Instagram and Whatsapp, Instagram because you want to check if the person is real or not, and Whatsapp because there you have control there and you do not have to be afraid of losing contact because the match disappeared or something*" (P17).

7.6 Limitations and reflections

This paper builds on existing research about the harms experienced by users on dating apps by highlighting the perceived algorithmic harm to users' relationships and their strategies for mitigating it. However, the study's sample

and methodology have inherent limitations. Firstly, we limited our focus to the Tinder dating app. This decision was made because each dating app employs different algorithms Ranzini and Lutz, 2017, and broadening our study across various platforms could have introduced complexity. Secondly, due to legal constraints in [country] regarding data scraping and the requirement to log into a personal account to access App Store data, we confined our review to publicly available data from the Google Play Store. Lastly, we acknowledge that app reviews do not encompass the complete user experience and may reflect a limited range of perspectives, alongside the challenge of inaccessible user demographics.

We aimed to address some of these limitations with the second part of our methodology: semi-structured interviews. These interviews revealed many blind spots not apparent in the reviews, allowed us to engage with iOS users, and ensured a diverse demographic representation. Notably, our inclusion of the LGBTQ+ community led to the identification of a unique type of perceived algorithmic harm specific to these communities, thus enriching our findings. However, we must acknowledge that our sample was predominantly cisgender, meaning the interviews provide only a partial representation of Tinder users. More involvement from trans users, who are highly impacted by algorithmic harm DeVito, 2022 and face unique self-presentation challenges in online dating Fernandez and Birnholtz, 2019, may yield additional perceived algorithmic harms and should thus be a focus of future work. Additionally, our data does not differentiate between premium and free users of the app. This is due to the lack of user account information in reviews and the tendency of interviewees to switch between account types. Moreover, while gender could play a significant role in the experience of algorithmic harm,

our study did not delve into gender-specific impacts, as this was beyond our research scope.

It is also important to note that the algorithmic harms mentioned in this paper are those that users perceive and attribute to the algorithm, and they may not necessarily reflect what is actually happening on the platform. Nevertheless, our research underscores the importance of considering these perceived harms when trying to fully understand user behavior and strategies on these platforms. Therefore, we urge the research community to continue investigating this topic. As dating app algorithms are continuously evolving, new types of perceived computer-mediated harms may emerge over time.

7.7 Discussion

In this study, we have broadened the existing view of potential harms on dating platforms by examining the algorithmic harms perceived by users at both personal and interpersonal levels. Given the extensive research on human-inflicted harm on dating apps, adopting an algorithmic lens to interpret perceived harms offers a fresh approach. This not only complements existing perspectives but also provides crucial insight into users' behavior and counter-strategies on such platforms. To further illustrate this, we first connect our discovered conflict-of-interest folk theory of Tinder's algorithms to the broader algorithmic folk theory literature to elucidate key differences in how users construct and respond to perceptions of algorithmic processes on dating apps (which users intend to eventually stop using after achieving their goals) and social media platforms that are intended for continual use. We then outline how our findings on algorithmic harms in online dating

build upon and broaden existing research on algorithmic harms in other contexts. Subsequently, we explore the interplay between human-inflicted and algorithmically-inflicted harm, demonstrating how algorithmic processes can mediate and perpetuate pre-existing human-inflicted harms and social biases. Finally, we examine current design approaches for harm mitigation to discuss the implications our study may hold for designing user experiences that are mindful of potential algorithmic harm on dating apps.

7.7.1 The Conflict-of-Interest Theory, Its Implications, and Associated User Behavior

In the face of perceived algorithmic harms, users constructed a folk theory about Tinder's algorithms that we refer to as the conflict-of-interest theory. The theory references two interests: the interest of users in finding a compatible romantic partner, and the the financial interests of Tinder as a company in sustaining an active, paying userbase. Mechanistically, users in our study believed that Tinder's algorithms were capable of computing romantically compatible matches, yet were engineered to deliberately avoid recommending such profiles together, or to otherwise sabotage promising interactions, to coax users into purchasing premium memberships and continuing their use of the application. This theory diverges from prior research into folk theories about Tinder's algorithms, which attributed profile recommendation to romantic compatibility and user feedback (Abel, Pei, et al., 2022) (in direct contradiction with our participants' folk theories), as well as pure randomness and desirability scores (S. A. Huang, Hancock, and Tong, 2022; Nader and Lee, 2022).

When looking at the broader algorithmic folk theory literature, our Tinder users' speculation of financially-driven algorithmic processes draws comparison to an *"economic motivation"* or *"cash grab"* suspected by Twitter users regarding a rumored change to its content curation algorithm (DeVito, Gergle, and Birnholtz, 2017). Yet the impetus of this suspicion is quite different - it is not a rumored change to an algorithm that is prompting the conclusion of financial motivation, but a user's own dating failure. This is an important difference because dating apps have a clear win state for many users that would result in discontinued app-use (finding a long-term partner) whereas social media platforms do not. Continued usage is actually desired by users of other social platforms studied in the algorithmic folk theory literature, such as goals of social support and connection on Facebook, Twitter, and Tumblr (DeVito, 2021), as well as visibility for the sake of helping LGBTQ+ people on Tiktok (DeVito, 2022). From our findings, suspicion of economically motivated algorithms on Tinder seems to uniquely derive from intent to justify why one's goals for using (and eventually leaving) a dating app have not been achieved.

User behavior on Tinder discovered in response to their conflict-of-interest theory would appear surprising in light of prior algorithmic folk theory literature. Devito used the concept of platform spirit to explain users' decisions to adapt to perceived algorithmic processes on various social platforms (DeVito, 2021, 2022). Platform spirit is defined as "*the user's perception of what a platform is and what it is for, as determined by the user's understanding of the platform's stated mission, its values and actions in practice over time, and the functionality which it allows as juxtaposed with the user's understanding of the platform's purpose*" (DeVito, 2021, p. 18). DeVito goes on to explain

that positive platform spirit - or algorithmic processes that users consider aligned with the platform's purpose - prompt users to continue platform-use and adapt to its algorithmic processes. Negative platform spirit - or algorithmic processes that users consider to conflict with the platform's purpose - motivate "*decisions to reduce or stop use of a system*" (DeVito, 2022, p. 6).

The perceived platform spirit of Tinder can more aptly be called a platform *promise* - a word that came up multiple times in our data - to connect users with a compatible partner. Users in our study overwhelmingly considered Tinder's algorithms to violate this platform promise, yet we saw little evidence of them leaving the platform or limiting participation, as DeVito's empirical research on social media platform spirit would suggest (DeVito, 2021, 2022). Perhaps the only example in line with this prior work is some Tinder users moving their messaging interactions to other social platforms early in the relationship-building process to avoid algorithmic sabotage of their match list. This complements impression management and formation strategies found in prior online dating research, in which users transitioned their interactions to other platforms - sometimes faster than they were comfortable with (Zytko, Grandhi, and Jones, 2014) - as a way to better evaluate compatibility and safety risks with a potential meeting partner (Gibbs, Ellison, and Lai, 2011).

A reason Tinder users may choose to stay, and even continue paying for memberships, is because the perceived platform promise can divest them of personal responsibility for their dating failures and instead attribute that responsibility to Tinder's algorithms. The conflict-of-interest theory is thus not simply an algorithmic folk theory, but a folk theory about why online daters have been unsuccessful at finding their desired partner in a way that places responsibility for that lack of success on the platform's broken

"promise" rather than their own actions or attractiveness: it's not me, it's the algorithm. The suspicion of economic motivation can then be understood as a hypothesis for why Tinder's algorithm would intentionally break its supposed promise.

The counteractions that we found Tinder users to engage with on the platform in response to their conflict-of-interest theory can be described as "*unfaithful adaptations*" (DeVito, 2021), or behaviors that run contrary to the intended use of the platform. Intentionally "liking" profiles that one is not attracted to and blocking users that have not done anything wrong would be two examples from our study. While practice of unfaithful adaptations in prior work appears to be for "experimentation" purposes (DeVito, 2021, p. 24) to test a revised algorithmic folk theory, we did not find much evidence of Tinder users experimenting or stress testing their perceptions of algorithmic processes - potentially because doing so would risk the conclusion that they, rather than the algorithm, are the reason they are not getting matches or dates.

If not experimentation, the unfaithful adaptations found in our study align with what Velkova and Kaun refer to as "algorithmic resistance," a form of complicit resistance that acknowledges the power of algorithms but operates within their framework for different ends (Velkova and Kaun, 2021). We see this exemplified with creation of new accounts to sidestep a perceived algorithmic throttling of profile recommendations as well as manipulating the range of profile recommendation filters; two strategies also evidenced in other online dating folk theory research (Abel, Pei, et al., 2022; Nader and Lee, 2022; Ward, 2017).

7.7.2 Synthesizing Perceived Algorithmic Harms in Online Dating with the Broader Algorithmic Harm Literature

This section synthesizes Tinder users' perceptions of algorithmic harm, identified in this study at both the intra- and interpersonal levels, with the broader themes found in algorithmic harm research. We particularly discuss the types of harms our study revealed, and how they connect to and expand the algorithmic harm literature.

At the intrapersonal level, the negative impact of dating apps on self-esteem aligns with well-documented effects of these platforms on mental health, self-compassion, and self-worth (e.g., Malz, 2020; Musan, 2020; Toma, 2022). Furthermore, it draws parallels with "allocational harm" from the algorithmic harm literature, where automated systems inequitably distribute resources or opportunities (Andalibi, Pyle, et al., 2023a). This type of harm is exemplified by dating platform algorithms that reduce profile visibility and the number of matches and likes as usage time increases, ostensibly to motivate users to buy premium services (A. Brown, 2018). This practice shares similarities with DeVito's description of how social media platforms manage content distribution, setting criteria for the visibility each piece of content receives (DeVito, 2022). However, in the dating context, users often internalize the scarcity of matches and likes as personal failings, describing themselves as "ugly" or "unsuitable" - a recurring theme in our findings. This internalization forces a dichotomy: remain "single" or spend money on premium features to enhance perceived desirability and self-esteem. This strategy not only damages self-worth but also perpetuates a system that disadvantages certain

groups, reflecting broader concerns about unfair resource distribution seen in other sectors like the gig economy (Z. Muller, 2019).

At the intrapersonal level our LGBTQ+ participants overwhelmingly expressed that the Tinder algorithm's binary filtering and its failure to consider their specific sexual preferences resulted in mental and physical harm. They detailed feelings of marginalization, frustration, anxiety, and insecurity, sometimes manifested in physical side effects. Supporting our observations, the broader social platform literature, including findings by Walker et al., DeVito, Phan, and Duguay (DeVito, Walker, and Birnholtz, 2018; DeVito, 2022; Duguay, 2017, 2016; Phan, Seigfried-Spellar, and Choo, 2021; Walker and DeVito, 2020), demonstrates that LGBTQ+ individuals often face identity stigmatization, invalidation, and erasure. In algorithmic harm literature, this perceived suppression and misrepresentation of specific gender and sexual identities by the algorithm, aligns with what Karizat, Delmonaco, et al., 2021 refer to as "algorithmic representational harm".

Unlike scenarios where LGBTQ+ users reduced their use of platforms, left them, or employed strategies like "*identity flattening*" or "*identity modulation*" (DeVito, Walker, and Birnholtz, 2018; Duguay, 2017; Walker and DeVito, 2020) to adjust their identity presentation to match platform norms or avoid stereotypes, or even attempted to "*domesticate*" the algorithm (Simpson, Hamann, and Semaan, 2022; Simpson and Semaan, 2021), our study highlights a different approach. We found that our participants utilized location-based filtering to mitigate harm and enhance safety, an approach that reflects broader usage trends on social media platforms, as discussed in DeVito, Walker, and Fernandez, 2021; Haimson, Brubaker, et al., 2015. However, in the context of dating apps, targeting the "*outright*" audience takes on

additional significance due to the platforms' heavy reliance on geographic location and specific user affordances. For example, our participants employed strategies such as using VPNs to change their location to less conservative areas or strategically using the app in queer-friendly environments. This is particularly important in areas where there is a risk of physical violence. While previous research (Birnholtz, Fitzpatrick, et al., 2014; Hardy and Lindtner, 2017) has shown the risks associated with location-based filtering, our findings suggest it can serve as a protective measure for LGBTQ+ users to mitigate potential harms.

The countermeasures found in our study to combat these perceived algorithmic harms echoed prior work, such as profile manipulation found by (Nader and Lee, 2022), but also revealed seemingly counter-intuitive behaviors, such as disliking profiles one is attracted to or blocking users who did nothing wrong. These actions could indirectly affect other users; for example, if Tinder misinterprets a block as a response to misconduct, it could lead to an account suspension or ban. This possibility extends the discussion beyond the "*latent harm*" concept described by (Walker and DeVito, 2020), which focuses on harms that develop over time due to coping strategies affecting access to support and resources, primarily impacting the mental and physical health of bi+ individuals. Our findings suggest that such behaviors could also inadvertently impact a broader user base, going beyond intrapersonal harm as it can cause interpersonal harm as well; e.g. by potentially restricting other users' access to platforms like Tinder, reflecting unintended consequences of these coping strategies.

7.7.3 The Interplay Between Human-Inflicted and Algorithm-Facilitated Harm

Prior research on dating apps has documented both intentional interpersonal harm (such as stalking, harassment, and even sexual violence (e.g., L. Blackwell, Dimond, et al., 2017; Doerfler, Forte, et al., 2021; Freeman, Zamanifard, et al., 2022; Jhaver, Ghoshal, et al., 2018; S. Kim, Razi, et al., 2021; Mladenović, Ošmjanski, and Stanković, 2021; Rubin, Blackwell, and Conley, 2020) and unintentional interpersonal harm (resulting from sexual scripts (Comunello, Parisi, and Ieracitano, 2020; Race, 2015) and consent practices Zytko, Furlo, et al., 2021). Our findings show that perceived algorithmic harms not only portray algorithms as additional "entities" that users feel the need to protect themselves from, but as enablers and facilitators of human-inflicted harms. In other words, humans and algorithms are not distinct sources of harm, but interconnected. There are several examples of this across our study.

Harassment through dating apps is well known (Albury, Byron, et al., 2019; Anderson, Vogels, and Turner, 2020; Shaw, 2016), yet our study adds a new dimension of understanding harassment through the role that algorithms may play in enabling it. Interview participants recounted several instances of being harassed due merely to discovery of their profile by others in inopportune or inappropriate social contexts. Examples included a teacher being harassed by students who were recommended her profile, a user being judged by colleagues who discovered their profile, and identity-based harassment to LGBTQ+ users in culturally conservative areas. Furthermore, our findings suggest that Tinder's matching algorithm may be *instigating*

harassment. For instance, research on hostile masculinity and "incels" finds that men sometimes engage in sexual harassment as retaliation for being "failed men" incapable of dating success (Halpin, 2022). Perceptions in our study that Tinder's algorithms may be throttling profile visibility and sabotaging relationship success could reinforce feelings of inevitable dating failure, which could be directed at other users in the form of harassment (harassment in retaliation for romantic rejection has been found elsewhere (Shaw, 2016).

Our findings also indicate that Tinder's algorithms may be augmenting capabilities for harm by malicious actors. The creation of fake accounts as a vehicle for hate crimes is particularly problematic. While dating apps typically offer some form of user verification (e.g., profile picture verification), use of such features is typically not obligatory and certainly not foolproof as evidenced by our LGBTQ+ participants who reported instances of being coerced into face-to-face encounters by fake profiles. This finding aligns with previous studies examining identity-based harm targeting gay individuals (Birnholtz, Rawat, et al., 2020; C. Blackwell, Birnholtz, and Abbott, 2014; Corriero and Tong, 2016) and deceptive practices (Whitty, 2015). Interestingly, protective features like the ability to block or report users were sometimes viewed as tools to be exploited by malicious actors. Users in our study suspected reporting features were being abused to trigger automatic account banning without proper investigation - although we should note that there is no evidence to our knowledge confirming this is actually how Tinder's banning process works.

7.7.4 Scrutinizing Perceived Algorithmic Harms and Related Algorithmic Processes

Our study focuses on users' *perceptions* of algorithmic harm on Tinder, which could be based on an inaccurate understanding of how Tinder's algorithms really work. While perceptions of algorithmic processes and harms can influence user behavior regardless of their accuracy (DeVito, Birnholtz, et al., 2018; DeVito, 2022; Simpson, Hamann, and Semaan, 2022), it is important to scrutinize the veracity of our Tinder users' perceptions of algorithmic processes before considering design implications to combat the supposed harms that follow from them. While Tinder remains relatively opaque about the exact workings of their algorithms, information made available by the company does provide some insight.

Tinder acknowledges that its matching algorithm does rank profiles, but not based on the Elo Score. Instead, profile activity is the key determinant, with the company stating, "We prioritize potential matches who are active, and active at the same time" (Tinderpress, 2022a). However, this claim is contradicted by Courtois et al., who used the Experience Sampling Method with 88 Tinder users to infer algorithmic processes (Courtois and Timmermans, 2018). Their findings suggest that usage frequency does not significantly influence the likelihood of receiving likes or matches. Moreover, they observed a counter-intuitive trend: "the more recurrent the usage of Tinder, the fewer the number of matches." These insights lend credence to users' perceptions of algorithmic allocational harm, particularly those targeted at long-term users of the app, suggesting that algorithmic processes may indeed be throttling a profile's visibility.

Tinder's own website also lends credibility to users' perceptions that their match preferences are ignored by the recommendation algorithm. Tinder acknowledges that its matching algorithm operates not only based on explicit filtering but also machine learning techniques that involve various factors to determine profile recommendations such as users' "likes and nopes" and "anonymized cues from photos" (Tinderpress, 2022a). In other words, users' filter preferences are not treated as absolute requirements for profile recommendation, but one set of several different factors used to determine which profiles to show.

There are alternative explanations for other experiences reported by users that were suspected to be the result of algorithmic processes. A notable one was the perception that Tinder's algorithms sabotage relationship formation by deleting messaging interactions that seem to be going well. This could simply be the result of a lack of interest from other users. Deleted messaging interactions could be due to a messaging partner voluntarily choosing to unmatch, rather than an intentional effort by an algorithm to sabotage relationship success. Relatedly, an absence of matches may really be due to other users finding one's profile unattractive, rather than profile visibility throttling. Another explanation is that Tinder's algorithms could have detected one's messaging partner to be a bot or a sex worker (neither of which are allowed according to Tinder's Terms of Service), and deleted such accounts without informing matched users of this detection. Tinder does acknowledge use of "automated decision-making and profiling" as part of their moderation efforts (for example, removing bots). The platform also employs automated tools such as safe message filters to identify and manage instances of harmful or illegal behavior, thereby preserving user privacy and security (Tinderpress,

2022b). Lastly, human-driven moderation policies and practices could be the cause of what participants believed to be algorithmic sabotage.

One reason users may be quick to attribute negative online dating experiences to algorithmic processes - however outlandish - is as a scapegoat for one's dating failures. In this light, blaming algorithms for dating failures could be understood as a coping strategy - we refer readers back to our reflection on users' decisions to remain on Tinder despite a perceived broken *promise* by its algorithms.

7.7.5 Implications for Designing Algorithmic Processes in Dating Apps

Designing dating apps to mitigate risk of harm is integral to the user experience, especially for individuals such as women and LGBTQ+ users who are at disproportionate risk of dating-related harm (Albury, Byron, et al., 2019; Aljasim and Zytko, 2022; Fernandez and Birnholtz, 2019). Users' experiences with risk and actualized interpersonal harm in online dating has informed research a variety of technical solutions to counter such harm e.g., nude image detector and harassment detectors (Simon, 2021; Thompson, 2022). However, there has been relatively little consideration of how dating apps can be designed to mitigate adverse algorithmic processes, largely due to an absence of empirical investigation of perceived algorithmic harms and associated user strategies.

The discovered conflict-of-interest theory may render this a seemingly futile task - can we expect dating app companies to alter design to mitigate

algorithmic processes that generate financial returns? Yet we would point out that the algorithmic processes that users *perceive* may not actually be occurring, and so we hesitate to propose counteractions to these supposed (and potentially nonexistent) algorithmic processes regardless of whether companies may be willing to adopt them. Nonetheless, we would argue that user and company interests are not always mutually exclusive, and there are potential modifications or additions to known algorithmic processes that could benefit both parties. Malicious acts enabled by (the absence of) algorithmic processes such as recommendation of fake profiles made by people that want to coerce marginalized groups into face-to-face meeting clearly benefit neither party - a user that leaves the platform due to harmful experiences neither generates profit nor a dating success story. Improving algorithmic detection of fake profiles, and improving matching algorithms to avoid recommendation of users with harmful beliefs against marginalized groups *to* said groups could foster a more enjoyable experience that leads to extended platform-use.

And while we cannot reliably assert that Tinder's algorithms do indeed inflict the harms attributed to them such as self-esteem damage, relationship sabotaging, and identity-based harm, such categories can be used as goals for new algorithmic implementations. For example, algorithms could help remedy negative impacts to self-esteem due to absence of matches through suggestions for profile construction (e.g., users that include a headshot are x% more likely to receive a match). To help users maximize the chances of a successful online interaction that leads to a face-to-face date, users could engage in AI-mediated messaging interactions with suggested conversation topics. To avoid identity-based harm and marginalization, algorithms could

provide more granular demographic details to facilitate more relevant profile recommendations.

7.8 Future work and conclusion

On online platforms, algorithms significantly influence and mediate our relationships. However, the bulk of research on dating platforms has focused on human-inflicted harm, leaving a gap in understanding the potential harm users perceive from algorithmic interventions. To fill this void and provide a comprehensive, empirical understanding of how users perceive algorithmic harm, how they make sense of it, and the strategies they apply to counteract it, we conducted an extensive analysis of 7,043 Tinder reviews. This was complemented by 30 semi-structured interviews with users of the platform.

Our study unveiled various categories of perceived algorithmic harm affecting both personal and interpersonal relationships. The perceived algorithmic harms ranged from damaging users' self-esteem, sabotaging potential relationships, encouraging antisocial behavior, to inflicting identity-based harm. We found that users formulated a conflict-of-interest theory to rationalize these perceived harms, suggesting a contradiction between a dating app's mission of finding the perfect partner and its commercial interests. This theory was supported by several user-perceived algorithmic processes such as throttling profile visibility and manipulating user matches. Additionally, we found that this belief influenced users' strategies to counter algorithmic harm, resulting in counter-intuitive behaviors on the platform or circumventing it altogether.

Our findings also reveal that users perceive algorithmic processes as enabling human-inflicted harm, suggesting that algorithmic and human-inflicted harms can often intertwine. This underscores our advocacy for future research on computer-mediated harm mitigation to consider both human and machine factors in their investigations. We believe this research will lay the groundwork for future studies exploring the potential harms caused by algorithmic processes across various societal platforms from the users' perspective. Furthermore, we encourage future work to build upon this study by considering and testing the design recommendations proposed herein, aiming to better support users in mitigating the potentional algorithmic harms, especially those which detrimentally affect users' dynamic relationships.

7.9 Appendix

Tab. 7.1.: An overview of the participants' demographics

PID	Age	Gender	Sexual orientation	Ethnicity	Cultural Background	Tinder usage
P1	29	Cisgender woman	Heterosexual	White	German	2 Years
P2	27	Non-binary	Pan-sexual	Asian	Chinese/American	3 Months
P3	32	Cisgender woman	Heterosexual	White	German	9 Months
P4	29	Cisgender man	Heterosexual	Black	Nigerian	2.5 years
P5	30	Cisgender man	Gay	Asian	Indian	5 years
P6	25	Cisgender woman	Heterosexual	White	German	3 Months
P7	29	Cisgender woman	Heterosexual	White	German	2 Years
P8	27	Cisgender man	Heterosexual	Asian	Indian	1 Year
P9	28	Cisgender man	Bisexual	White	Serbian/German	1 Year
P10	33	Cisgender woman	Heterosexual	Asian	Iranian	3 Months
P11	32	Cisgender woman	Heterosexual	Asian	Pakistani	2,5 Years
P12	27	Non-Binary	Bi-sexual	Asian	Iranian	6 Months
P13	33	Cisgender man	Heterosexual	White	German	2 Months
P14	40	Cisgender man	Heterosexual	White	Albanian	2 Years
P15	30	Cisgender man	Heterosexual	Mixed race	German/Iranian	3 Months
P16	26	Cisgender woman	Heterosexual	White	German	6 Months
P17	30	Cisgender woman	Heterosexual	Asian	Iranian	2 Years
P18	44	Cisgender man	Heterosexual	White	German	3 Years
P19	23	Non-binary	Queer	Black	Nigerian	5 Years
P20	26	Cisgender man	Heterosexual	Asian	Iranian	2 Years
P21	30	Cisgender woman	Heterosexual	Asian	German/Iranian	3 Months
P22	27	Cisgender man	Gay	Asian	Vietnamese	2 Years
P23	37	Cisgender woman	Heterosexual	White	German	7 Months
P24	23	Cisgender man	Heterosexual	Asian	Iranian	3 Months
P25	38	Cisgender man	Heterosexual	White	German	3 years
P26	33	Cisgender woman	Heterosexual	Asian	Iranian	6 Months
P27	25	Cisgender man	Heterosexual	Asian	Chinese	6 Months
P28	24	Cisgender woman	Heterosexual	Asian	Pakistani	6 Months
P29	34	Cisgender woman	Heterosexual	White	German	10 Months
P30	34	Cisgender man	Heterosexual	White	German	1 year

Part V

Agentic AI Technologies

Introduction

> "*The future is not some place we are going to, but one we are creating. The paths are not to be found, but made, and the activity of making them changes both the maker and the destination.*" – John Schaar

In this part, I explore the emerging opportunities for sensemaking in agentic AI systems, particularly voice assistants, through data access and direct dialogue. As voice assistant users can now request the data collected by the system or ask it directly why it behaved a certain way, I aimed to understand whether these new possibilities address users' needs for empowerment. I also looked at design approaches that treat users as active participants in the design process or even as experimenters of the design.

Both studies presented here were part of a three-year project involving 33 households in a living lab setup in Germany. This setting allowed for a deeper exploration of users' contexts, their situatedness, and their everyday routines and practices.

Chapter 8 explores data as both a design resource and a material, examining how users make sense of their daily interactions with Alexa in the home. The chapter also touches on themes like data literacy and competence-building, as users often seek additional information to better understand their experiences with the system.

Building on this, Chapter 9 investigates how explanations, or "accounts," provided by voice assistants can enhance sensemaking during dialogue-based interactions. The focus here is on improving users' understanding of the assistant's behavior in specific situations.

Together, these chapters show the potential for agentic AI technologies to empower users and foster more participatory and meaningful engagements with technology in their everyday lives.

8

Alexa, We Need to Talk

Alexa, We Need to Talk: A Data Literacy Approach on Voice Assistants

8.1 Abstract

Voice assistants (VA) collect data about users' daily life including interactions with other connected devices, musical preferences, and unintended interactions. While users appreciate the convenience of VAs, their understanding and expectations of data collection by vendors are often vague and incomplete. By making the collected data explorable for consumers, our research-through-design approach seeks to unveil design resources for fostering data literacy and help users in making better informed decisions regarding their use of VAs. In this paper, we present the design of an interactive prototype that visualizes the conversations with VAs on a timeline and provides end users with basic means to engage with data, for instance allowing for filtering and categorization. Based on an evaluation with eleven households, our paper provides insights on how users reflect upon their data trails and presents design guidelines for supporting data literacy of consumers in the context of VAs.

8.2 Introduction

Imagine the keyboard of your computer would log every word or sentence you enter in a file. What would this data reveal about you? Fortunately, a common keyboard does not store this data (at least not until now). Voice assistants (VA), however, are perceived by users only as simple input devices to ask for information and control smart home devices without any greater importance (Pins, Boden, et al., 2020; Voit, Niess, et al., 2020). Unlike keyboards, VAs do log every interaction by default. Combined with the fact that VAs are prone to cause unintended interactions (Schönherr, Golla, et al., 2020) and are an integral part of daily life as they are often being positioned in social places such as kitchens, bed- or living rooms (Sciuto, Saini, et al., 2018), they are potentially very privacy-sensitive technologies. Yet, users are often not fully aware about the kind of data being collected about them and their homes by a VA. Consequently, they do not know how the data are processed and what they reveal about them. This further makes it hard to make informed decisions. Hence, trust in the vendors instead of informed privacy consent has been identified as the central factor affecting the use of VAs (Fruchter and Liccardi, 2018; Worthy, Matthews, and Viller, 2016b).

As data is increasingly collected and processed by almost all kinds of digital services–ranging from surrounding embedded sensors to broadly established cloud services (Stevens and Bossauer, 2017; Sundar, Kim, and Cho, 2019)–consumers tend to feel both vulnerable and unable to trace their data or to protect themselves (Kang, Dabbish, et al., 2015b). This is especially important in the case of VAs, which are placed in users' private spaces and collect data about them (e.g., recordings, preferences, and location),

revealing their daily routines and the (usage) behavior in their homes. Right now, VA vendors only provide abstract information about the collected data in their privacy policies or allow users to see simple lists of interactions with their VA devices hidden somewhere deep in their account settings. By empowering users in their data literacy on voice interaction data, we want to make the operations and processes of VAs more transparent, thereby helping users to make better informed decisions regarding data and privacy protection.

Adopting a research-through-design approach (Davidoff, Zimmerman, and Dey, 2010), the prototype we present aims at making VA data accessible to end users and unveil design elements suited to support sense-making of privacy implications relevant to users. For gaining data of VAs, we made use of the right to access articulated in Art. 15 GDPR, granting users the opportunity to request all personal data collected by a device in a machine-readable form. These takeouts are non-standardized and typically consist of compressed folder structures with a combination of audio, json, html, and other file types. At that time, the prototype was able to process the conversation related data from Google Assistant and Amazon Alexa. These interaction data were visualized on a timeline, accompanied by basic functions for interacting with data, such as categorization and filtering for terms from the commands.

We evaluated our prototype with 11 users of VAs (8 Amazons' Alexa and 3 Google Assistant users) using a combination of interviews and observation as proposed, e.g., by (Tolmie, Crabtree, Rodden, Colley, and Luger, 2016). The interviews focused on users' expectations, awareness, and concerns about data collection by VAs. Following up, we used a research-through-design approach and let users explore their data freely. We observed their interaction

with our prototype based on tasks and questions we have set and by thinking aloud to analyze their sense-making of the data.

Beyond drawing attention to VAs as an important field for research on supporting data literacy, we contribute with insights and guidelines for HCI practitioners to support consumers in their data work and sense-making of sensitive data in general and particularly audio/transcribed data to reduce the existing imbalance of power and knowledge between consumers and vendors:

- Through providing transparency on the variety of collected data and information to participants, we investigate users' data collection awareness regarding their VAs,
- By providing visualizations and means of interaction with the data, we analyze how consumers make sense of data collection and processing as well as how it affects their perceptions, and
- Based on these insights, we propose and demonstrate design approaches for visualizing VA data to support data work to ultimately strengthen consumers in their data literacy.

8.3 Related Work

With VAs becoming common in private households, recent work has shown that users often only have superficial awareness of how a VA works (Luger and Sellen, 2016) and how the collected data is being used (Lau, Zimmerman, and Schaub, 2018). That is why we take up recent work about the adoption and use of VAs, illustrate why it is important to foster user awareness of

the extent of data collection in users' private spaces, and report how recent research explores ways to make the use of VAs more privacy-friendly. We further address how users understand data collection as they often do not know or even do not care about that (Lau, Zimmerman, and Schaub, 2018; Tabassum, Kosinski, and Lipford, 2019). According to current research, the latter is mainly due to the lack of transparency and complexity of terms and conditions as well as privacy policies (Cadogan et al., 2004; Jakobi, Stevens, et al., 2020). This may lead to feeling overwhelmed or disappointed and cause users' distrust. As presented in 8.3.3, we contribute to the ongoing research in HCI by providing insights about the sense-making of data.

8.3.1 Adopting VAs in Daily Life and Private Spaces

As previous work depicts, VAs are often perceived as simple input devices. For example, Porcheron et al. investigated voice interfaces in everyday life and claimed that interactions with VAs are not of conversational but rather of rational nature (Porcheron, Fischer, et al., 2018). Nevertheless, when adopting a VA, users often face high expectations regarding their functionalities (Cranor, 2008; Pins, Boden, et al., 2020; Voit, Niess, et al., 2020). In some cases, VAs were even anthropomorphized (Lopatovska and Williams, 2018; Purington, Taft, et al., 2017).

As recent work has shown (Bentley, Luvogt, et al., 2018; M. Cho, Lee, and Lee, 2019; Porcheron, Fischer, et al., 2018; Sciuto, Saini, et al., 2018; Voit, Niess, et al., 2020), users need to learn how to integrate the VA into their daily and social life at home. By studying long-term use of VAs, (Bentley, Luvogt, et al., 2018) found that after a longer period of time, routines of functions

and usage times can be identified. (Sciuto, Saini, et al., 2018) supported this insight and concluded that intensity of use is strongly dependent on the daily routines of the users as well as the devices' locations. This leads to the assumption that VA data allows for comprehensively picturing users and their activities. Moreover, VAs suffer from unintended activations as Schonherr et al. found (Schönherr, Golla, et al., 2020). However, VAs do not treat these interactions differently and collect every word followed by the misinterpreted activation word. Combined with the fact that users tend to forget that a voice assistant is present (M. Cho, Lee, and Lee, 2019), the awareness about the actual extent of data collection may be reduced even further.

The use of these so-called 'always-listening' (Lau, Zimmerman, and Schaub, 2018; Tabassum, Kosinski, and Lipford, 2019) devices and the potential for using the collected data has not only been a sensitive topic in general society for some time now (Søndergaard and Hansen, 2018; Tabassum, Kosinski, and Lipford, 2019). It has also been of interest among the HCI community regarding privacy preserving technologies (Abdi, Ramokapane, and Such, 2019; Chung, Iorga, et al., 2017; Diao, Liu, et al., 2014; Lau, Zimmerman, and Schaub, 2018). In addition, the collection and use of VA data is also of importance from a legal perspective as it may be of evidential value for law enforcement (Orr and Sanchez, 2018; Pfeifle, 2018).

In the HCI community, this aspect motivated the exploration of possible measures to avoid data collection, e.g., by using wearable microphone jammers (Chen, Li, et al., 2020), implementing solutions for data protection like a detector for security-critical commands (Kwak, Huh, et al., 2019), and

improving activation-word detection by taking the gaze direction and voice volume level into account (Mhaidli, Venkatesh, et al., 2020).

8.3.2 Consumer Understanding of Data Collection and Processing

In their study on data collection practices of VAs, (Malkin, Deatrick, et al., 2019) found that about half of their participants (51,7% of 116) were unaware of or unsure about companies' keeping their recordings for an indefinite time. Additionally, users only vaguely understand the extent of data collection and their processing by the system or vendors (Abdi, Ramokapane, and Such, 2019; Tabassum, Kosinski, and Lipford, 2019). This opaqueness is a known phenomenon which also exists for other services and is attributed to the fact that terms of use and privacy policies, which include information about just these practices, are commonly not read or understood (Kang, Dabbish, et al., 2015a), fostering misconceptions on the customer side (Reidenberg, Breaux, et al., 2015; Tabassum, Kosinski, and Lipford, 2019). In this vein, many customers were surprised when it came to public attention that subcontractors of VA vendors listened to actual user-input audios to train VAs (Carr, Day, et al., 2019).

Previous studies have shown how existing measures for transparency in VAs seem to fail consumer needs (Alizadeh, Jakobi, et al., 2019; Jakobi, Stevens, et al., 2020; Merrill, Chuang, and Cheshire, 2019). For example, Jakobi et al. revealed that consumers did not perceive the purpose limitations provided by VA data protection policies to be informative with regard to the extent and consequences of data processing (Jakobi, Stevens, et al., 2020).

Vendors already provide a dashboard that offers tracing data and the system status of ongoing interactions with a VA. Another way to interact with data was made possible with the adoption of the GDPR that gives consumers the right to access, manage, and transfer their data . Recent work on feeding back data to consumers based on the right to access shows how mismatches between data collection and users' expectations can have a variety of consequences: For example, Alizadeh et al. reported that consumers mistrust loyalty card companies, assuming that they were hiding something from their takeouts (Alizadeh, Jakobi, et al., 2019).

A similar conclusion was also reached by (Merrill, Chuang, and Cheshire, 2019) who investigated the perceived capabilities of biosensors (e.g., face expressions or GPS and accelerometer) and pointed out that on the one hand, devices may seem creepier than they are, whereas on the other hand, in fact, seemingly harmless devices may be used to infer intimate details such as emotions or mental health. They further concluded that users' perceived capabilities about technology often rely on the physical shape or interface of the device, which will become an important issue when designing and implementing tools gadgets with ubiquitous sensors.

VAs and other smart home devices are also described as black boxes to their users – especially if they are not tech-savvy. While these inter-connected devices blend into consumers' daily life, lack of control and accountability of their inner workings lead to various concerns, for instance about remote access control or reliance on services or vendors (Brush, Lee, et al., 2011). With the home being the epitome of privacy (Habermas, 2015; Warren and Brandeis, 1890), consumers should be able to see and control what actually happens there (Castelli, Ogonowski, et al., 2017; Tolmie, Crabtree,

Rodden, Colley, and Luger, 2016). For example, throughout the phases of technology appropriation (Silverstone and Haddon, 1996) smart home technology acceptance hinges on appropriate ways of providing insights in the system status (Jakobi, Stevens, et al., 2018).

Jakobi et al. already identified design resources for fostering controllability and accountability of do-it-yourself smart home systems that increase data literacy, eventually helping users to learn about system behavior and processes such as data disclosure (Jakobi, Stevens, et al., 2018). Arguably, similar principles may be applicable to the use of VAs in the home but have not been getting as much attention. That is why we want to pick up this thread on accountability of connected devices and focus on the conversation data of VAs in this paper. By doing so, we increase both visibility and accessibility of the collected and stored data and draw attention to executed actions and hidden processes (especially unaware activations) (Bellotti and Edwards, 2001; Schönherr, Golla, et al., 2020).

8.3.3 Designing for Sense-Making of Data

The lack of transparency of data collected by sensors at home and fostering consumers' data literacy is a common topic in research (Worthy, Matthews, and Viller, 2016a). For example, users of smart home systems need solutions to maintain their systems to understand how it works and behaves, especially to get informed about unexpected events (Jakobi, Stevens, et al., 2018). However, users need support in exploring and making sense of their data collection and use (Tolmie, Crabtree, Rodden, Colley, and Luger, 2016) beyond receiving notifications.

Whereas companies are able to employ data scientists, households lack both skills and tools (e.g., for natural language processing) to identify and judge sensitivity of data (Castelli, Ogonowski, et al., 2017; J. E. Fischer, Crabtree, et al., 2016; Tolmie, Crabtree, Rodden, Colley, and Luger, 2016). Hence, designing tools that support users in their data work and its challenges are still a matter of ongoing research (Chhetri, 2019; S. Garfinkel and Lipford, 2014). This clearly illustrates the imbalance of knowledge and power over user data between users and companies.

One noticeable example is the study of (Castelli, Ogonowski, et al., 2017) that shows how flexibility and compatibility to routines and practices are crucial for a visualization of connected device data for smart home systems. Another important study was conducted by (Kurze, Bischof, et al., 2020), who developed a method called 'Guess the data'. In their work they observed the individual but also collective sense-making of live data of people's homes. Other approaches address privacy precautions specific to the smart home like storing and processing the collected data in the devices locally (Yao, Basdeo, et al., 2019) or the minimization of data collection. Furthermore, there are already tools existing that notify consumers when a device shares data (Seymour, 2019) or that summarize the shared information (Balebako, Jung, et al., 2013).

Initial studies on data work with dashboards developed by VA vendors themselves show – while raising awareness in general – how provided information in the form of chronological lists was not well suited for users to explore data and build an understanding of their VA data footprint (Pins, Boden, et al., 2020). Moreover, VA data may introduce a different challenge: Since VAs can operate as a connective platform of smart home devices, apps, and services

that bundles a variety of sources and information, interaction data of such devices may hold a rich database about activities at home. (Vitale, Chen, et al., 2020), for example, already investigated a centralization of different information on a data dashboard to support users to curate their data from across platforms and devices due to the mix of devices and cloud platforms used by customers.

The field of usable privacy (A. Adams and Sasse, 1999; Brodie, Karat, et al., 2005; S. Garfinkel and Lipford, 2014) is, from an HCI perspective, about enabling consumers to be more aware of their personal data disclosure and providing usable controls for privacy management (A. Adams and Sasse, 1999; Cranor, 2008; Whitten and Tygar, 1999). This includes both usable tools for accessing and managing personal data as well as support in understanding how and for what purposes the data are processed (S. Garfinkel and Lipford, 2014; Van Kleek, Liccardi, et al., 2017). Research often focuses on sensor data like energy consumption, system status, stability of smart home systems or the presence of people in a room as there already are tools supporting sense making of smart home data (Castelli, Ogonowski, et al., 2017; J. E. Fischer, Crabtree, et al., 2016; Kurze, Bischof, et al., 2020).

With regard to VAs, there is already some research on designing natural language processing tools to help consumers find hidden patterns and meanings of their (potentially sensitive) data at home (Seymour, 2019). However, to best of our knowledge, there are no tools that support sense-making of voice interaction data specifically.

We argue that data of voice assistants is a non-trivial task and sensitive issue: In fact, there exist many design resources for smart home data analysis that can be reapplied for designing sense-making tools for VAs. For example,

as it is the case for many data collecting devices, users only have a vague understanding of the type of collected data as well as timing, storage format, and intended use by the service provider (Abdi, Ramokapane, and Such, 2019).

However, VAs need to be treated different due to the natural language interaction via voice, the accidental activations, and the thereby collected data that arguably make VAs different. Combined with the known problems of opaqueness and lack of usable control of embedded devices, VAs, at least in their current form, are suited to further weaken consumers' informational autonomy over insights into activities at home and the data collected about these activities and routines. Therefore, much like with other smart home devices, a promising approach in this respect is researching usable tools for supporting consumers in analyzing their voice interaction data to foster awareness and sensitivity for privacy implications of conversation data with VAs.

8.4 Data Visualization Prototype

In previous work, we investigated the use and appropriation of voice assistants in private homes (Pins, Boden, et al., 2020). In this context, we presented users their interaction histories on the dashboards provided by the vendors. This led to some interesting insights for the users but turned out to be too limited for deeper reflections in the sense of fostering data literacy or awareness. Based on this work, we developed a prototype (see Figure 8.1) which is able to visualize, browse, cluster, and filter conversation data. It

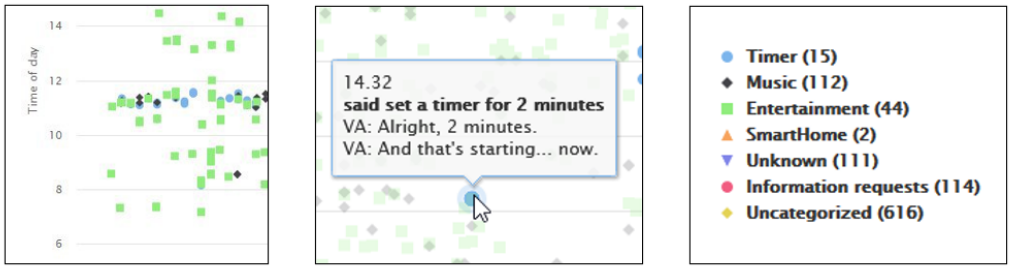


Fig. 8.1.: Screenshots of our prototype for supporting consumers in understanding their VA data. Left: Graph with interaction events, sorted by day and time of day. Middle: Information of an individual event. Right: User-defined categorizations to structure data.

accepts data takeouts as input as provided by Google and Amazon via the right to access as postulated in GDPR.

For the implementation, we compared both datasets – especially the interaction data transcriptions – to identify visualization possibilities and decided to focus on time stamp and transcribed commands in our initial prototype since both variables were consistently present. Takeouts from Google further contained the answers of the VA which were missing from the Amazon takeout. However, by now, Amazon also provides such information for some commands. Based on those conditions, we decided to show the interactions as points on a graph, sorted by time with the possibility of categorization by terms extracted from the transcribed commands to identify patterns. Other studies support this decision as they have already shown that time series data and the visualization of patterns and actions are usable cues for supporting people in reasoning about data. Further, many automated actions that were programmed were mostly triggered by the time of day (Davidoff, Zimmerman, and Dey, 2010; Jakobi, Stevens, et al., 2018; Mennicken, Hofer, et al., 2014).

The tool visualizes every interaction on a timeline, where the x-axis shows the date and the y-axis shows the time of day of the interaction. Hence, each column in the graph represents a day from 0:00 at the bottom line to 23:59 on the top line. The user can zoom into the graph and move the view to look at clusters of interactions in detail (see Figure 8.1, left). If a single interaction marker is selected, a tooltip shows the user's command as recorded by the device in combination with the VA's response (see Figure 8.1, middle).

To further support sense-making, the tool allows users to sort their data into categories based on terms from the users' commands, for instance 'timer' would include all user commands that contain this expression. These categories are then coded with individual colors and symbols (see Figure 8.1, right). Users can toggle visibility of each category to focus on single events or categories. For supporting the classification of interactions, the tool implements a simple natural language processing algorithm which tokenizes the user commands and presents the terms as a list ordered by frequency. By doing this, users can see which terms appeared frequently in their voice interactions and might be interesting to use as selectors for categories, such as 'music', 'play', 'start', etc., as illustrated in Figure 8.2.

We supplemented a video as a visual demonstration of our working prototype to this paper.

8.5 Research Methodology

Between June and August 2020, we conducted a remote study with 11 participants to observe how our prototype affects consumers in their sense-making process. The study took place remotely via a video-conferencing tool

because of the COVID-19-situation and was led by two interviewers: one moderator and one note taker.

Prior to the data exploration session, we supported the participants in requesting their data and interviewed them about their expectations regarding the provided data takeout. In case of Alexa, the data takeout can take up to several weeks until it is made available. We therefore continued with the second part of the study after the data takeouts could be accessed. Although this was an important and interesting step towards the conduct of the exploration, this paper focuses on the data exploration in order to keep the scope and focus narrow.

During the data exploration, we asked participants to share their screen to be able to trace their activities and provide support if needed. The shared screen was recorded with their consent. Both interviewers wrote down their observations about critical incidents, interests, and needs expressed by the participants regarding data visualization and usability. We asked them to verbally reflect on their data and freely express any questions. Each exploration took about 60 to 90 minutes. Our analysis was based on an open-coding approach in the sense of Grounded theory (Strauss, Corbin, et al., 1996) which helped us to analyze users' behavior and needs regarding data work and sense-making.

The data exploration session started with a 10-to-15-minute exploration of the data takeout where we discussed the clarity of the data and files they received. We also asked about the relation of the information regarding the users' expectation recorded in the first round of interviews. After that, we gave them access to the prototype to explore the data. The procedure was structured openly so that our participants could freely explore the prototype

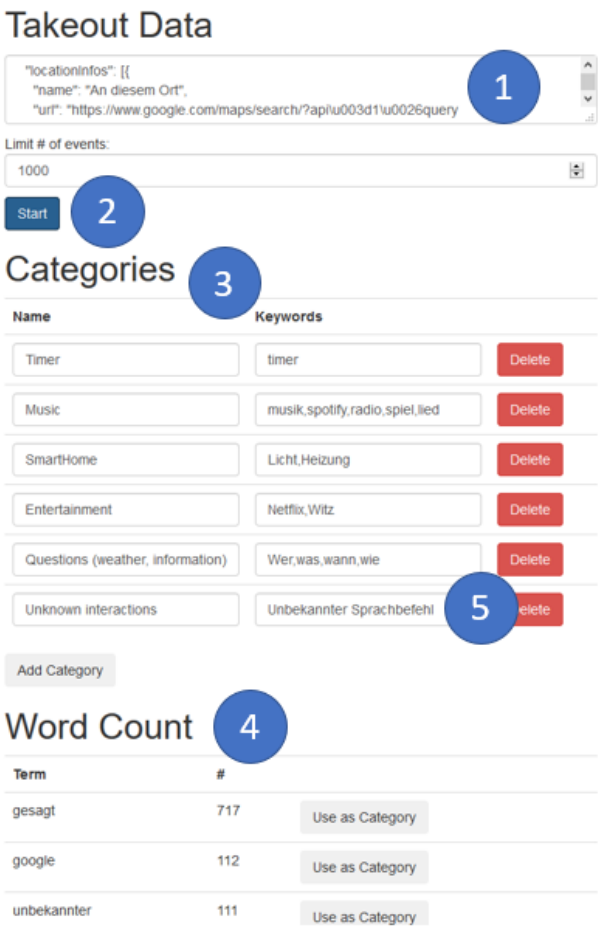


Fig. 8.2.: Settings and functionalities of the prototype for visualizing the data.

and their data. Nevertheless, we prepared a guideline with four tasks to guide the users through the exploration. The observation started with task 1 by copy-pasting the conversation data into a respective form field on the prototype (see Figure 8.2, number 1). After clicking on 'Start' (number 2), a first visualization appeared with 'Music' and 'Timer' as two pre-defined categories. We gave the participants some time to freely explore the prototype and data visualization while thinking aloud. The second task was to create own categories (number 3) by considering the graph, the containing interactions, and the word list (number 4). Finally, the last task was to create another category for all data that relate to unassignable interactions (number

5) which are declared in the data takeout as 'data not available' for Alexa users and 'unknown voice command' for Google Assistant. This was planned as a separate task as we assumed that the participants were not aware of the existence of that data. Furthermore, we wanted to investigate if they understood the meaning of this tag and thereby wanted to ensure that they became aware of these data with potentially sensitive content.

8.5.1 Sample

The 11 participants of our study (see Table 8.1) were 9 male and 2 female German native speakers with a mean age of 37. All of them were VA users. They either lived in single households or shared their homes with their family, partners, or flatmates. Participants were recruited via social environments of the authors or via snowball sampling. The most important criterion here was a general interest in the data that is or might be collected by a VA, regardless of their level of concern regarding privacy.

Furthermore, qualified participants needed to have access to the collected data. Therefore, an additional precondition for participation was being the account holder and qualified of requesting a takeout via the right to access (Søndergaard and Hansen, 2018). We informed our participants about a post-study 15 € voucher only after they had agreed to participate in our study. This approach ensured that our interviewees were taking part in our research primarily because they were interested in the data and not motivated by financial incentive. The account holders who we talked to turned out to be mostly the male household member. This is, as we also reflected in the

#	Age	Sex	Household	VA System
P1	35	male	3 people, Family	Amazon
P2	27	male	3 people, Partnership / Shared flat	Google
P3	48	male	4 people, Family	Amazon
P4	26	female	1 person, Single	Amazon
P5	35	male	4 people, Family	Amazon
P6	24	male	1 person, Single	Amazon
P7	34	male	3 people, Family	Amazon
P8	56	male	3 people, Shared flat	Google
P9	34	male	2 people, Partnership	Amazon
P10	58	female	2 people, Partnership	Amazon
P11	31	male	5 people, Shared flat	Google

Tab. 8.1.: Participant information of our study

analysis, in line with the general user group composition of smart home devices (Shin, Park, and Lee, 2018).

The participants had used either Amazon Alexa or Google Assistant for one up to four years. Our sample includes 8 users of Alexa and 3 of Google Assistant, which reflects the fact that Alexa is more popular in Germany. However, we also included Google users to contrast different data trails. The main use cases turned out to be rather similar in all households and included mainly smart home control, especially lights and TV, listening to music or radio, and asking for information like news, weather, or knowledge inquiries. Some used it for playing interactive games, especially quizzes, others for entertainment like watching series or telling jokes. Two households connected their VAs with their calendar to regularly ask about upcoming events. Another household used the VA to maintain a shared shopping list.

8.6 Results

In this section, we present the results of the data exploration. We first give a short overview about the process of receiving the data and its exploration to explain users' expectations regarding the data. After that, we continue with the visualization and sense-making with the help of the prototype.

8.6.1 Expectations on the Takeout Format

After participants had received their data takeout, we started with a 10-to-15-minute exploration of the raw data. In contrast to Amazon Alexa users, Google Assistant users had the option to choose their data takeout format, for instance, json or html, and could therefore already develop an idea of how the data will be presented. However, our interviewees who used Alexa expected to receive a more accessible and well-formatted takeout since they obtained no information about the potential content, format, or structure prior to the download. As high-level expectations, some thought of receiving a link to a website where the data would be prepared and fed back: These data were imagined to be statements about recommendations of songs they listened to, or information about who had access to this data. However, the expectations regarding the scope and depth of detail were different: With respect to the data format, five participants mentioned they expect a single extensive pdf document or excel sheets with tables. Overall, participants were afraid that the data might be aggregated or otherwise incomprehensible:

> *"I think I get them in the simplest possible form. I would be surprised if they were complete or even meaningful. I am really curious what I get."* (P5)

There was consensus that, similar to the interaction history provided by vendors through their proprietary dashboards, the file would include information about the timestamp, the command, and the answer. Moreover, most of the participants thought of an option to listen to the record as it can be done in the vendors delivered interaction history. Two participants also imagined that information about the devices used would also be provided, for example how long the TV was on or whether and what data exchange had taken place with the devices.

8.6.2 Exploring the Data via the Takeout

The dispersed structure of the directory in the data takeout made it harder to gain an overview and understanding of the provided data as the labelling of folders and files were not self-descriptive. Html files, if available, were easier to understand than csv or json files. As expected, three of our participants who reported previous experiences with programming and / or worked in the IT sector had less problems to go through the takeout or open and read the files. All other participants, however, described the content as very technical, difficult to read, and hard or impossible to understand – especially because the labels and terms were provided in English.

Next to the format of the presentation, some participants were also surprised by the contents of the data takeout; for example, becoming aware of the mere amount of data in the takeout from Google (and all its services):

> *"[There are] 9,427 files. [. . .] That means 30 files per day over a year, that's quite a lot."* (P8)

Another participant was especially surprised about the audio files of the commands that were ever spoken to the VA which were a part of the takeout:

> *"I'm just unzipping the directory and that's a lot [...] There are alone 3,170 mp3 files [of 6,979 in total] in the folder of Google Assistant, that's totally crazy. [. . .] And this means, there have to be a lot of other interesting files in the directory."* (P11)

After the open exploration of the directory, we asked our participants to identify the interaction related data, especially the recordings and transcriptions to use the latter for further processing with our prototype. Again, the participants had problems to identify the correct files or were unsure about where to look for them. The files of the Google Assistant, for example, were located within the folder "My activities" whereas the participants expected to find them on a directory level similar to other Google services like Gmail or Maps. Alexa users were unsure if they had found the correct files since they were confused by the missing structure of the provided csv file – it was presented as a single string containing all information.

The raw data format provided by the vendors underlined our assumed need for implementing a tool that visualizes and supports in sense-making of the data.

8.6.3 Exploring the Data via the Prototype

Within the prototype, the participants were able to visualize their interaction data in a graph with the date noted on the x-axis and the time of the day shown on the y-axis. This already helped participants to get a birds'-eyes overview of their spoken commands since the beginning of use:

> *"The graph and the categories are pretty cool. I thought of pie charts about general use. But this allows me to see what I said and when [. . .] As you can see here 'turn that [light] off' and as here is shown, you can see when I went to bed at these times."* (P2)

The participants quickly understood the graph and identified their typical usage times by noticing accumulations of events in the morning and the later afternoon or evening with empty areas in between.

We further implemented a word list that shows regularly used terms based on users' commands ordered by frequency of use (see Figure 3). The word list together with looking at the graph was always considered to be helpful for building new categories. They also followed their mentioned use cases such as 'Control' (of lights or volume), 'Weather', 'TV', 'News' or 'Shopping' (for shopping list). Their reactions to the visualizations were always positive:

> *"I have now actively searched for 'shopping' and the tool now shows me when I added tissues and toilet paper to the list. I like it, I really like it."* (P5)

Four of our participants expressed the wish to have a list of complete commands for building categories but were also concerned that commands for the same task can vary regarding wording or syntax. None of our participants felt the need to apply the category for unknown voice commands by themselves since their categorization reflected their actual use. Therefore, we created a specific task consisting of assigning all interactions caused by unintended processing of the VA to the respective unknown voice commands category because we also wanted to discuss this in our interviews. Some Alexa users

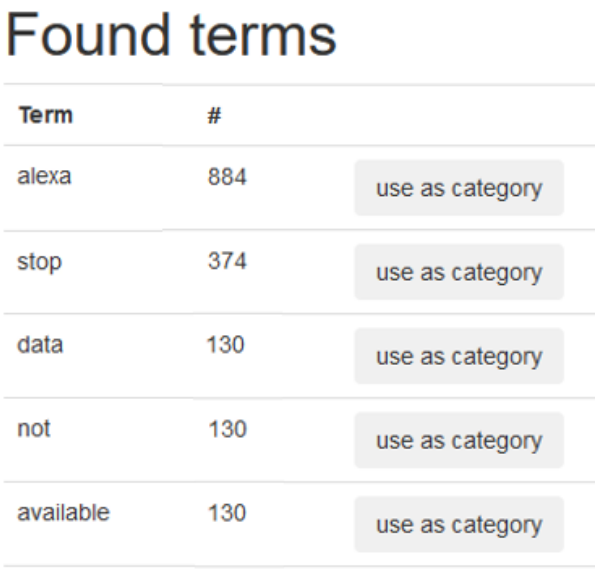


Fig. 8.3.: List with identified terms out from the interactions with the terms 'data', 'not' and 'available' that refer to unknown voice commands (own example)

were already confused about the English terms in the word list (see Figure 8.3) even before we directed their attention in our dedicated task:

> *"I don't know now what this means. I don't think we have spoken in English to her, but maybe Alexa said that when something happened. But I can't explain it to myself."* (P1)

The participants quickly came to conclusion that these commands might have occurred due to technical problems like internet or connection issues or that these might have been incomplete or unintended interactions:

> *"Either they are the ones where my flat mates were saying something or those are commands where it [the VA] was unable to assign it to a command."* (P8)

Nevertheless, these interactions attracted their interest and made some participants reconsider their views about the collected data. For example,

one participant asked about the necessity of storing especially the data on unknown commands:

> *"I wonder why the data is stored for such a long time, especially those unknown voice commands."* (P3)

When Alexa users listened to a few audio files connected to these unknown commands, they sometimes only recognized the activation word or TV noises from the background. On the one hand, this calmed them down. On the other hand, however, it also made them realize that there could be more sensitive information recorded. As a direct reaction to this encounter, four participants stated they might be more careful about what they say with a VA present in the future to avoid its unintended activation:

> *"When I give a command in the future, I would try to make sure that I am in a neutral environment or that I am paying attention to what is being said in the background."* (P6)

Google Assistant did not attach any recordings to their unknown voice commands, and users were therefore wondering why the information was appearing in the data at all without further clarification.

The visualization of the data was appreciated by the participants and allowed them to reflect on their current use of the VA. It also supported them to better understand what data was collected about them but they rarely saw the need to use a tool like our prototype on a regular basis:

> *"I don't know why I would use it again in a couple of weeks, after I had a look at the tool today, because then I won't see any difference. [...] At the moment the tool is still rather good as something 'Nice-To-See'."* (P11)

This might be because the prototype did not allow further interpretation and analysis of the data. Therefore, we asked participants about their interests and needs in the next step as reported in the following section.

8.6.4 Supporting Sense-Making

During the thinking aloud part of the exploration, we got meaningful insights into how the visualizations inspired people to reflect on the use of their VA. Furthermore, the visualization motivated them to ask further questions, for example about what other things Google or Amazon might know about them. We identified four different areas of envisioned use of a VA data awareness tool:

For maintaining an overview. The prototype helped participants to get aware of their usage behavior and times by categorizing the commands. The different colors and forms for each category further helped to trace the interactions. By hovering an interaction point in the graph, the participants could also see every single interaction with its date and the conversation in a little tooltip next to the event (see Figure 8.1, in the middle). This visualization was evaluated as helpful although Alexa users were missing the answers of the VA since this feature was not yet available during the implementation of the prototype.

The participants expressed further needs to get information about the person voicing the command, the used device, and its location. Besides VA interaction data, participants were especially interested in data disclosure to involved devices (e.g., smart home devices like lights or TV) or third parties (e.g., when using skills / actions like Spotify):

> *"It would still be interesting to see who the [addressed] app provider is or who is behind it. But more I cannot do with it yet. I think it would be really exciting to see if there is another interaction list between the devices."* (P6)

P6 envisioned the mentioned list to contain information about the data that were transferred between the devices and also show if the information was only a signal or the whole command. It would also be helpful to see if an interaction was executed successfully.

For looking back at past activities. The graph and the timeline further enabled participants to look back to see all interactions, even those that happened when the user was not at home. As a result, the data takeout revealed some information about what other household members were doing during the absence of the account holder:

> *"As a user [account holder], I can see data from periods of time even when I was not at home. In terms of privacy, this is already precarious and like a small eavesdropping attack. If I were to suspect that my wife was receiving male visitors, then I could practically start looking for them now. You know: 'At this and that time I wasn't at home, and there is an audio file where I can hear a male voice in the background!' or something like that."* (P3)

While perceived critically from a privacy perspective and the existing advantage of knowledge about other household members, it was also perceived helpful for specific issues like rediscovering information that had been looked up with the VAs previously:

> *"For example, if I want to find a song I heard in the past. The category 'music' and the date could help me find the song again."* (P10)

For learning about interactions. As mentioned in the results, the word list was helpful to identify frequently used terms, and it supported participants in building categories. Some participants asked for a list of all used commands, which then could be used for categorization or clarification about potential differences in interpreting the voice command and the intended interaction from a technical point of view:

> *"It currently shows me the identified expressions as single words. It would be interesting if you could say as an alternative option 'show me this as it was understood', as a whole command."* (P9)

The participants were interested in getting to know about differences or changes like if the way of talking is becoming more human-like or not, for example, by analyzing the duration of the commands as another participant concluded:

> *"The length of the commands is also saved in the data. [...] a hypothesis might be that people talk more complexly with it over time, what I actually do not believe. But then you could show how the number of seconds per command has changed over time."* (P11)

P11 also proposed a better presorting of the terms or commands by topics for a clearer structure of the list and the data:

> *"It would be good if data were as much as possible pre-categorized. These single terms are actually not that relevant. Rather, they are*

more relevant [as categories like] smart home control, alarm clock, everyday organization, or multimedia [...].“ (P11)

For re-balancing power and knowledge. Our participants were interested in the visualization of the data, but also expressed a need to know what the data tells about the users and how the vendors might analyze their behavior – in other words, how others use the data and what they learn about the users:

“First of all, I would like to know what they are doing with the information they have gathered. That would be interesting, and it would also be interesting to know what they do with terms like ‘television’, if they use it in any way.” (P9)

Here, a need can be observed to understand which patterns can be identified and even how single terms or commands affect their analysis. For example, participants expressed the need to get further analyses about the music they listened to, based on recommendations of their music app:

“The VA also knows for every song which similar songs exist or what I would like to hear next.” (P7).

Besides the analysis of the content and the way of talking to the VA, two participants also assumed that the audio file could be used to analyze the speaker or the situation s/he was in:

“The audio file probably tells if I am alone, if there are other voices in the background, if there is a party going on or an exuberant mood. [...] The vendor can already generate a good profile about me.” (P6)

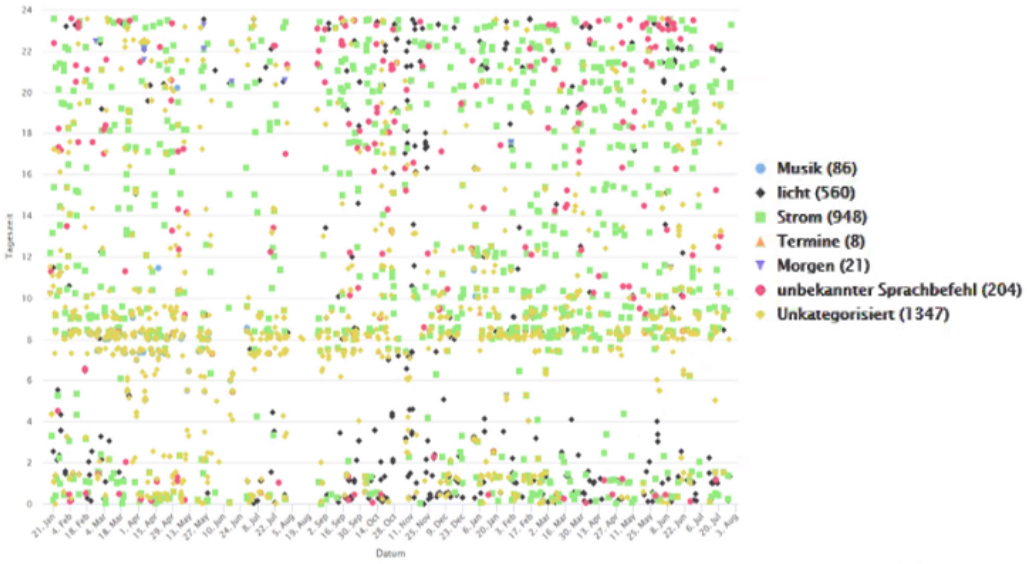


Fig. 8.4.: A visualization of VA data from P11 with 3,172 interaction points in total since his first use in January 2019. The categories on the right with number of events in brackets are as follow: Music (86), Light (560), Power (948), Events (8), Morning (implemented routine of actions) (21), Unknown Voice Commands (204), Uncategorized (1,347)

Only one participant noted that it is less a question of the type and manner of the analysis than of the conclusions that can be drawn from them:

> *"I can imagine they store other, more profound data too. For example, analytics about the voice [...] as well as other basic things such as the algorithms that I as a layperson might not be able to use [...]."* (P4)

As a concluding example for a users' visualization, Figure 8.4 depicts a graphical representation of the data from P11 who used his Google Assistant for about two years in his room in a shared flat. Therefore, he was the only user besides his girlfriend who would sometimes use it. P11 mainly controlled smart home devices, specifically lights and power consumption of other devices (category 'power').

8.7 Discussion

The results show that, much like for other embedded devices, also for VAs, user expectations about data collection often differ from the data which is actually being collected (Merrill, Chuang, and Cheshire, 2019).

In the following, we therefore show the participants' learning process and discuss how a tool like the prototype used in our study supports users in increasing their data literacy if it is enhanced further.

8.7.1 Awareness Creates Data Literacy

Throughout the study, our participants underwent a learning process as we supported them in reflecting about their personal VA data. This information was previously described as difficult to grasp. Their expectations regarding what data might be collected were based on the vendors' interaction log and their mental models about common usage situations. Our tool partially acknowledged these views as the general categories fit the use cases that had been expected, but also made our participants aware of both the amount of use and especially the unintended interactions.

Bringing attention to recordings. However, the users were surprised to find information such as full audio recordings including, e.g., background noises, which allow for drawing conclusions about the situation in their home. In this regard, some participants highlighted additional privacy implications for themselves as well as for bystanders (Silverstone and Haddon, 1996). Besides the awareness of the interactions with the VA – intended or unintended – we asked our participants about concerns regarding privacy when they had

a look at their data. As other studies have demonstrated, our users felt generally aware of the privacy implications of voice recordings (Chhetri, 2019; Frohlich and Kraut, 2002; Schönherr, Golla, et al., 2020), especially of those that contain sensitive content (Chhetri, 2019; Diao, Liu, et al., 2014).

Revealing unintended interactions. Another issue evolved around the existence of unknown voice commands and their meaning and privacy implications: Being able to look through and listen to these recordings turned out to pose a trigger for participants to wonder about the sensitivity of VA recordings from a new perspective. It also confirms our initial assumption that VA data is not as trivial as suggested in other research (Kurze, Bischof, et al., 2020). We therefore argue that these data sources are of high relevance for further development of awareness tools for VAs.

Addressing user needs for sense-making. As our previous work had shown (Pins, Boden, et al., 2020), the list of commands in the vendors' dashboards offers few practical functionalities to navigate through interactions. Furthermore, despite our sample consisting partly of tech-savvy users, the raw data format posed a serious challenge in terms of intelligibility. There is a need for more tools like the prototype presented in this work to help users making sense of their data and raise awareness about what IoT devices detect, save, and process. Here, our study pinpoints the fact that consumers seem to be less interested in how the data is processed (Tabassum, Kosinski, and Lipford, 2019; Zeng, Mare, and Roesner, 2017). However, they do care about what is revealed about them or what additional data layers are created through the processing (Frohlich and Kraut, 2002). Usable tools for providing awareness and transparency in terms of privacy implications will have to find a balance

between showing detailed (e.g., raw) data or technical detail and high-level visualizations.

An interesting question in this regard is to what extent such tools interfere with or match the algorithms that companies use to interpret data. On the one hand, our findings indicate that business secrets are not affected because they are anyway exempt from the regulation and do not have to be revealed (the Council, 2018). On the other hand, it is unclear how we can provide adequate analyses of companies' data for users with the limited possibilities of working with individual takeouts in contrast to the possibilities of vendors who have access to data of all users. One approach could be to focus more on shared analysis such as building a database of keywords and categories that help users analyze their data collectively other than performing individual analysis on single data sources.

8.7.2 Data Literacy as a Skill that Concerns Everyone

Based on the Privacy Personas by (Dupree, Devries, et al., 2016) we could identify mostly personas ranging from 'lazy experts' and 'Amateurs' in our study as they trust or rely on the security of the services they use and often mentioned they have nothing to hide or there are other devices or services that already collect their data (Zeng, Mare, and Roesner, 2017). One participant in our study (P10) was more concerned about the always-listening device and therefore stored it in the drawer when it was not in use. This indicates that P10 showed a higher motivation to protect her privacy and security than our remaining participants by choosing security

over convenience. A closer look into the interview data in follow-up research will give more insights about the attitudes of our sample.

Although awareness about data safety and security issues including the potential of profiling people is gaining popularity, affected users often do not have the means to track and understand data collected about them via smart home devices such as VAs. Additionally, data takeouts which are requestable from the devices' vendor require a certain technical understanding for users to make sense of them. It is therefore necessary to provide helpful and easy tools to regularly check data transfers for users at home (Castelli, Ogonowski, et al., 2017; Davidoff, Zimmerman, and Dey, 2010; Jakobi, Stevens, et al., 2018; Mennicken, Hofer, et al., 2014) –just like it should be done with ones' financial tracks. That is why, based on the Design-for-All principle Persson, Åhman, et al., 2015, we intended to create a tool which enables interested VA users to investigate what the voice assistant is picking up at home.

As our results show and what can be confirmed by other related work, the needs and interests regarding data can be different among users and may change over time (Jakobi, Stevens, et al., 2018; Vitale, Chen, et al., 2020). For instance, some participants were only interested in unexpected events whereas others directly started to ask further questions about how the speech behavior might change over time and how this adaption could be represented. This further implies that not the full set of available information will be necessary for every use of the prototype and for each user. For example, users may just need a notification telling them that everything works as expected (Jakobi, Stevens, et al., 2018) but sometimes they want to investigate further or explore more detailed questions about their data. This change in needs and type of data investigation should be considered for the implementation

of future functionalities, for example by adding exploration tools for more advanced users or by offering personalization options.

8.7.3 Guidelines for VA Data Literacy

The prototype turned out to be a useful first step towards helping users to better understand their data from conversations with VAs as well as the information within the data collected. Therefore, it can support strengthening data literacy. The prototype enabled our participants to interact with the otherwise rather unintelligible raw data more intuitively and raised awareness about unintended recordings. By conducting the observation of data exploration, we were able to show the tool's utility and gather further needs and ideas from our participants to improve the prototype:

Support right from the start. The dispersed structure of the takeout made it hard to understand the data and complicated the navigation within the directory. Therefore, supporting users in exploring and understanding their data should start with explaining the content of the various files and folders and offer follow-up actions a user can take, such as importing the required file from the original data takeout file structure into the visualization tool. On the one hand, this would be more comfortable for the users; on the other hand, it means that the tool should be precise about how the imported data is further processed to inform users about privacy implications. Finally, privacy settings and preferences are often only prompted when a user encounters an unknown product or service for the first time and are likely to be forgotten and never shown or informed about during long-term use. Awareness about

the existence of collected data including relevant policies should be evoked from time to time to prohibit unwanted surprises.

Share categories and support to structure the data. The prototype offered categories and a word list to structure the data. During the user test, we detected categories which were relevant to at least the majority of our participants. Our users also directly asked for more pre-defined categories to speed up the process of structuring the data. Requested categories were, for example 'Smart Home control', 'Lights', 'Unknown commands', 'TV', 'Shopping' or 'Weather'. Implementing more robust automated detection and improved algorithms for suggesting categories may be one viable option. Additionally, allowing users to exchange categories with each other might be a privacy-friendly way to cooperatively support them with sense-making. In general, improving the way categories are built with smarter support functions like usage times, voice recognition or categories based on tasks and their related commands would help users considerably in making sense of their data.

Reconstruct the context. The visualization of voice interaction data in a graph supports consumers to have an overview and to understand their data in contrast to a chronological list as offered in the vendors' apps. Besides the information about the time, the content, and the answer of the VA for each voice command, participants were interested in additional information such as an identification of the person voicing the command. This would help to reconstruct situations by showing who used which of the devices. Furthermore, it would enable users to identify unknown voices or unusual locations as potential threats. In addition, participants found the location or the used device to be interesting since VAs can also be used in mobile devices. Moreover, some devices can also be used in different rooms at home. If audio

recordings are available, it would be useful to make them easily accessible and link them to the interactions to enable users, for instance, to immediately listen to unintended interactions or failed automatic transcriptions.

Draw attention to unintended interactions. Events and interactions that deviate from normal use or were recorded unintentionally were the most concerning data for our participants. These events should be highlighted so that consumers can reflect on these and maybe initiate further steps. Current labels like 'data not available' as offered by Amazon Alexa provide no information and force users to listen to every file manually. The transcribed voice recognition of these data would be a good starting point to further support consumers in quickly grasping these data.

Tell the consumer how the vendors (might) see them. Participants also asked what the data tells vendors and third parties about their use and elicited profiles. Our participants were aware of the fact that companies use VA data, for improving voice recognition and personalized advertising. In this vein, they wanted to understand how their data affects the analysis and what can be concluded about the consumer (Rao, Schaub, and Sadeh, 2015). This can be an analysis of preferences in music based on the users' playlists or even a personalized profile that can be drawn from the use of the VA.

Disclose the communication between devices and services. As a platform, a VA offers access to a variety of third-party services and an application ecosystem. Hereby, our participants were interested in exploring their own data traces beyond the boundaries of single services. A VA command including a skill or third-party device is likely to share data with these third parties and can consequently not be accessed by requesting a data takeout from Amazon or Google.

8.7.4 Limitations

Our study was focused on the impressions of account holders as they were the ones with access to the VA data. However, VAs collect voice data from all household members disregarding who has activated the device or who can be heard in the background. Future research should also consider other household members during the design process since VAs also collect data about them. This would be an important direction to broaden the scope of our study (Geeng and Roesner, 2019). It must also be emphasized that we used only the conversation data and left out other content and data from the takeout. Furthermore, it would be interesting to test our findings with a bigger and more diverse set of users. Especially in the case of more vulnerable groups of users, collaborative aspects might play a bigger role in the context of computing, controlling, and managing the smart home in general (Frohlich and Kraut, 2002; Randall, 2003) as well as VAs in particular (Y. Huang, Obada-Obieh, and Beznosov, 2020).

8.8 Conclusion

The aim of our study was to learn more about what users of VAs know and understand about data collection and their use by the vendor. We therefore developed a prototype to explore the conversation data with eleven users and gained three major insights: First, the exploration revealed that users were not or hardly aware of the data collection of unintended interactions. Second, users were surprised about the storage of the recordings and what other data or sounds might be associated with it. And third, resulting from the previous points and the other insights, we determine the need for a tool to strengthen

data literacy making conversations with VA systems and its context more transparent. This should lead to reducing the imbalance of knowledge and power to access and understanding users' data footprints at home. As related work had already shown the relevance of providing transparency about the amount and use of smart home data in general, this paper provides further insights with special regard to VAs. Furthermore, by confronting users with their data, we challenged their assumptions and supported their learning processes in understanding their usage behavior by showing them what kind of data VAs store.

For future research, we want to identify more needs regarding the unintended events and further allow analytics about topics and patterns users are interested in and of which they require more information (Lau, Zimmerman, and Schaub, 2018). As research on mobility (Jakobi, Stevens, et al., 2018) as well as energy data (Castelli, Ogonowski, et al., 2017) highlights, it is less necessary to present all data, but more important to alert consumers when unexpected events have occurred, specific kinds of information need to be requested or even to tell if everything works as expected. Hence, we plan to implement a dashboard based on the prototype with additional functions from which the users can choose the ones they need. This will go beyond the previous options of the prototype (visualization and categorization) to support further sense-making and management of other types of user data like audio recordings.

9

Alexa Why Did You Do That?

Exploring Voice Assistants' Accountability through Explanatory Dialogues

9.1 Abstract

As voice assistants (VAs) become more advanced leveraging Large Language Models (LLMs) and natural language processing, their potential for accountable behavior expands. Yet, the long-term situational effectiveness of VAs' accounts when errors occur remains unclear. In our 19-month exploratory study with 19 households, we investigated the impact of an Alexa feature that allows users to inquire about the reasons behind its actions. Our findings indicate that Alexa's accounts are often single, decontextualized responses that led to users' alternative repair strategies over the long-term, such as turning off the device, rather than initiating a dialogue about what went wrong. Through role-playing workshops, we demonstrate that VA interactions should facilitate explanatory dialogues as dynamic exchanges that consider a range of speech acts, recognizing users' emotional states and the

context of interaction. We conclude by discussing the implications of our findings for the design of accountable VAs.

9.2 Introduction

Modern VAs extend far beyond the scope of Suchman's seminal research on human-machine communication with the copy machine (Suchman, 1987), pushing the boundaries of interaction yet leaving users to navigate communication challenges. One major challenge with VAs is that unlike human interactions, where both parties can immediately ask questions to clarify potential misunderstandings and call one another to account for their actions, VAs are often less clearly solution-oriented and tend to end the interaction abruptly (Voit, Niess, et al., 2020). Previous studies have shown that because of the low level of support from the VAs in resolving communication breakdowns, the burden of ensuring successful communication falls on users, who have to adapt their communication patterns to a VA's needs (Beneteau, Richards, et al., 2019; Y. Cheng, Yen, et al., 2018). However, this process requires a lot of effort and is not always successful (Bentley, Luvogt, et al., 2018). As a result, VA users are often confused or even frustrated when errors occur.

As artificial intelligence (AI), natural language processing (NLP), and sensing technologies rapidly advance, VAs are becoming more proficient at understanding users' behaviors, intentions, and environments (Zargham, Reicherts, et al., 2022). This progress presents a wide array of possibilities for ensuring VAs' accountability when errors occur.

Accountability of VAs is considered as VAs' "obligation or willingness to take ownership or ultimate responsibility for the outcome of a task, including one that has been delegated" (Kianzad, Kim, et al., n.d., p. 1, 16). Researchers have begun to examine VAs' accountability through self-repair and providing explanations, and have found that VAs' explanations alleviate peoples' privacy concerns and greatly improve their assessment of an intelligent voice assistant if a mistake has been made (Cuadra, Li, et al., 2021; Seymour and Van Kleek, 2021). Others have looked into VAs' accountability as an interactional achievement rather than an intrinsic feature of actions and argue that designers should consider the *projected accountability* of requests in terms of the contextual circumstances that the user is supposed to experience (J. Cho and Rader, 2020; Porcheron, Fischer, et al., 2018; Reeves, Porcheron, et al., 2018).

In 2019 Amazon announced a new Alexa feature, allowing users to call Alexa to account for its specific action by asking "why did you do that?" (Amazon, 2024; Richard Nieva, 2019). In response, Alexa provides a short explanation about its response to users' recent voice requests, Bluetooth connections, routine triggers, and more. However, to date, there is a lack of studies into how such interactions with users demonstrate accountability and their situational effectiveness over the long term. To close this gap, we ask the following questions:

- **RQ1) In what ways do VAs currently demonstrate accountability with users through dialogues?**
- **RQ2) How might these dialogues be rendered to be more situationally effective?**

To explore these questions, we conducted a three-stage exploratory study involving 19 households over 19 months. This included fieldwork to evaluate the Alexa "why did you do that" feature in real-world settings, followed by two co-design workshops to capture situational aspects through role-playing, and a final interview to assess the users' long-term experiences. We identified that Alexa's responses to users' queries are often single, decontextualized responses to what it understood as a basic request without grasping what is being done with that request. This led users to adopt alternative repair strategies, such as turning off the device, rather than initiating a dialogue about what went wrong. Through our co-design workshops with both Alexa and Google Home users, we demonstrate that such interactions with VAs should be viewed as *explanatory dialogues*—dynamic exchanges that encompass a range of speech acts to recognize users' emotional states and the context of the interaction, thereby being situationally more effective.

9.3 Background

Interaction with VAs is a current topic in research on human-computer interaction (HCI) and conversational user interfaces (CUI). Our research intersects with three areas of VA research: (1) the exploration of VAs' errors, and (2) the need for VAs' accountability in such scenarios, as well as (3) the studies on the design of accountable VAs in practice.

9.3.1 Errors in Interaction with VAs

Errors, misunderstandings, and communication breakdowns are significant challenges in Human-VA interactions, with a wide array of research documenting these issues and exploring their underlying causes (e.g., Beneteau, Richards, et al., 2019; Clausen, Kyhn, et al., 2023; Daim, Ha, et al., 2012; Du, Zhang, et al., 2021; Mavrina, Szczuka, et al., 2022). Studies such as Schönherr, Golla, et al., 2020 highlight how VAs can be unintentionally activated by phrases resembling their wake words, not only causing annoyance but also posing privacy concerns (Pins and Alizadeh, 2021; Pins, Jakobi, et al., 2021). This problem is compounded by errors resulting from VAs misinterpreting unrelated user utterances, which is particularly problematic given their capabilities for making purchases or initiating calls (Mavrina, Szczuka, et al., 2022). The complexity of user commands and environmental factors further exacerbate these issues, with research indicating a tendency among users to simplify their queries to key words to avoid recognition errors (Bloom, 1980; Jiang, Jeng, and He, 2013; Kennedy, Wilkes, et al., 1988). In the context of family use, misunderstandings by VAs often require collaborative efforts from family members to rectify, illustrating the practical implications of these errors (Beneteau, Richards, et al., 2019).

On the analytical front, researchers have sought to identify potential reasons behind VA errors and develop frameworks for their categorization. For example, variations in intonation have been found to increase the likelihood of accidental VA activations (Siegert, 2021). Bohus and Rudnicky, 2008 created a taxonomy for understanding errors caused by both environmental noise and misaligned user-VA goals. Further research by Higashinaka et al. has emphasized the importance of context in understanding dialogue

breakdowns, suggesting that a more nuanced approach to categorizing these errors is necessary (Higashinaka, D'Haro, et al., 2021).

In sum, the literature underscores the prevalence of various errors in Human-VA interactions, highlighting the crucial need for VAs to offer users clear guidance on resolving these issues, thereby enhancing the user experience.

9.3.2 Accountability of VAs

People routinely try to perform social actions in a way that make them appear as accountable to others and to the situation. This is an ongoing concern for members of society as they routinely attempt to account for what they are actually doing in the face of possible inadequacies in accountability for their actions (Porcheron, Fischer, et al., 2018). Accountability is defined as an entity's obligation or willingness to assume responsibility for the outcomes of a task, including those that are delegated (Y. Cheng, Yen, et al., 2018).

Researchers have found that users often interact with VAs in a manner reminiscent of human-to-human communication, imbuing these interactions with expectations of accountability similar to those held for other people (Porcheron, Fischer, et al., 2018). For example, studies by Cho and Rader have shown that users treat VA responses as they would human feedback, using it to refine their understanding of the VA's capabilities and to establish a common ground for achieving their goals, regardless of the response's relevance (J. Cho and Rader, 2020).

Lahoual et al. explored user behavior in scenarios where VAs misunderstood requests or acted unexpectedly, uncovering seven reasons users "forgive" VA errors, known as the "usage paradox" (Lahoual and Frejus, 2019). This

paradox reflects the counterintuitive behavior where users continue to accept and use VAs despite encountering failures. These seven reasons include users forgiving isolated errors due to overall satisfactory system performance, tolerating failures in anticipation of future improvements, valuing the VA's integration into daily routines, showing initial forgiveness by understanding error causes, overlooking malfunctions due to the ability to correct them via conversation history, seeing error correction as an opportunity to 'train' the VA, and acknowledging the VA's potential to learn and evolve.

Further, in a five-weeks study of the behavior of adult and child users in situations where communications fail, (Mavrina, Szczuka, et al., 2022) illustrated a significant negative effect of users' abandoned requests from VA on their satisfaction in interaction with the device. The authors also showed that users' motivation to apply specific repair strategies (such as the use of repetition) decreased significantly over time. Similarly, Caudra et al. manipulated Alexa's responses after mistakes to demonstrate that a VA's ability to recognize and attempt to correct its own errors (self-repair) significantly enhances users' perceptions of an intelligent voice assistant (Cuadra, Li, et al., 2021).

Collectively, these studies illustrate a unanimous conclusion among researchers: users not only expect VAs to demonstrate accountable behavior in the face of errors but also find that such accountable responses elevate the quality of interaction.

9.3.3 Designing for Accountable VAs

The literature relating to explainable, trustworthy, and accountable AI has become extensive (e.g., Alizadeh, Mniestri, and Stevens, 2022a; Alizadeh, Pins, and Stevens, 2023; Alizadeh, Stevens, and Esau, 2021; Ehsan and Riedl, 2020; Ehsan, Tambwekar, et al., 2019; Q. V. Liao and Varshney, 2021; Ras, van Gerven, and Haselager, 2018), much of which is not of direct relevance to this paper. Nonetheless, the focus of this discussion is particularly pertinent given the increasing use of VAs for more high-risk tasks, such as medical diagnostics or financial decision-making, where errors can have significant consequences. This trend underscores the need for designing VAs that not only perform tasks but also do so with transparency and accountability. For instance, Hocutt's examination of Alexa's decision-making process in source selection revealed a default preference for retail solutions. He underscored the importance of providing users with reasons for source selection beyond mere citations or URLs (Hocutt, 2021).

In a move towards greater VA's accountability, Amazon introduced a feature that enables users to inquire about Alexa's actions. This view of VA's accountability as an interactional achievement, was also explored by Porcheron et al. in the practical design of user requests. The authors argued that responses from digital assistants should be resources for users to gauge the system's state for further interaction (Porcheron, Fischer, et al., 2018). Echoing this, Cho and Rader advocated for designing responses to serve as "evidence" that coordinates a conversational process, as opposed to optimized outputs (J. Cho and Rader, 2020). This approach would allow users to continually explore a system's functionality and capabilities over the long term.

Kianzad et al. explored links between degree of automation, accountability of voice interfaces, and user satisfaction (Kianzad, Kim, et al., n.d.). The authors moved beyond understanding users' perception of accountability to direct it through design. They described accountability-aware design of VA as design that considers context, "specifically task complexity, type of user(s), and ambiguity of the interactions" (p. 7, 32). This study also showed that recommendations and instructions from the system generate a sense of "shared accountability" among the users. Seymour investigated designing VA accountability through trust and privacy explanations, recommending the use of analogies to enhance user understanding (Seymour and Van Kleek, 2021).

Nevertheless, research focusing on the long-term accountability of VAs, spanning over a year, remains limited. In this paper, we move beyond the existing research on designing accountable VAs by investigating how the accounts offered by VAs are perceived by users in long term and how those accounts might be rendered more situationally effective, specifically with Alexa's accountability feature as a case.

9.4 Method

Acknowledging the domestication and integration of VAs into users' daily routines and practices, previously researchers have advocated for studying VAs in their natural environments (Beirl, Rogers, and Yuill, 2019; Pins, 2019; Siegert, 2020). This study was part of a larger 3-year project, focused on developing privacy-friendly and trustworthy VAs (see Figure9.1).

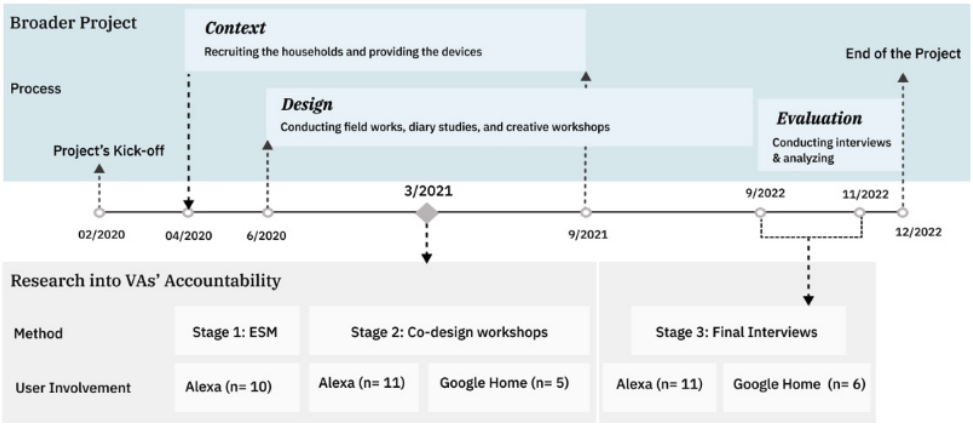


Fig. 9.1.: Overview of the process.

Starting in April 2020, we provided both novice and experienced users with VA devices like Amazon's Alexa (Echo, Echo Dot, and Echo Show) and Google Home VAs (Google Home, Google Nest Mini, and Google Nest Hub), tailoring our selection to households' preferences. While the broader project's recruitment concluded in September 2021, our specific investigation into VAs' accountability was triggered in March 2021 following an announcement from Amazon about a new feature for Alexa released in Germany (see 9.3 in the Appendix 9.8.1). Therefore, this study engaged participants who had joined the project by June 2020 via social media, personal contacts, and university mailing lists. Our study ultimately involved 19 households—13 utilizing Alexa and 6 using Google Home, encompassing diverse living arrangements: four single occupants, five living with a partner, five in shared housing with roommates, and five residing with their families and children. Detailed demographics are provided in Table 9.8.2 in the Appendix 9.8.2.

Although we encouraged all family members to engage, our study specifically involved interactions with the devices' primary users—one per household—resulting in a participant group of 14 men and 5 women, with an

average age of 33. The ethical committee of University of Siegen granted approval for this study.

9.4.1 Procedure

In February 2021, German Alexa users were notified by Amazon about a new feature allowing them to inquire why Alexa performed certain actions (see Figure 9.3 in the Appendix 9.8.1), roughly a year after they started using the devices. To test the situational effectiveness of this new feature we applied the Experience Sampling Method (ESM) (Hektner, 2007), prompting participants to test the Alexa "why did you do that" feature and document their experiences in an online folder immediately after completion (stage 1). To explore what users expected from the VA's accountability through this feature and how such interactions could be rendered to address users' expectations, we then held two co-design workshops (Zamenopoulos and Alexiou, 2018) with users of both Alexa and Google assistants (stage 2). Finally, after 19 months, we interviewed users of both Alexa and Google Home to explore how VA behavior in response to errors influenced their interactions and accountability needs over the long term (stage 3).

ESM: Testing Alexa's Features

In early March 2021, we tasked Alexa users to query Alexa about its actions when errors occurred ("Alexa, why did you do that?") and document their experiences immediately. They had a month to use this feature, ensuring real-time application when errors occurred. The instruction email encouraged users to detail the context of using the feature, Alexa's response, and its

effectiveness. We also requested information from those unable to use the feature by the deadline, asking for reasons. Out of 13 Alexa users, we received 10 complete reports.

Co-design workshops: Understanding Users' Expectations

We advanced our study by inviting all 19 households to participate in co-design workshops, with 16 accepting. To foster interactive sessions and balanced participation, we split them into two groups (Workshop 1: 6 men and 2 women, avg. age 34; Workshop 2: 5 men and 3 women, avg. age 33) (Refer to Table 9.8.2 in the Appendix 9.8.2 for participant details). Due to the COVID-19 pandemic, these workshops were held remotely via Zoom and Miro [7]. in late March 2021, each lasting three hours and recorded with participants' consent.

The workshops were structured in six steps: (1) A warm-up exercise facilitated introductions among participants. (2) Participants shared their experiences with VA's errors and outlined scenarios where VAs should account for their behavior. (3) Using the affinity diagram method (Lucero, 2015), participants grouped these scenarios into categories based on their similarity to each other. (4) To diversify design activity considerations, we asked the participants to select two scenarios across different categories, formulating two inquiry questions per scenario, each with five potential responses from the VAs. (5) Using HCI's performative methods for exploring human-object relationships (e.g., Dörrenbächer, Löffler, and Hassenzahl, 2020; Elsden, Chatting, et al., 2017), we guided participants to role-play their top two

[7]Miro is a digital collaboration platform designed to facilitate remote and distributed communication and project management.

responses in pairs, alternating between user and VA roles, with the option to mimic VA voices. Imitating another being helps to understand similarities and differences between oneself and the other being (Dörrenbächer, Löffler, and Hassenzahl, 2020). Hence, the imitation and role-playing aimed to enhance understanding of the similarities and differences between humans and the VAs, specifically focusing on how each accounts for their behavior differently. (6) Finally, the participants reflected about their expectations of VA's accountable behavior by comparing their own design with the VA's current feature.

Final interviews: Tracking the Changes

In the final phase, from September to November 2022, we conducted semi-structured interviews with both Alexa and Google Home users. We explored whether errors recurred and if users attempted to hold the VAs accountable for their actions. The goal was to assess any changes in the long-term usage of Alexa's feature and its impact on users' accountability needs. We also examined changes in the usage of Google Home, focusing on how it addressed or failed to address its behavior and users' accountability needs. With participants' consent, these interviews were conducted remotely via Microsoft Teams and were recorded. Each interview lasted between 40 minutes to an hour and was conducted in German.

9.4.2 Analysis

In the analysis phase, we anonymized the data by transcribing, translating, and removing personal identifiers from the interviews, assigning each

participant a unique identifier (P#). Following the principles of inductive thematic analysis (Clarke and Braun, 2017), two researchers coded each ESM report in-vivo. This initial phase culminated in achieving consensus and setting a foundation for structuring, discussing, and interpreting the data, leading to the creation of an initial codebook with two main categories (decontextualized, single responses and "I'm just learning" responses) and their subcategories. During the in-depth coding of the final interviews, these categories were refined and adjusted as needed, with a primary focus on the alternative repair strategy and its subcategories such as repositioning and turning off devices.

In two co-design workshops, participants generated 80 different responses for 16 different scenarios. Figure 9.2 shows 40 of these dialogues, which are the results from Workshop 1. These were analyzed alongside Miro board data to further refine the codebook by adjusting existing themes and adding new categories to better reflect user expectations. In the next phase, we employed MAXQDA Analyzer to systematically map all interview statements onto our established categories and subcategories. The final phase saw researchers reaching consensus on the themes during two interpretation sessions, each lasting approximately two hours, systematically refining our understanding of the relationships between the identified themes. The following sections focus on the themes relevant to our research questions.

9.5 Findings

In this section, we first start with results from the experience sampling study and interviews then discuss the findings from our co-design workshops.

9.5.1 Exploring VA's Accounts over Time (RQ1)

Analysis of user reports suggests that initial interactions with Alexa's feature frequently led to immediate frustration. This often resulted in users abandoning the feature after just one use or a few attempts. Over time, such frustration contributed to a broader disengagement with the device, prompting users to resort to alternative repair strategies, such as unplugging the device or repositioning it. This reaction was not unique to Alexa users; Google Home users, who lacked specific features for addressing errors, exhibited similar responses.

The primary issue with Alexa's feature was that the responses were decontextualized, designed more for single interactions rather than considering the need for a conversational flow that includes back-and-forth communication between the user and the device. In the subsequent subsections, we will further examine these issues with VA responses and discuss their long-term implications.

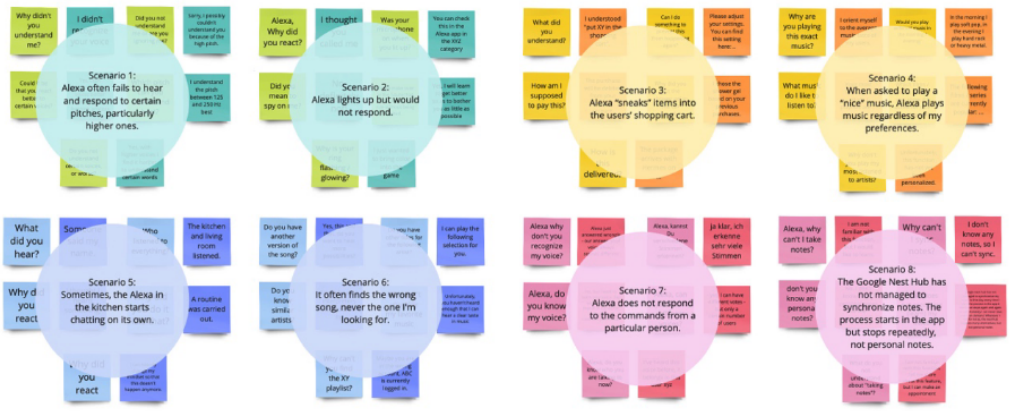


Fig. 9.2.: Eight scenarios and five dialogues generated for each during Workshop 1

Decontextualized single responses lead to abonnement of the feature.

Successful conversations between humans require exchanges signaling mutual understanding, allowing the conversation partners to easily correct misunderstandings by updating shared knowledge and beliefs (Clark, 1996). However, Alexa's accountability feature was limited to recognizing only a narrow set of programmed responses, hindering its ability to resolve misunderstandings by adapting to context. For instance, P1 initially reported, when he asked Alexa why the light ring was activated unexpectedly, Alexa replied, "*I can only repeat the last command,*" not addressing P1's actual query. One year and a half later, P1 explained that this error still occurred, but he no longer tried to figure out why: "*The ring still lights up suddenly, but I don't try to find out the reason. I just unplug Alexa when this happens*" (P1). This scenario highlights how Alexa's decontextualized responses can discourage users from engaging further with holding the device accountable for its action and lead to alternative repair strategy of unplugging the device.

A conversation involves not just the spoken words but also the unspoken context, relying on the partners' interpretive abilities and imagination to fill the gaps (M. Lee, 2020). The core issue with Alexa was its inability to interpret the situational context and respond appropriately, failing to fill the gaps in conversation and address queries based on situational elements. P5's experience illustrates this challenge. At one point, P5 asked Alexa what time it was. Alexa responded by telling her the time and wishing her a nice day. P5 then asked, "*Why did you do that?*" and Alexa replied that it heard a query about the time. Alexa failed to grasp that P5 was inquiring about the added wish, not the time itself, showing a lack of mutual understanding and

contextual insight. P5 recounted, "*She [Alexa] didn't explain why she wished me a nice day. When I pressed, she suggested I try a skill, 'A stupid spell,' which left me more confused about her reasoning and why she recommended that skill.*" Instead of addressing the misunderstanding by establishing common ground, Alexa immediately suggested a solution unrelated to P5's query. A year and a half later, P5 noted that while misinterpretations continued, she ceased to engage with Alexa on these issues, opting to disable the problematic functions: "*Alexa still misinterprets my request for Spotify as Amazon Music... Now, I just turn the Music app off*" (P5).

While Alexa's initial responses to queries were not necessarily unrelated to the situation, participants felt these answers were inadequate for fully understanding errors or misunderstandings, expressing a desire for more dialogic interaction to explore what went wrong. However, the feature's design, which predominantly supports single interactions, limited opportunities for further dialogue. P10's experience underscores this limitation: after inquiring why Alexa spontaneously played the radio, he received a response that failed to facilitate an ongoing conversation. "*Alexa's response to my question was, 'I thought I heard a command to play the radio. Would you like to know what I heard?' When I answered 'Yes', Alexa simply echoed 'Yes'. The result was sobering.*"

Over time, issues like device miscommunication and misplaced commands persisted, leading participants like P10 to adopt pragmatic yet unsatisfactory solutions. "*A year later, Alexa still misinterprets commands, like activating timers or music in unintended rooms... I've stopped asking why; now, I just deactivate Alexa in those rooms.*" This narrative highlights the systems' inability to engage users in meaningful dialogue or effectively address misunderstand-

ings. Similarly, Google Home users lacking the specific feature to query the device about its actions resorted to alternative strategies. For example, P9 noted in the final interview, "*I still encounter situations where I ask Google to set a timer, and it sets it in the wrong room. Like, I'm in the kitchen, but the timer goes off in the bedroom. I just turn Google off when this happens*" (P9).

"I'm just learning" responses keep users temporary hopeful.

A common issue users faced was the "I'm just learning" response, which, while falling short of explaining errors, offered hope for future improvements. For instance, P6 shared, "*For simple queries like 'Alexa, turn the light on,' I received the command I had given. For more complex commands, I got 'I can only explain recent voice requests. I'm just learning how to explain more about my actions. You can find more information in the Alexa app'*." Despite labeling the feature "unhelpful," he remained optimistic about future improvements: "... I will continue to try it [the feature] with more complex voice commands, hoping for better results."

P8's frustration stemmed from specific issues, such as setting timers incorrectly or playing the wrong music, met with the same generic response. Yet, she speculated on the response's potential purpose: "*Perhaps Amazon is using users' interactions as feedback to improve Alexa. I'll continue to see if Alexa gets better at this*."

However, this type of response did not translate into long-term benefits either. P6 recounted: "*There was one year of relief from this type of errors [Alexa suddenly lights up], and recently they are happening again.*" This led P6 to reposition the device rather than seek further explanations: "*We moved Alexa to the dining table, thinking it was unintentionally triggered by TV ads. But I*

don't know why the error still happens; I don't ask why. The feature remains unchanged and doesn't offer any insights." P8 also abandoned the problematic function, opting for avoidance over resolution: "*I just don't use the functions that don't work well anymore. I don't try to fix them. Like, for example, having a shopping list. I don't use it anymore*." These narratives reveal that while the "I'm just learning" response initially kept users hopeful, it ultimately led to dissatisfaction and disuse of certain features due to persistent issues and lack of improvement.

In sum, Alexa's responses were often narrowly programmed and decontextualized, failing to establish a conversational foundation relevant to the situation or clarify misunderstandings. While the "I'm just learning" responses initially raised hopes for future improvements and encouraged short-term user engagement, the lack of progress in addressing misunderstandings or adapting to context led users to revert to their own repair strategies, such as disabling or ignoring problematic features (also documented in previous research, e.g., (Ashktorab, Jain, et al., 2019; Beneteau, Richards, et al., 2019; Y. Cheng, Yen, et al., 2018; Clausen, Kyhn, et al., 2023; Mavrina, Szczuka, et al., 2022).

9.5.2 Users' Expectations from VA's Accountability Feature (RQ2)

We outline our findings from the co-design workshops in subsequent sections. Initially, we explore insights regarding the user-designed VA responses, comparing these with Alexa's features. Following this, we transition to our findings from the role-play activity.

Insights from the dialogues

Our analysis of the generated dialogues during the co-design workshops was guided by our research question, focusing on how responses accounted for errors rather than whether they provided valid explanations for what went wrong. Comparing user-designed dialogues with Alexa's accountability feature, we identified four main differences:

1. Accounting for users' feelings: User-designed VA responses frequently accounted for users' feelings by owning up to mistakes. These accounts were often conveyed through apologies for the error (e.g., "Sorry") or by using prefatory terms like "unfortunately" (see the Examples in Table 9.1).

Tab. 9.1.: User-designed dialogues in which the VA accounts for users' feelings.

#	Scenario	User's query	VA's response
1	When asked to play a "nice" music, Alexa plays music regardless of my preferences.	Why don't you play my most played songs?	Unfortunately, you have not listened to enough songs, so this feature is not personalized for you.
2	Alexa does not respond to the commands from a particular person.	Alexa, can't you recognize my voice?	Sorry, I didn't hear you. I find it difficult to hear certain words at a higher pitch of voice.

2. Situational awareness: In many instances, participants did not specify the exact problems or errors they referenced. This ambiguity might stem from the activities design, wherein participants were aware of the scenario, leading to less explicit communication. But reports on Alexa's accountability feature indicates an expectation also for the devices to possess a shared contextual and situational understanding, echoing findings from previous research (J. Cho and Rader, 2020; Murad, Candello, and Munteanu, 2023; Porcheron, Fischer, et al., 2018). This expectation is highlighted in the designed responses; for instance, in

Example 3 from Table 9.2, both the user and Alexa understand which specific shower gel is being discussed, and in Example 4, the user anticipates Alexa to play a recording based on her own interpretation, without the need for explicit instruction.

Tab. 9.2.: User-designed dialogues in which the VA possesses situational awareness.

#	Scenario	User's query	VA's response
3	Alexa "sneaks" items into the users' shopping cart.	Why did you put exactly this shower gel in my shopping cart?	I chose the shower gel based on your past purchases.
4	Alexa lights up but would not respond.	Was your microphone on when your activity light was on?	Yes, because I received this command: [plays her interpretation of the microphone's recording].

3. Progressive but minimal: The accounts designed by our participants highlight the value of minimalism. Despite not being directed to craft concise responses, participants naturally opted for brevity in their designs. However, they subtly introduced the concept of interaction progressivity (J. E. Fischer, Reeves, et al., 2019), such as by having the device guide the user on subsequent steps (refer to the examples in Table 9.3).

Tab. 9.3.: The progressivity, as well as co-performance and shared accountability in the user-designed dialogues.

#	Scenario	User's query	VA's response
5	VA recognizes the voice of a random person as the device owner.	Why did you do that?	The person has a similar voice to yours. We can retain your voice.
6	Wrong device is activated by the command.	How can we solve this problem?	Let us test the rooms and the name of the devices. We can then test the configuration of multi-room music.
7	VA is activated randomly.	How can we break you of the habit of getting activated randomly?	By reading these sentences to me, so that I can improve my speech recognition.

4. Co-performance and shared accountability: As can be seen in Table 9.3, our participants went beyond crafting progressive responses that would instruct on what to do next. They framed their responses based

on the concepts of shared accountability (Kianzad, Kim, et al., n.d. and co-performance (Kuijer and Giaccardi, 2018). This means that the VA attempts to learn about the problem through interplaying with the user (co-performance) and that VA's instructions are presented as suggestions (shared accountability). This approach invites participants to work together towards a goal, while giving them the choice to make the final decision on whether they want to solve the error or not. Here, participants are still accountable for their decisions, embodying the principle of shared accountability.

Insights from the role-playing

Participants' engagement in the role-playing task revealed a range of emotional reactions to the VAs' errors, encompassing not only verbal responses but also non-verbal behaviors such as gestures, facial expressions, and body movements, reflecting their communication partners' emotional states. Although the role-playing was conducted online—a medium initially thought to hinder the transfer of movements and emotions, potentially leading to a less immersive experience—we were surprised to find that participants still became acutely aware of these communication nuances. They also recognized the importance of the context in which the dialogue was happening and its potential effects, leading them to make adjustments accordingly.

For instance, they found that in scenarios where the VA had added items to the users' cart without consent, a response like 'I chose the shower gel based on your past purchases' could be perceived as invasive and creepy, similar to privacy concerns highlighted in previous research (Seberger, Shklovski, et al., 2022). To mitigate this, participants revised the response to be more

reassuring: *"I selected this item based on your previous purchases, but you can turn off this feature if you prefer. Would you like to know how?"* This adjustment reflects participants' dilemma between informing and reassuring the (fictive) user, a topic also discussed by Seymour and Van Kleek, 2021 in the context of explanation design. Participants' decisions to lean towards one approach over the other were influenced by the specific circumstances in which the errors occurred and the perceived emotional state of the (fictive) user.

Furthermore, participants acknowledged that not every situation allows for a direct solution. At times, the technical capabilities of devices do not meet the behavioral expectations set for VAs. However, they expressed a desire to be informed about such limitations, distinguishing between issues that could potentially be resolved through specific repair strategies and those that could not. In these discussions, participants stressed the importance of clear communication about the problem's cause, favoring the elimination of uncertain terms like 'maybe' to provide more definitive reassurance.

9.6 Discussion

In this section, we borrow from the concept of explanatory dialogues (Arioua and Croitoru, 2015) and apply it to the accountable design of VAs (section 9.6.1). Building on this, we explore how explanatory dialogues can be designed to bring to life the possibility of designing VAs as engaged counterparts and to meet users' expectations of being understood and supported when communication with VAs fails (Section 9.6.2). We conclude with an admission of our shortcomings and a call for future research to apply, test, and add

to the preliminary checklist proposed in this paper for designing explanatory dialogues (Section 9.6.3).

9.6.1 Explanatory Dialogues as Species of Accounts

In this work, we explored the accountability of VAs by adopting the approach of Porcheron, Fischer, et al., 2018 to examine accountability as an interactional achievement. However, we expanded this perspective by intensively investigating the situatedness of accountability over the long term. This approach enabled us to update J. Cho and Rader, 2020's findings regarding unrelated responses. We demonstrated that although such responses —responses like 'I'm just learning'— may appear helpful in the short term for understanding system status and accepting system failures (Lahoual and Frejus, 2019) in the long run, users prefer alternative repair strategies that are genuinely more effective.

Our approach confirms studies that users' repair strategies change over time showing less willingness from the users to engage in dialogues with VAs in the face of errors over time and applying alternative strategies for repair. Although we observed "usage paradox" in short term, specifically tolerating failures in anticipation of future improvements as promised by the device, we saw that in the long users limited their usage (Lahoual and Frejus, 2019) in those situations where the device styed unaccountable over longer period of time.

On the basis of our findings, especially the gap between actual Alexa accounts and desired accounts uncovered in co-design workshops, we aim to move beyond the existing literature with the concept of explanatory dialogues.

Explanatory dialogues are defined as speech acts in which an explanation is requested and given, and are represented in the system as dialogue moves (Alizadeh, Pins, and Stevens, 2023; Arioua and Croitoru, 2015). In this model, two different sets of moves are allowed: one for the explainer and one for the explainee. The dialogue starts with the explainer giving a statement, whereupon the explainee requests an explanation. The explainer can then try to explain or declare that they are unable to explain. In the former case, the explainee may ask for further explanation or confirm their understanding (Arioua and Croitoru, 2015). Characterizing such interactions with VAs as explanatory dialogues draws attention to two important aspects of interaction:

1. Unlike humans, who can communicate and account for their actions both verbally and non-verbally, the communicability of VAs is restricted. They operate from a position where they can respond but not initiate actions or account for them in a human-like manner (Lind and Dickel, 2024). The concept of explanatory dialogues broadens our understanding by shifting our focus from merely what is said in an utterance to what is accomplished with it in a given situation (Grice, 1975). This shift enables us to move beyond merely factual explanations—declarative responses that inform users about the state of affairs (e.g., "Because I received the following command")—to incorporating various types of speech acts tailored to the situation. These can include expressive statements that acknowledge the user's emotional state (e.g., "I'm sorry"), and commissive statements in which the VA commits to an action (e.g., "By processing these sentences, I can improve my speech recognition capabilities"). Considering what is being done with the

utterance in specific situation helps us to move towards the projective accountability of the response i.e. the contextual circumstances that the user is supposed to experience (Porcheron, Fischer, et al., 2018).

2. Second, this concept emphasizes the back-and-forth nature of speaking and listening between the explainer and the explainee. Our research indicates that users anticipate a dialogic interaction with VAs during error scenarios. The dialogic character of their crafted responses is designed to be open to sequential elaboration. These elaborations play a crucial role in contextual sense-making and the repair process. However, currently interactions with VAs often lead to conversational dead-ends, lacking the opportunity for further elaboration. This suggests to the user that the VA was not listening in earnest and is not a conversational partner, but merely a medium through which communication takes place. Therefore, the principles of explanatory dialogues guide us toward developing VAs as active listeners, which in turn enhances users' perceptions of emotional support (E. Cho, Motalebi, et al., 2022).

We would like to point out that although explanatory dialogues account for VA behavior, not every explanatory dialogue is a good account. We found that from the user's perspective, good explanatory dialogues are contextual (J. Cho and Rader, 2020; Murad, Candello, and Munteanu, 2023; Porcheron, Fischer, et al., 2018), non-creepy (Seberger, Shklovski, et al., 2022), informative (J. Cho and Rader, 2020; Seymour and Such, 2023), minimal (J. E. Fischer, Reeves, et al., 2019; Porcheron, Fischer, et al., 2018; Reeves, Porcheron, et al., 2018), progressive (J. E. Fischer, Reeves, et al., 2019), cooperative (Dybkjær, Bernsen, and Dybkjær, 1996; Grice, 1975), reassuring (Seymour and Such, 2023), and responsive to the user's emotional state

(Spitale and Garzotto, 2020; Wilhelm, Otten, et al., 2022). However, user satisfaction is only one goal in the design of explanatory dialogues. A satisfactory explanatory dialogue may be misleading or even completely wrong. Ideally, explanatory dialogues would present users meaningful opportunities to identify undesirable behavior by VAs and prompt them to improve their current and future interactions (Eslami, Krishna Kumaran, et al., 2018).

9.6.2 Design Implications: Towards Contextual Sense-making and Collaborative Repair

Given that users of VA can see the interaction history when they log into their account, we should ask what benefit this modality change has, especially if it only provides them with the same level of information without the possibility for further elaboration. One perspective is that this change in modality could enhance situational effectiveness and facilitate better contextual sense-making for users. When users' interaction histories are presented detached from the context of the errors, it creates greater opportunities for collaborative repair with the VA.

Previous research has offered guidelines and heuristics for designing voice user interfaces (VUI) (Murad, 2019; Murad, Candello, and Munteanu, 2023; Murad, Munteanu, et al., 2021), with some studies focusing on more specific scenarios, such as designing requests (Porcheron, Fischer, et al., 2018) or addressing interactional elements like progressivity (J. E. Fischer, Reeves, et al., 2019). Continuing in this vein, we explore the implications of our study for design, both narrowly and broadly. In a narrower sense, we discuss how our findings can inform the design of accountability mechanisms through

explanatory dialogues. On a broader scale, we examine what our results reveal about effectively communicating complex VA behaviors to users.

A checklist for designing accountability through explanatory dialogues.

Previous research has suggested that an accountability-aware design should consider the context, especially with respect to task complexity, types of users, and the ambiguity of interactions (Kianzad, Kim, et al., n.d.). In this paper, we examine context and situation more closely to investigate how we can incorporate accountability into dialogues during error occurrences. Drawing on our findings and in line with the approaches of prior studies (e.g., J. E. Fischer, Reeves, et al., 2019; Porcheron, Fischer, et al., 2018), we propose six questions that designers can ask themselves in the design process of accountability through explanatory dialogues (see Table 9.4). These questions, while previously acknowledged within existing general CUI guidelines, assume a specific importance in our discussion on designing accountability through explanatory dialogues. We have compiled these questions into what we term the "Explanatory Dialogue Checklist":

First, our study shows that users prefer explanatory dialogues that incorporate a range of speech acts, such as expressions acknowledging feelings like anger or confusion, along with apologies. Therefore, designers should first consider (Q1) which speech acts are relevant for this situation. Additionally, users found the current Alexa feature "not helpful" as it only provides information they "already knew." Thus, designers should identify (Q2) what new information to include in dialogues. Also, the role-playing activity high-

lighted the importance of careful wording in error situations, revealing a user preference for reassuring language in such contexts (Q3).

Tab. 9.4.: Explanatory dialogue checklist.

#	Questions to design accountability through explanatory dialogues
Q1	What types of speech acts are relevant for this situation?
Q2	What information does the user not know? Does the explanatory dialogue include any information that the user does not already know?
Q3	Is the explanatory dialogue reassuring in terms of wording?
Q4	Is the explanatory dialogue open to sequential elaborations?
Q5	Is the explanatory dialogue consistent with the capabilities of the VA?
Q6	What possibilities for action does the user have after the explanatory dialogue?

Furthermore, while user-designed responses were minimal, they were open to sequential elaborations (Q4), a point designers should note. We observed that users understand technical limitations and prefer clear communication about the feasibility of resolving issues rather than vague assurances of device learning in the long run. This requires ensuring (Q5) that the dialogue aligns with the VA's actual and potential capabilities, and that the responses do not promise learning that is not reflected in the device's future interactions. Lastly, effective dialogues should encourage users to collaborate with the device towards a goal, checking how dialogues prompt user actions after communication (Q6).

From shared accountability to shared agency.

The concept of agency in VAs suggests a degree of autonomy and intentionality; however, their actual capacity is limited to responses rather than initiating actions or accounting for their behaviors. We view agency and actors not as inherent qualities but as characteristics that emerge through communicative acts (Lind and Dickel, 2024). Our findings indicate that users

expect VAs to account for their behavior and engage in explanatory dialogues when errors occur, a premise that aligns with existing research portraying VAs as social actors (Porcheron, Fischer, et al., 2018; Purington, Taft, et al., 2017). Furthermore, our study illustrates that users anticipate a collaborative performance with VAs. The notion of co-performance treats artifacts as participants capable of exerting agency alongside humans in social practices (Kuijer and Giaccardi, 2018). This approach encourages the development of AI systems that evolve through human interaction, creating meaningful partnerships that extend beyond simple autonomous task completion.

Cho and Rader advocate for a partnership model between humans and VAs, grounded in co-performance (J. Cho and Rader, 2020). This model highlights the importance of automatic communication from humans and the strategic selection of responses by VAs to bolster interactions. Its goals include enhancing data collection, unveiling the VA's capabilities to users, and establishing realistic expectations to support learning. However, mechanisms based on instructions and recommendations, while initially improving learnability, might become overly repetitive over time. Therefore, Kianzad et al. suggests that designers should implement transitional mechanisms, where system operations gradually offer fewer explanations and become more automated, advancing the notion of shared accountability through continuous learning (Kianzad, Kim, et al., n.d.).

We advance this conversation by emphasizing the situated nature of interactions. The aspiration that VA devices behave in such a way that explanations are superfluous is ultimately unattainable. Instead, we suggest reimagining explanations as dialogues with the VA tailored to specific situations. Explanatory dialogues open up new avenues not only for interaction designers, but

also for CUI researchers, and promote a broader interdisciplinary discourse on the multifaceted role of VAs' accounts. This includes considering whether communicative initiatives should always originate from the explainee, or whether the VA should proactively provide accounts of its behavior in certain contexts. Prior research has shown that users prefer proactive VAs in emergency situations or when technical support is needed (Zargham, Reicherts, et al., 2022). In light of this, we argue for a post-humanist approach that cultivates shared agency through interaction and marks a shift from traditional paradigms towards a more integrated and interactive future.

9.6.3 Limitations and Future work

Our work has broadened the scope of research on situational effectiveness of VAs' accounts by framing such interactions as explanatory dialogues and providing insights into the usage of Alexa's accountability features in real-world settings. However, we would like to point out some unavoidable limitations of this research.

First, our sample does not represent all VA user groups. In particular, our data collection was primarily conducted with the administrator of the devices in each household. We attempted to address this issue by asking questions about all household members, but it is likely that the administrators had their own view of things that may not have been shared by others in the household. Therefore, our theoretical findings need to be tested across different user groups to validate our findings.

Our approach to VA accountability diverges from human-human interaction, focusing on creating explanatory dialogues that address users' basic needs

rather than achieving Turing test-like realism. This concept, while modellable, varies in necessity across different user groups as demonstrated by the wider explainable AI research (e.g., Ribera and Lapedriza, 2019; Riveiro and Thill, 2021), underscoring that design solutions cannot be one-size-fits-all (Arya, Bellamy, et al., 2019; Sokol and Flach, 2020). Consequently, we advocate for flexibility and adaptability in catering to varying contexts and user needs to effectively tackle the challenge of designing situationally effective VAs' accounts.

Our study also suggests that the value of the information given in an explanatory dialogue is not always the same and depends on situational factors. In some situations, the reason for what happened is of great value while, in other situations, it is more important for users to figure out how to prevent the situation. Future work could explore what aspects of a situation determine the explanatory value of the different information provided.

Lastly, we recognize that the explanatory dialogue checklist introduced here is not exhaustive but serves as a foundation for future exploration into accountability via explanatory dialogues. Moving forward, our goal is to refine and validate the checklist through practical application with a broader user base.

9.7 Conclusion

In this paper, we report on a three-stage exploratory study conducted over a 19-month period in 19 households. This included a field study of users' experiences with Alexa's "Why did you do that?" feature, two co-design workshops employing role-playing to explore situational nuances, and inter-

views assessing long-term experiences. We observed that Alexa often issued single, decontextualized responses, leading users towards alternative repair strategies instead of engaging in dialogue. Advocating for these responses to evolve into interactive explanatory dialogues, we aim to facilitate contextualized understanding and collaborative repair. Our findings echo Suchman's prediction of shared accountability between humans and machines (Suchman, 1987), underscoring the need for research into shared agency through effective accounts. This study aims to spur further exploration into the design of accountable and transparent systems.

9.8 Appendices

9.8.1 Appendix 1

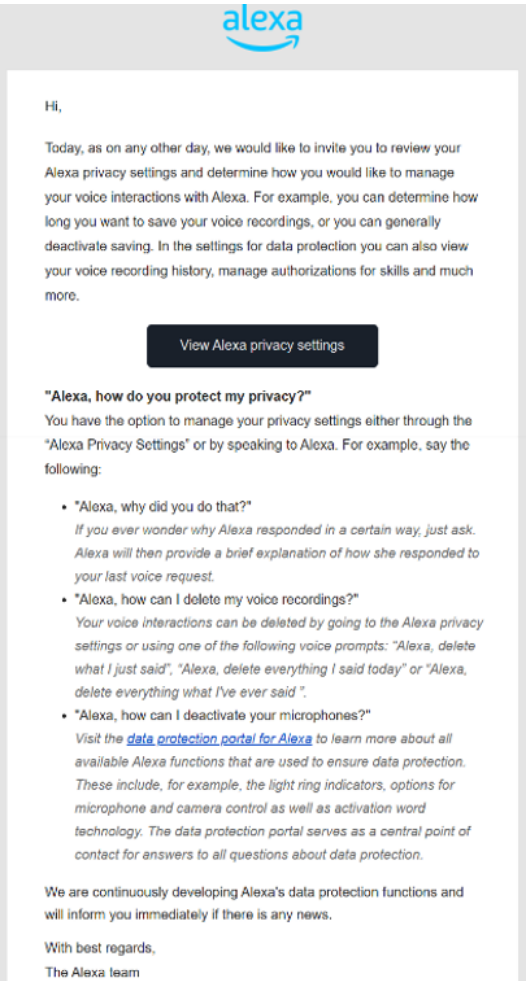


Fig. 9.3.: The email that Alexa users in Germany received from Amazon on February 2, 2021, after the new feature had been released (note that the original email was in German).

9.8.2 Appendix 2

Tab. 9.5.: Overview of the participants and their involvement in the project

ID #	Age	Gender	Co-inhabitants	VA Device	ESM	Co-design Workshops	Final Interview
P1	31	M	Partner	2 Alexa	Yes	Yes	Yes
P2	35	M	Wife & child	8 Alexa	Yes	Yes	Yes
P3	50	M	None	1 Alexa	No	Yes	No
P4	30	M	4 flatmates	2 Google Home	–	Yes	Yes
P5	32	F	2 flatmates	2 Alexa	Yes	Yes	Yes
P6	24	M	2 flatmates	2 Alexa	Yes	Yes	Yes
P7	28	M	Partner & child	5 Google Home	–	Yes	Yes
P8	26	F	None	1 Alexa	Yes	Yes	Yes
P9	56	M	Partner	5 Google Home	–	Yes	Yes
P10	35	M	Wife & 2 children	4 Alexa	Yes	Yes	Yes
P11	27	M	None	1 Google Home	–	Yes	Yes
P12	18	M	Parents & sister	2 Google Home	–	No	Yes
P13	45	M	3 flatmates	1 Alexa	Yes	Yes	Yes
P14	32	F	None	1 Alexa	No	Yes	No
P15	27	F	2 flatmates	1 Alexa	Yes	Yes	Yes
P16	27	M	Partner	2 Alexa	Yes	No	Yes
P17	34	M	Wife & child	4 Alexa	No	Yes	Yes
P18	23	M	Partner	3 Alexa	Yes	No	Yes
P19	40	F	Partner	1 Google Home	–	Yes	Yes

Part VI

Research Outcome

10

Discussion and Implications

Traveler, your footsteps are the road, nothing more;
traveler, there is no path,
the path is made by walking.
By walking, you make the road,
and turning to look behind
you see the path you never again will step upon.— Antonio Machado

In this dissertation, I investigated how people perceive and make sense of AI in their daily encounters across various application domains as these technologies become integrated into their lives. This chapter aims to synthesize the findings from the different studies conducted in this dissertation to address the research questions posed in this work:

1. RQ1) How do users make sense of AI in their everyday life encounters?
2. RQ2) What are the needs for empowerment that emerge through users' everyday encounters with AI?
3. RQ3) In what ways can we apply design to empower users' in everyday life encounters with AI?

To answer RQ1, Section 10.1 describes the findings and insights from the studies through the key components of sensemaking according to Weick's framework (Weick and Weick, 1995) — meaning ascription and enactment— and summarizes some distinctive features of sensemaking of AI systems.

Section 10.2 addresses the second research question by applying Zimmerman's theory of psychological empowerment (Zimmerman, 1995) to discuss users' empowerment needs, focusing on three dimensions—*feeling*, *knowing*, and *doing*—and how these needs manifest across different AI application domains described in this work.

Section 10.3 builds on the findings to suggest design implications for cultivating sensemaking by re-framing it as an interactional element within situated contexts and by conceptualizing users as active AI sensemakers. This section concludes with Section 10.4, which discusses how we can apply prospective sensemaking in the context of AI to handle the uncertainty it presents.

Ultimately, this dissertation invites us to embrace the unknown not as a void, but as an opportunity for meaning making in our daily encounters with AI, encouraging resilience and adaptability in the face of an ever-evolving digital landscape.

10.1 Sensemaking in Everyday life Encounters with AI (RQ1)

Perception is more than just a passive process of receiving information—it's how we actively interpret and make sense of the world around us (Norwich, 1982). Shaped by prior knowledge, experiences, and cultural narratives,

perception helps us turn sensory inputs into meaningful insights. When it comes to AI, perception isn't just about recognizing machines or algorithms; it's about how we comprehend their role in society, their capabilities, and their implications for the future.

My findings indicate that on a broader abstract level, users' perceptions of AI—what I term *perceived AI*—are shaped by both scientific and technical framings (e.g., AI mimicking human intelligence or recognizing patterns) and cultural narratives, which often portray AI as either a superpower or a potential threat (e.g., Cohen, 2024; Gáti, 2023). These perceptions are dynamic, broad, and not necessarily influenced by emerging technical trends. For example, users may categorize even simple automated, sensor-based devices as AI, illustrating the flexible and expansive nature of their sense-making process. This interplay of technical definitions, cultural narratives, and individual interpretations is central to how users construct their understanding of AI. As users encounter AI in diverse contexts, they integrate prior knowledge, external narratives, and firsthand experiences to form their own nuanced definitions of what AI is.

Another notable observation is how users differentiate between AI's competencies, performances, and behaviors, and the physical entities in which AI might reside. They separate the concept of AI from tangible machines, understanding that AI is not confined to a physical device but rather emerges through its underlying code—though even the program alone may not always be perceived as AI. This aligns with Schank's decades-old observation of users asking, "Where is the AI?" (Schank, 1987). Users continue to grapple with pinpointing the exact "location" of AI, which is reflected in the broad and

often vague definitions they offer, such as "AI is something, a thing, or an entity."

The following sections will explore the RQ1 in more detail, using Weick's framework of sensemaking, focusing on two the components of meaning ascription and enactment.

10.1.1 Sensemaking of AI through Meaning Ascription, and Definitions

Meaning ascription is central to Weick's concept of sensemaking, described as a social process in which individuals extract cues from their environment and construct shared meaning through cycles of interpretation and action (Weick, Sutcliffe, and Obstfeld, 2005). In Part II and Part III I examined AI sensemaking through the lens of meaning ascription and user definitions. This section builds on these findings, extending prior research by applying the characteristics of Weick's framework to AI and identifying distinctive features unique to AI sensemaking.

1. *Sensemaking of AI definition is shaped by interconnected folk concepts.*

   In Chapter 4, I identified six folk concepts that form the foundation of users' expectations for their present and future interactions with AI. These include: 1) AI can support humans by fulfilling a pre-programmed task, 2) AI can work independently and make decisions for humans, 3) AI can act and react like humans, 4) AI can simulate human competencies, 5) AI can improve human intelligence, and finally 6) AI can learn and develop by itself. By introducing the term *folk*

concepts in relation to AI, I extend the existing research on users' folk theorization of AI (e.g., DeVito, Gergle, and Birnholtz, 2017; Eslami, Karahalios, et al., 2016; Eslami, Rickman, et al., 2015) and propose that these concepts serve as the foundational building blocks for such theories.

In the social sciences, the distinction between concepts and theories is fundamental to understanding how knowledge is structured and developed. Concepts are the basic building blocks of theory; they are abstract ideas or general notions that represent phenomena within a field of study (Gerring, 1999). Theories, on the other hand, are more complex structures that explain relationships between concepts. They provide systematic explanations of how and why certain phenomena occur, often involving hypotheses that can be tested through empirical research (Schutz, 1970). Theories are thus constructed upon concepts, using them to build models that predict and interpret social phenomena.

Unlike folk *theories* of AI, which tend to be more reflective, systematic, and internally consistent (Eslami, Rickman, et al., 2015), folk *concepts* of AI tend to be more intuitive and less developed. They are not equivalent to scientific concepts but instead function as *sensitizing concepts*—offering individuals a flexible framework for interpreting and discussing AI. Folk concepts act as the foundation for folk theories of AI, helping us understand how users make sense of AI and how they construct more elaborate folk theories. Where folk theories of AI aim to explain specific algorithmic processes or system behaviors (DeVito, Gergle, and Birnholtz, 2017), folk concepts of AI focus more

broadly on users' assumptions and expectations about AI's capabilities and limitations, which ultimately shape those specific theories.

By comparing various folk concepts related to AI, I demonstrated that while users' definitions and perceptions of AI vary, there is a common thread in how they conceptualize it. Nearly all folk concepts frame AI in relation to human characteristics. For instance, users describe AI as mimicking human behavior, simulating human intelligence, or being created by humans to assist or replace certain tasks. These human-centric views of AI align with both early scientific definitions, such as Turing's framing of AI in terms of human intelligence and behavior (Turing, 2009) and media and fictional narratives surrounding AI (e.g., Bartneck, 2013). This alignment is also reflected in research on the anthropomorphism of AI, which highlights the public's tendency to attribute human-like traits and motivations to AI systems (e.g., Brauner, Hick, et al., 2023; Deshpande, Rajpurohit, et al., 2023). This suggests that despite individual variations in how users perceive AI, these perceptions are interconnected through a persistent human-centered lens. Folk concepts of AI, therefore, not only reflect users' intuitive sensemaking but also highlight the cultural and historical narratives that shape these perceptions.

2. *Sensemaking of AI's potential harm is influenced by overarching folk theories that extend beyond specific system behaviors.*

   In chapter 7 I discussed the *conflict-of-interest theory* regarding Tinder's algorithms, which posits that while these algorithms can identify compatible matches, they deliberately withhold or disrupt them to retain a paying user base. This theory extends beyond typical folk theories,

which are often informal, unscientific beliefs users develop to explain the *specific behaviors* of complex systems (DeVito, Gergle, and Birnholtz, 2017). Instead, the conflict-of-interest theory provides an overarching explanation, linking various folk theories about algorithmic behaviors under a common framework.

Users hold diverse beliefs about how Tinder's algorithms function—ranging from recommending profiles based on romantic compatibility and user feedback (Abel, Pei, et al., 2022), to utilizing desirability scores or randomness in matching (S. A. Huang, Hancock, and Tong, 2022; Nader and Lee, 2022) —the *conflict-of-interest theory* acts as an umbrella, uniting these differing perspectives by attributing the platform's outcomes to its financial motivations. This theory suggests that algorithmic behaviors, regardless of their specifics, are ultimately shaped by Tinder's interest in maintaining an engaged and paying user base.

Additionally, the conflict-of-interest theory transcends being merely an *algorithmic* folk theory which explains how the algorithm functions. It also addresses users' perceptions of why they may feel unsuccessful in finding a partner. This broader explanation shifts responsibility for failure away from users' actions or desirability and onto the platform's broken *promise*—the expectation that Tinder will connect them with compatible partners. In other words, it frames the problem as "it's not me, it's the algorithm."

Central to the conflict-of-interest theory is the suspicion of economic motivation. Users hypothesize that Tinder's financial interests—such as fostering subscriptions to premium services—override its stated mission of facilitating connections. This leads to a belief that the platform inten-

tionally disrupts promising matches, aligning with broader scholarship on platform dynamics. In such cases, users' trust is influenced not just by technical transparency but by their perception of the platform's integrity and alignment with their goals (Gillespie, 2018).

Thus, overarching folk theories, such as the conflict-of-interest theory, go beyond technical and algorithmic aspects of AI systems. They are deeply intertwined with a platform's mission, values, and accountability, offering users a way to understand not just *how* these systems work but also *why* they may fail to align with their expectations and potentially cause harm.

3. *Sensemaking of AI's definition and Trust in AI are strongly influenced by defining human intelligence and interpersonal trust.*

   It is well-established that humans often interact with machines socially, much like they do with other people, as demonstrated by the Computers Are Social Actors (CASA) paradigm. For example, people greet computers even though such behavior is unnecessary for machines (Fogg and Nass, 1997; Nass, Steuer, and Tauber, 1994). We extend social behaviors like greeting to both humans and computers as if they were alike (Moon and Nass, 1996). But how do we assign meaning to human-centric concepts when applied to machines? For instance, how do we interpret concepts like intelligence and trust in the context of machines?

   When examining folk concepts in chapter 4, I not only observed frequent references to human traits—such as "made by humans," "simulating humans," "like humans," or "independent of humans."—but also a

notable trend in how intelligence is conceptualized. Metaphorically, AI in my first study the participants often discussed in terms of adolescent development—initially influenced by and oriented toward its creators (humans), but gradually claiming independence. This progression mirrors the stages of human cognitive development, where intelligence evolves over time (Commons and Miller, 2002). Thus, folk concepts regarding the emergence of machine intelligence seem to be influenced by this developmental trajectory of human intelligence.

A similar pattern emerged in how my study participants conceptualized trust in AI, as discussed in Part III. Both practitioners and decision subjects referenced factors commonly associated with interpersonal trust. In the social sciences, trust is frequently linked to the concept of actor vulnerability (Schafheitle, Weibel, and Möllering, 2023), as exemplified in Mayer, Davis, and Schoorman, 1995, definition: trust is "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party."

In the study on trust in AI-assisted decision-making, both decision subjects and practitioners highlighted vulnerability and positive expectations as central components for trust in AI to develop. While their interpretations of these concepts varied depending on their roles (see the subsequent discussion on stakeholder-dependent sensemaking of trust), it is reasonable to suggest that expectations about how trust in AI emerges are shaped by existing frameworks for establishing trust in interpersonal relationships.

4. *Sensemaking of trust in AI is stakeholder-dependent*

Previous studies on Human-AI trust have predominantly focused on the perspectives of direct users—those who interact directly with AI systems (Lai, Chen, et al., 2023). In part III I explored how AI practitioners involved in system design and deployment (such as developers and project managers) and decision subjects—individuals who do not interact with AI systems directly but are affected by their decisions—may hold differing perspectives on factors influencing trust. While both groups acknowledged the importance of positive expectations and vulnerability as foundational components of trust, their interpretations of these factors differed.

For instance, the concept of vulnerability varied based on how participants defined situations in which they felt exposed to harm or perceived themselves to be in a position of weakness. One participant noted that judging users based on physical appearance could constitute a form of vulnerability, highlighting emotional harm. This contrasts with other participants who referenced financial harm as an example. These findings suggest that vulnerability as a prerequisite for trust is context-dependent and influenced by the individual or community involved. Cases of algorithmic discrimination, such as biases against Black women (Buolamwini and Gebru, 2018), highlight how certain populations may be placed in a more vulnerable position than others.

Discrepancies were also observed in factors like AI transparency, AI literacy, and system interactivity. For example, AI practitioners emphasized the importance of explaining AI recommendations and ensuring transparency in system operations. In contrast, decision subjects ex-

pressed skepticism about whether such transparency or explanations would genuinely improve their understanding of the decisions being made.

These findings suggest that sensemaking and trust-building around AI are influenced by stakeholder roles and experiences. Consequently, an individualistic and cognitivist approach to sensemaking should be broadened to incorporate a sociological perspective. This shift can address the need for further investigation into the diverse viewpoints and experiences of various stakeholders involved in or impacted by AI systems.

5. *Sensemaking of AI and Trust in AI is strongly influenced by other stakeholders.*

   According to Blumer, human beings act toward things based on the meanings those things hold for them, and these meanings arise from social interaction with others and society (Blumer, 1986). Weick similarly emphasizes that sensemaking is inherently social, involving the exchange of cues and narratives among individuals. It is not just an individual cognitive process but one of social construction, where meanings are negotiated and co-created through discourse and interaction (Weick and Weick, 1995). This view is also reflected in Battarbee's notion of *co-experience* (Battarbee, 2003), which "takes place as experiences are created together or shared with others."

   Building on my findings from Parts II and III I argue that the sensemaking of AI is shaped not only by social interactions but also by a complex network of stakeholders. My analysis revealed that both decision sub-

jects and AI practitioners regarded trust in different stakeholder groups as a significant factor influencing their trust in the AI system.

For example, AI practitioners emphasized the importance of trust between users and the AI team responsible for developing the system. In contrast, decision subjects placed greater importance on their trust in the direct users of the system as a critical factor in fostering trust in the AI. These findings suggest that in a multi-stakeholder socio-technical system, trust in other stakeholders plays a key role in shaping trust in the AI system itself. However, the extent of this influence varies depending on each stakeholder's position and role within the system.

These findings align with the concept of social transparency in AI systems (Ehsan, Liao, et al., 2021b), which shifts the focus from AI's internal workings to the broader social context of its use. Social transparency emphasizes understanding the history of other users' interactions with AI recommendations, acknowledging that trust and sensemaking are shaped by these interactions as much as by the system itself.

The focus on stakeholder-dependent trust in my findings complements social transparency, as both highlight the importance of situating AI systems within their social contexts. By making stakeholder interactions visible, social transparency can help bridge trust gaps, offering users insights into how AI systems are used and understood by others.

10.1.2 Sensemaking of AI and Enactment

Weick, Sutcliffe, and Obstfeld, 2005, emphasize that sensemaking is not just about interpreting existing data but also about enacting the environment

through actions that generate new data and insights. This process, known as "enactment," means that individuals and organizations actively shape their realities through their actions, which in turn influence their interpretations and subsequent actions (Weick, 1988). By engaging in action, people can test their hypotheses and adjust their understanding based on the outcomes, making sensemaking a dynamic and iterative process (Stevens and Pipek, 2018; Weick and Weick, 1995). This interplay between action and interpretation is essential for navigating complex and uncertain situations, as it allows for continuous learning and adaptation (Weick, 2009).

Folk theories, as a lens to capture users' sensemaking in algorithmic environments, have been shown to guide users' actions and interactions in these spaces. For instance, as discussed in chapter 6 , when users face algorithmic policing on social media—such as losing access to the platform without an explanation—folk theories about why the algorithm blocked or banned them guide their behavior. Users adjust by avoiding actions they perceive as "machine-like" from the algorithm's perspective. Similarly, as shown in chapter 5, folk theories guide users on dating platforms to push back against perceived algorithmic harms or processes they view as unfavorable—for example, by modifying their profiles or even moving to other platforms where they feel less scrutinized.

In this section, I explore *enactment* as a key component of AI sensemaking and contextualize it within the broader literature on human sensemaking. By examining how individuals act on their interpretations of AI and adjust their actions based on the outcomes, I provide insights into the unique characteristics of AI sensemaking and how it aligns with, or diverges from, traditional sensemaking theories.

1. *Sensemaking of AI is triggered by unexpected behaviors or unfavorable outcomes.*

   Weick, Sutcliffe, and Obstfeld, 2005, argue that explicit efforts at sensemaking tend to arise when the current state of the world deviates from expectations or when there is no clear way to engage with the situation—essentially, when the flow of action becomes unintelligible. In response to such disruptions, people first seek reasons that allow them to resume interrupted activity and stay in action.

   Chapters 8 and 9 illustrate a similar pattern in the sensemaking of AI systems. When users encountered unexpected behavior from VAs, they often asked the device for explanations, reflecting Weick's concept of seeking reasons to stay in action. Similarly, when confronted with data collected by VAs, users explicitly sought to make sense of unknown commands or unintended triggers (Part V). However, unlike Weick's argument that these *reasons*are often drawn from frameworks like institutional constraints, organizational premises, or traditions (Weick, Sutcliffe, and Obstfeld, 2005), in the context of AI systems, users expected the device itself to provide accounts or reasoning during the interaction. This reasoning extended beyond understanding the immediate disruption to offering guidance on how to respond or prevent similar situations in the future.

   This pattern of sensemaking triggered by unexpected or unfavorable outcomes also emerged in other contexts, such as users banned from social media or Tinder users experiencing unfavorable outcomes (Part IV). In these instances, users were left without a clear understanding of why the unfavorable event occurred or how to proceed. In the following

section, I will discuss how these sensemaking efforts may not only drive subsequent actions but also result in counteractions against AI systems, potentially impacting user engagement and trust.

2. *Sensemaking of AI is about plausibility which potentially creates distrust.*

   Sensemaking is not about uncovering objective truth, but rather about constructing plausible explanations that help people navigate their experiences (Weick and Weick, 1995). Plausibility, rather than accuracy, drives much of human sensemaking, as people continuously refine their understanding to incorporate new information and maintain resilience against uncertainty or criticism.

   In the context of AI, users' folk theories similarly focus on plausible explanations that fit their observations, even if these explanations are not entirely accurate (Karizat, Delmonaco, et al., 2021). As discussed in chapter 7 Tinder users may attribute missing messages or lack of matches to algorithmic sabotage, when in reality, these may be due to user-driven actions like unmatching or account deletions for violating the platform's policies. Tinder, like many AI systems, uses automated decision-making and moderation tools that may explain these outcomes (Tinderpress, 2022b) but, without transparency, users default to perhaps more plausible but possibly inaccurate explanations.

   Previous research (e.g., Mezias and Starbuck, 2003; Winter, 2003) has argued that plausibility generally benefits sensemaking by encouraging action and fostering learning. However, my findings suggest that this is not always the case with AI. Folk theories such as the *conflict-of-interest theory* which frames users' interests as being at odds with the platform's

goals, can undermine trust in the platform and prompt counteractions with potentially negative consequences. The next section will examine these counteractions in greater detail.

3. *Sensemaking of AI involves not only actions but also counteractions in response to perceived algorithmic behaviors.*

   The enactment process in sensemaking involves two steps: individuals first bracket portions of their experience for closer attention based on preconceptions and then act within the context of these bracketed elements. Through their actions, they shape the environment in ways that tend to confirm these preconceptions (W. Powers, 1973). In this process, action not only responds to the environment but also creates the conditions for further interpretation, shaping the sensemaking process itself (Weick, Sutcliffe, and Obstfeld, 2005).

   In the context of unfavorable algorithmic behavior, my findings reveal that users often go beyond merely understanding or adapting to the system. Instead, they engage in counteractions—behaviors aimed at disrupting or opposing perceived algorithmic processes. These actions, driven by uncertainty and opaque system behavior, align with Weick's view that sensemaking is an active process where individuals test assumptions and attempt to regain control (Weick and Weick, 1995). However, my analysis extends this framework by demonstrating how such enactments can lead to unintended consequences that affect both users and others in the system.

   For example, as shown in Chapter 7 and Nader and Lee, 2022, users employed countermeasures like profile manipulation, disliking attractive

profiles, or blocking users without cause to combat perceived algorithmic harms. While these actions were intended to counteract the system, they could mislead the algorithm, leading to unintended outcomes such as wrongful suspensions or bans of other users.

These findings also connect to the concept of *latent harm* (as described by Walker and DeVito, 2020), which focuses on the long-term impact of coping strategies that limit access to resources or support. My analysis suggests that users' counteractions can go beyond personal harm to cause broader interpersonal harm within the user base. For instance, counteractive behaviors may inadvertently restrict others' access to platforms like Tinder, creating cascading effects that reflect the broader, unintended consequences of such actions.

By building on Weick's framework, I argue that sensemaking through enactment not only clarifies the environment but also actively reshapes it, often in ways that complicate the experience for others.

10.2 Sensemaking of AI and User Empowerment (RQ2)

Sensemaking processes inherently involve interpreting and navigating the dynamics of influence and control within organizations and systems (Schildt, Mantere, and Cornelissen, 2020). In the context of system design, Dervin argues that sensemaking should attend to users' struggles, constraints, and barriers by fostering spaces for dissent and integrating deviant observations into the discourse (Dervin and Naumer, 2009). She highlights the need to

address entrenched power structures and balance agreement with contest, ensuring that diverse perspectives are considered.

To address the RQ2, I build on this perspective by applying (Zimmerman, 1995) —spanning the dimensions of feeling, knowing, and acting—in my studies of sensemaking to explore users' struggles, constraints, and barriers in interactions with AI systems. Doing so, I extend this framework by showing that sensemaking in AI involves addressing both **episodic** and **systemic** power. Episodic power emerges in specific interactions with AI, such as when algorithms make decisions about access or content moderation. Systemic power, however, operates more subtly, shaping user experiences and perceptions through institutional norms embedded in algorithmic design, opaque decision-making processes, and the broader socio-technical systems that reinforce existing inequalities.

By recognizing this dual nature of power, my research underscores the importance of designing AI systems that empower users not only by enabling agency in individual interactions but also by addressing systemic constraints that limit their ability to navigate and shape their experiences meaningfully.

10.2.1 Empowering Users by Addressing Emotional Needs

In this work, I examined various encounters with AI systems in which users felt disempowered. I showed that in such situations, users respond emotionally and need the system to acknowledge and account for the emotions they are experiencing. These emotional responses often shape how users interact with and perceive the system. For example, Tinder users experiencing a decline in matches and likes may suffer from self-esteem issues, leading

them to develop folk theories attributing their dating failures to algorithmic sabotage rather than personal factors. Similarly, when social media users are blocked due to invisible AI policing, they often interpret the algorithmic action emotionally, feeling punished or excluded (Register, Qin, et al., 2023). These reactions highlight the **episodic power** of AI, where specific system actions directly affect users' sense of agency and control. However, they also reflect **systemic power**, as the design of algorithms often embeds and reinforces norms, such as defining desirability or perpetuating narrow understandings of the human body.

A more direct example of users' need for systems to address emotional responses is explored in chapter 9, where a newly added feature in VAs aimed at addressing unexpected user behavior—such as accidental triggers or unrecognized commands—initially caused frustration. Users struggled with understanding how their feedback on the feature would be processed, leading to feelings of uncertainty and disempowerment. Interestingly, users were more willing to engage with this feature when they believed their data would contribute to meaningful system improvements in the future. This demonstrates how **episodic power** manifests in specific system interactions that directly impact users' trust and willingness to engage. At the same time, it highlights **systemic power** in the lack of transparency regarding how user feedback is utilized, which creates structural barriers to users' empowerment by reinforcing uncertainty and limiting their sense of control.

The need for AI systems to address emotional responses aligns with research on affective computing, which emphasizes that systems should respond to users' emotional contexts to enhance their experience (e.g., Calvo and Peters, 2014; Picard, 2000). Failing to do so can undermine users' trust (Binns,

Van Kleek, et al., 2018) and sense of empowerment, as feelings of frustration, self-doubt, or punishment often stem not only from episodic encounters but also from systemic constraints embedded in the design and operation of AI system.

10.2.2 Empowering Users Through Knowledge and Understanding

knowing is a complex concept and should not be confused with simply providing transparency. Schneider suggests that empowering users requires systems designed around knowing to help users acquire new skills to increase their chances for achieving their goals (Schneider, Eiband, et al., 2018).

In my research, I demonstrated that users often struggle to gather the information and skills necessary to achieve their goals, such as being matched with compatible partners. Many created fake accounts to learn about their comparators and optimized their profiles by adding pictures or adjusting filters. Similarly, users confronted with invisible AI policing on social media sought skills to solve CAPTCHAs and avoid behaviors perceived as "machine-like."

This behavior is not limited to direct users of AI systems but extends to those indirectly affected. For instance, job applicants tried to optimize their CVs to align with assumed algorithmic criteria, making their applications "machine-readable." These examples reflect the **episodic power**, where users adapt their strategies to overcome immediate obstacles.

I argue that platforms need not reveal the intricate workings of their algorithms (i.e., their "secret sauce"). Transparency without intelligibility, as seen

in the context of voice assistants, can overwhelm and confuse users. Simply revealing interaction data without context increases the need for guidance and can lead to unintended outcomes (Diakopoulos, 2016).

Instead, AI systems could empower users by facilitating skill-building through actionable support. For example, a dating app could provide recommendations on improving users' chances of securing meaningful interactions, while voice assistants could offer contextual information or visual tools to help users map data to their specific situations.

These examples illustrate how users' episodic efforts to adapt intersect with the systemic power structures embedded in AI systems. This underscores the importance of actionable design interventions that bridge the gap between transparency and intelligibility, ultimately empowering users to navigate these systems effectively.

10.2.3 Empowering Users in their Actions

Schneider describes a system that empowers users by focusing on *doing*, enabling users to complete tasks more efficiently through automation (Schneider, Eiband, et al., 2018). However, automation introduces challenges related to the "automation paradox," where increased reliance on automated systems can decrease user involvement, leading to a loss of situational awareness and control (Bainbridge, 1983).

This concern is particularly evident in systems employing invisible AI policing (Chapeter 6) where I argued for the inclusion of meaningful user controls to address the **episodic power** users seek in immediate interactions. These controls include: 1) ensuring users understand they are interacting with

an AI system, 2) enabling them to express frustrations and contest results, and 3) providing access to information about how their data, including biometric data, is handled. While the EU AI Act grants users rights regarding AI interaction and data use, it provides limited mechanisms for contesting AI-driven outcomes (the Future of Life Institute (FLI), 2024). For example, as discussed in Chapter 7 Tinder users expressed frustration at repeatedly encountering the same profiles but lacked meaningful tools to address the issue, often resorting to blocking or banning others as their only recourse. Similarly, blocked users on social media have described appeal processes as futile attempts to "shout into the void" with no response from the platform (Myers West, 2018).

Automation also poses challenges for indirect stakeholders who lack opportunities to engage with the system but are affected by its decisions. In loan applications, for instance, applicants had little chance to make sense of their outcomes, often unaware of the data fed into the system by employees. This contrasts with contexts like job recruitment, where users can exercise episodic power by tailoring their inputs—such as resumes—to align with the system's perceived expectations (Mittelstadt, Allo, et al., 2016).

Empowerment, therefore, must extend beyond direct users to include all stakeholders affected by AI systems. This need is particularly urgent in contexts where **systemic power** limits stakeholder agency, as observed in loan applications, compared to scenarios like job recruitment, where users have greater influence over outcomes.

The importance of empowering users in their actions became clearer in Chapter 9 where users of VAs benefited from explanations that progressively guided them with actionable advice. These explanations helped users take

informed actions to achieve their goals, aligning with research on decision support systems that highlights the value of clear recommendations and actionable insights (S. Lee, Kim, and Sundar, 2015).

Empowerment in sensemaking emerges when systems respond to users' perspectives and needs rather than imposing expert-driven frameworks (Naumer, Fisher, and Dervin, 2008). By addressing users' emotional, cognitive, and behavioral dimensions—*feeling*, *knowing*, and *doing*—AI systems can enhance users' sense of control and agency. This dual focus on addressing episodic challenges while mitigating systemic constraints creates a more empowering experience for users and other stakeholders alike. In the next subsection, I will build on these insights to explore implications for designing systems that better support these needs.

10.3 Designing for Sensemaking and User Empowerment (RQ3)

Sensemaking is not merely a cognitive process but an inherently interactional one, shaped by the dynamics of participation and action. As a research area, sensemaking emphasizes the creation of meaningful, context-specific practices, moving away from the top-down imposition of expertise on users (Naumer, Fisher, and Dervin, 2008). Instead, sensemaking involves continuous interaction, where individuals actively shape their understanding through engagement with both their environment and the system. In this sense, sensemaking is an evolving process that emerges through interactions,

with users enacting their way into meaning by balancing structure and agency (Turner, Allen, et al., 2023).

My research highlights two key implications for designing AI systems that better support sensemaking. First, by drawing on Social Practice Theory, it frames sensemaking of AI as an interactional feature that should be prioritized in design. This can be achieved by fostering users' competencies in their specific contexts, enabling them to better navigate and make sense of AI systems in their everyday situations. Second, it emphasizes the importance of diversifying user participation in the design process. By involving users more actively, designers can empower them to engage in sensemaking, sense-unmaking, sense-giving, and sense-breaking processes, thereby fostering a more collaborative and inclusive approach to AI design.

Finally, I extend Weick's framework of sensemaking to include **prospective sensemaking** in the context of AI. This approach offers a method for designing systems that address the uncertainty inherent in AI technologies, helping users anticipate, adapt to, and shape the evolving nature of their interactions with these systems.

10.3.1 Framing AI Sensemaking as an Interactional Element in Practice

By exploring different AI systems in users' everyday lives, I demonstrated that AI systems are embedded within a socio-technical network of users' social settings and practices. *Social practice theory (SPT)* views human activity as

embedded in routines shaped by three interconnected elements (Reckwitz, 2003):

- Materials: The physical and technological artifacts involved in a practice. In the context of AI, this includes algorithms, interfaces, and devices voice assistants (Maller, 2015).
- Meanings: The shared understandings, norms, and cultural narratives underpinning a practice. In the context of AI, these are shaped by users' perceptions of AI's capabilities and roles, as well as broader societal narratives around AI (Durey, Ward, et al., 2024).
- Competencies: The skills and knowledge required to perform practices effectively. In the context of AI, this involves users understanding system feedback, adapting behaviors, and engaging in trial-and-error experimentation (Durey, Ward, et al., 2024).

Applying SPT to AI design emphasizes the **situatedness of interactions** and highlights the critical role of competencies in connecting materials (the AI systems) and meanings (sensemaking). Competencies should be regarded as an integral part of practice, enabling users to engage meaningfully with AI systems. At the same time, AI systems (materials) should support and scaffold these competencies to facilitate sensemaking in everyday contexts. Based on this approach, some design recommendations from my research could be:

- Skill Development: In dating apps, when users experience decreased matches or likes, algorithms could suggest profile enhancements to improve match potential (e.g., "Adding a headshot improves matches by X%").

- Experimentation: In CAPTCHA tests or social media interactions, systems should allow users some degree of trial and error before imposing harsh consequences like permanent blocking. This approach helps users make sense of the system's functionality and addresses sense-unmaking caused by false assumptions.
- Contextual Feedback: For unexpected behavior in voice assistants, actionable, context-aware guidance—such as dialogic feedback explaining why an action occurred and how to avoid similar outcomes—supports sensemaking and aligns AI interactions with the user's specific context.

In sum, by drawing on *practice theory* and viewing AI sensemaking as an interactional feature emerging from engagement in specific situations, we can design technologies that facilitate deeper, more effective sensemaking. These technologies, when integrated into users' daily routines, hold the potential to foster empowerment by aligning with the situated and evolving nature of human practices.

10.3.2 Diversifying Participation for Empowered Sensemaking

As discussed earlier (see Chapter 2) the sensemaking methodology conceptualizes users as "theorists and knowledge-makers in their worlds" (Dervin, 2015; Naumer, Fisher, and Dervin, 2008; Stevens, 2009), who struggle to make sense of an incomplete reality rather than simply seeking an objective "reality" (Naumer, Fisher, and Dervin, 2008). In the context of AI, this perspective suggests that users are not passive recipients of information but

actively shape their understanding of AI systems based on their engagement with them.

Participation in the design process is a critical avenue for addressing the power structures inherent in sensemaking. Participation determines who has the authority to define, disrupt (sense-breaking), or contribute (sense-giving) to shared meanings in socio-technical systems.

In my work, I explored users' engagement with VAs as sensemakers, leveraging two design frameworks that operate on different levels of participation. These approaches illuminated the evolving relationships users have with AI and the power dynamics embedded within these interactions.

- RtD: Sense-Breaking and Self-Discovery: In an RtD approach, I examined the effects of visualizing the data collected by VAs for users. When exposed to the extent and nature of the data collected, users often experienced a sense of lack of control and disempowerment, particularly regarding their privacy. This reaction highlights the limitations of transparency-based design; simply showing users the inner workings of a system does not necessarily lead to comprehension or empowerment (Ylirisku, Halttunen, et al., 2009).

  However, this intervention also enabled users to engage in practices of self-discovery and reflection. As users observed patterns in their routines and habits through the VA-collected data, they began to connect these insights to their identity, aligning with Weick's assertion that sensemaking links ambiguous stimuli to identity needs (Weick and Weick, 1995). For instance, users started reflecting on their behavioral patterns and daily routines, reframing their relationship with the VA

as a tool for introspection. This sense-breaking—where users were forced to confront assumptions about the VA—became a catalyst for sense-giving, where they redefined their relationship with the system in ways that resonated with their self-concept.

- Co-Design: Sense-Giving and Shared Accountability: In the co-design process, I explored a participatory approach where users treated VAs as conversational partners. Agency in this context was not inherent to the system but emerged through communicative acts between users and the VA (Lind and Dickel, 2024). The collaborative solutions users co-created, such as personalized recommendations or instructional designs, fostered a sense of *shared accountability*.

  Shared accountability is a dynamic in which users perceive responsibility not only in their own actions but also in how the system operates. Through co-design, users actively shaped the AI's outputs, establishing a relationship with the VA that aligned more closely with a counterpart or companion rather than a subservient tool (Q. V. Liao, Wang, et al., 2019; Sayago, 2020). This process empowered users to engage in sense-giving—imposing their interpretations and preferences onto the AI system—and even sense-breaking when they challenged its outputs or their own assumptions.

Based on these findings, I propose that designing human-centered AI requires diversifying the roles users play in the design process. Participation strategies should not only aim to engage users as testers or evaluators but also as co-creators and reflectors who can redefine their own relationships with AI. Different levels of participation foster distinct relational stances toward AI

systems, from introspective tools for self-discovery to collaborative partners that share accountability.

Such diversification captures the nuances of users' relational positions, empowering them to actively engage in sensemaking, sense-unmaking, sense-giving, and sense-breaking processes. This, in turn, enables users to navigate and shape their own realities with AI, addressing both their identity needs and the broader power structures embedded in AI design.

10.4 Dealing with Uncertainty: Prospective AI Sensemaking (RQ3)

I began this work with the question of how we handle the growing uncertainty posed by AI and whether we can leverage the uncertainty to prepare for an unforeseeable future with this technology. I then applied sensemaking as a framework to explore how people fluidly navigate between states of *Certainty vs. Uncertainty*, shaping both their understanding of AI and their interactions with it (Weick, Sutcliffe, and Obstfeld, 2005).

Traditionally, sensemaking has been viewed as a retrospective process, where individuals make sense of their actions and experiences only after they've occurred. Weick's framework largely reinforces this retrospective focus, and much of the literature on sensemaking (e.g., Baran and Scott, 2010; Rantatalo, 2013; Teo, Lee, and Lim, 2017) aligns with this notion.

Technological advancements, particularly in AI, are accelerating at an unprecedented pace, embedding themselves deeply in our everyday lives and routines. The rapid turnover of devices and applications fosters a sense of

fleeting familiarity, as technologies quickly become integral only to fade, creating a shifting landscape that shapes our experiences and life stages (Alizadeh, Mniestri, Uhde, and Stevens, 2022). Beyond influencing personal identity, these technologies also impact collective identity, shaping shared experiences and societal norms. Studies show that 'smart' home technologies adapt to and shape users' routines, transforming the home from a static environment into a fluid, evolving space that adjusts with its inhabitants (Mennicken, Hwang, et al., 2015; Reddy, 2020; Taylor, Harper, et al., 2007). Similarly, with AI agents like ChatGPT, robots, and virtual avatars playing diverse roles in homes and other settings, technology has moved from being a passive backdrop to becoming an active, adaptive participant (Alizadeh, Randall, et al., 2024). Yet, this constant state of technological change also creates continuous uncertainty, where change itself becomes the only constant.

In such times of dynamic complexity, where yesterday's assumptions quickly become obsolete, a retrospective approach to sensemaking may no longer be sufficient. As Barton, Sutcliffe, et al., 2015, argue, a proactive, future-oriented approach is essential, especially for leaders managing and guiding people through uncertain contexts. This proactive stance on sensemaking, or {prospective sensemaking}, is rooted in responding to present cues and taking actions that prepare us for emerging possibilities.

To move toward prospective sensemaking, I draw on Colville et al.'s concept of *simplexity* (Colville, Brown, and Pye, 2012), which balances the complexity of adaptive thinking with the clarity of actionable responses. Rather than relying solely on past experiences, simplexity encourages openness to multiple future possibilities and uses present cues to anticipate and influence outcomes. Applying this framework to AI design involves focusing on two key principles:

the *quality of attention* (Tsoukas, 2017) and the *iterative process of learning and adapting* (Kolb, 2014).

Prospective sensemaking prioritizes a high *quality of attention*—not only to the unfolding situation but also to how users interpret and respond to their experiences. This requires AI systems to go beyond providing information or outputs and instead support users in engaging meaningfully with the *how* of their interactions: how they interpret feedback, adapt their behavior, and construct meaning in real-time (Colville, Pye, and Brown, 2016). Mindfulness (Langer and Ngnoumen, 2017) and imagination play critical roles here, enabling users to remain responsive to uncertainty and changes while exploring potential outcomes.

A prospective approach also emphasizes *continuous learning* over arriving at fixed answers. As AI systems evolve, they must foster an ongoing learning process for users, enabling them to experiment, observe outcomes, and iteratively refine their interactions. This iterative process aligns users' needs with the evolving functionality of the system, allowing for small, ongoing adjustments that enhance understanding and resilience in navigating uncertainty (Colville, Pye, and Brown, 2016). For AI to support prospective sensemaking effectively, its design must embrace *adaptability.* Systems should not "design" or rigidly shape users but instead co-evolve with them, facilitating mutual adaptation. This could involve designing AI systems that allow users to experiment safely, observe outcomes, and iteratively refine their interactions. By embracing a framework of continuous adaptation, AI systems can support users in navigating unfamiliar or unprecedented events, ultimately fostering empowerment and resilience in an era of rapid technological change.

In summary, this work emphasizes that the process of sensemaking—the meaningful interpretation of technologies—can no longer remain solely reliant on retrospective interpretations, in which users only understand how and why specific actions were performed *after* the fact. Instead, a prospective orientation is needed: users must also be empowered to anticipate future developments and shape their own roles in interactions with technology.

The design of AI systems, and the support for their appropriation (Stevens and Pipek, 2018; Stevens, Pipek, and Wulf, 2009), must enable a **dialectic** between retrospective and prospective sensemaking. Only by embracing this balance can human-centered AI empower users to navigate uncertainty, adapt to change, and play an active role in shaping the evolution and impact of AI systems.

11

Conclusion

11.1 Summary of the Thesis

This dissertation uses sensemaking as a framework to explore how people interact with AI in their daily lives and the empowerment needs that come up along the way. By framing AI as a socio-technical concept, I look at how its definition has evolved and how users perceive it in different contexts.

Part I lays the groundwork for the research with two chapters. Chapter 2 dives into definitions and perceptions of AI, reviews sensemaking theories (with a focus on Weick's work), and explores how people make sense of AI in three areas: AI-assisted decision-making, AI-mediated social platforms, and agentic AI technologies. It also covers approaches to user empowerment in HCI, including Zimmerman's holistic framework, and looks at how these ideas can support sensemaking. Chapter 3 outlines the methods I used, including design case studies, co-design, and Research through Design, and gives an overview of the different research settings involved.

The middle sections—**Part II**, **Part III**, **Part IV**, and **Part V** —present a mix of conference articles that cover different case studies, each capturing sensemaking from various angles.

I start with Chapter 4 which uses symbolic interactionism to explore how people perceive and make sense of the term "AI". From there, I dig into specific application domains:

- **Part III** focuses on AI-assisted decision-making systems, looking at trust not just from the users' perspective but also from the point of view of decision subjects and AI practitioners. It provides a comparison of these perspectives.
- **Part IV** looks at AI-mediated social platforms. In Chapter 6, I explore how users deal with invisible AI policing and what happens when algorithms make mistakes. Chapter 7 takes Tinder as an example to show how users make sense of perceived harms and how these perceptions affect their interactions on the platform.
- **Part V** focuses on agentic AI technologies. Chapter 8 examines how users make sense of the data collected by Alexa, while Chapter 9 explores users' sensemaking of unexpected behavior from voice assistants through dialogic interactions.

Part VI wraps up with two chapters that cover the outcomes of my research. Chapter 10 discusses the findings in relation to the research questions, and the final Chapter 11 sums up the dissertation, highlights its main contributions, and acknowledges its limitations.

11.2 Contributions

The objectives and research questions of this thesis, as outlined earlier, build on and extend prior work in three key areas: AI sensemaking, user

empowerment, and designing for sensemaking. With these focus areas in mind, I've organized and summarized the contributions along these lines:

11.2.1 Sensemaking of AI

Understanding AI remains a challenge for many people, partly because there's no clear and accessible definition of it and partly due to media narratives that often misrepresent its reality. Studies show that many users still don't realize when they're interacting with AI (e.g., Eslami, Karahalios, et al., 2016; Long and Magerko, 2020; Pegasystems, 2019), which leads to frustration when these systems don't meet their expectations and makes it harder to engage with and use the technology effectively (Fast and Horvitz, 2017).

Researchers in the HCI community have started tackling these misconceptions by investigating how people make sense of AI (e.g., DeVito, Gergle, and Birnholtz, 2017; Druga, Williams, et al., 2017; Eslami, Vaccaro, et al., 2017). However, studies on mental models and folk theories related to AI are often fragmented, as they vary across different application domains (Alizadeh, Stevens, and Esau, 2021).

In this dissertation, I treat AI as a socio-technical construct and apply Weick's framework of sensemaking to explore the question: *RQ1) How do users make sense of AI in their everyday life encounters?* I investigate this across three different application domains that arise in daily life. A major contribution of this work is outlining the unique characteristics of AI sensemaking through a cross-comparison of contexts.

Weick's framework describes sensemaking as a process that is grounded in identity construction, retrospective, enactive, inherently social, focused on

extracted cues, and driven by plausibility rather than accuracy (Weick and Weick, 1995). My research shows that AI sensemaking shares several of these characteristics:

1. Enactive: AI sensemaking is an active process where users' actions shape their understanding of the system and, in turn, how the system behaves.
2. Driven by Plausibility: Users often settle for plausible explanations of how AI works rather than striving for a complete understanding.
3. Inherently Social: Sensemaking is shaped by social interactions, shared narratives, and societal discussions about AI.

At the same time, AI sensemaking also displays unique characteristics that extend Weick's framework:

1. Interactional and Situated – not just triggered by disruption, but embedded in ongoing, everyday encounters with AI systems.
2. Dynamic and Iterative – combining retrospective reflection with prospective adaptation as users navigate AI over time.
3. Epistemically Grounded in Folk Concepts – drawing on lived experiences, cultural narratives, and subjective interpretations that may be conflicting, evolving, or speculative.
4. Goal-Oriented toward Coping and Agency – helping users cope with uncertainty, adapt their roles, and, where possible, exercise power over the interaction.

5. Role-Diverse – users act not only as sense-makers but also as sense-givers, unmakers, and breakers, actively shaping and challenging the meanings of AI in their context.

This research also reveals important implications of these characteristics. For example, while plausibility is central to AI sensemaking, it can sometimes undermine trust in AI systems or inadvertently harm other users.

11.2.2 User Empowerment

Empowerment has become one main goal in the field of HCAI, reflected in both research and regulatory initiatives like the EU's AI Act (Shneiderman, 2022; the Future of Life Institute (FLI), 2024). The focus is on designing systems that amplify human capabilities, protect fundamental rights, and promote equitable outcomes (Usmani, Happonen, and Watada, 2023). Research also highlights the emerging risks AI poses, including algorithmic biases and harm (e.g., Andalibi, Pyle, et al., 2023a; Karizat, Delmonaco, et al., 2021). A key contribution of my work is applying *sensemaking* as a lens to examine how power operates in users' interactions with AI, specifically addressing both episodic and systemic power. I use this perspective to answer *RQ2) What are the needs for empowerment that emerge through users' everyday encounters with AI?*

Through this, I contribute to HCI research on user empowerment by using Zimmerman's holistic framework of empowerment (Zimmerman, 1995). My work categorizes users' needs for empowerment across AI application domains into three dimensions—feeling, knowing, and doing—while clar-

ifying how episodic and systemic power intersect within each dimension (Figure 11.1).

1. Feeling: AI systems should acknowledge and respond to users' emotional contexts, as negative emotions like frustration, self-doubt, and exclusion not only undermine trust and agency in specific interactions (episodic power) but also reinforce systemic norms and constraints that limit broader empowerment.

2. Knowing: AI systems should actively help users build the knowledge and skills they need to achieve their goals. There are two key dimensions to this: *episodic power* and *systemic power*. Episodic power comes into play when users face immediate barriers, like being blocked on social media without knowing the reason. Systemic power, on the other hand, involves the broader structural limitations that keep other potential stakeholders from understanding the outcomes. For instance, job applicants may be rejected without ever interacting with the system or knowing the criteria that led to that decision.

3. Doing: Empowerment isn't just about transparency—AI systems should enable users to take meaningful action through actionable feedback and effective controls. Episodic power involves addressing immediate challenges, such as contesting algorithmic decisions, while systemic power lies in ensuring users can act within complex, multi-stakeholder systems that often obscure pathways for real change.

It goes without saying that the different dimensions of empowerment are interconnected, but a key contribution of my work is showing why it's essential to address all of them to achieve meaningful user empowerment. By integrat-

ing these dimensions—and focusing on both episodic and systemic power—I offer a more nuanced perspective on empowerment. This emphasizes the importance of addressing both immediate user interactions and the broader systems that shape those experiences.

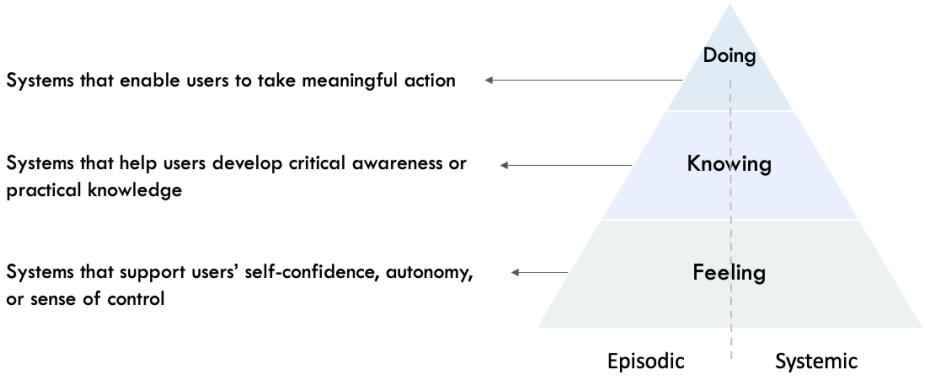


Fig. 11.1.: Adaptation of Zimmerman's (1995) framework of empowerment, illustrating three dimensions—feeling, knowing, and doing—and how they intersect with episodic and systemic power in AI sensemaking.

11.2.3 Designing for Sensemaking: Main Contribution to HCI

So far, HCI researchers have explored a variety of approaches to design for sensemaking and user empowerment. These range from enhancing users' direct interactions with existing systems—like accountability-focused design (e.g., Kianzad, Kim, et al., n.d.; Porcheron, Fischer, et al., 2018), which helps users trace AI decision-making processes, and explainability-focused design (e.g., Ehsan and Riedl, 2020; Hoffman, Mueller, et al., 2018), aimed at refining mental models and clarifying system behaviors—to transparency-focused design (e.g., Chalmers, MacColl, and Bell, 2003; Eslami, Karahalios, et al., 2016), which uses strategies such as seamful design to reveal the imperfec-

tions and limitations of AI systems. Broader approaches, like data-enabled design, combine data analytics and user-centered methods to improve how users interpret and interact with AI systems.

Addressing the question of *RQ3: In what ways can we apply design to empower users in their everyday life encounters with AI?*, the contribution of this dissertation lies at the intersection of these approaches.

First, by drawing on Social Practice Theory, it frames sensemaking as a situated, context-specific element that should be addressed as situations arise. This involves fostering users' competencies in their unique contexts, helping them navigate and understand AI systems in their daily interactions. For instance, providing opportunities for explanatory dialogues during moments of unexpected behavior in voice assistants can enhance users' ability to make sense of the situation.

Second, this work bridges different design approaches by highlighting the importance of diversifying user participation in the design process. For instance, voice assistant users were involved at two distinct levels: first, through research through design, where they experimented with new ways of interacting with data to make sense of the system's behavior; and second, through participatory design, where they took a more active role in shaping that behavior. These varying levels of involvement gave users opportunities to exercise power in different ways

The **main contribution to HCI** of this dissertation is the conceptualization of **four expanded user roles in AI sensemaking**—sense-maker, sense-unmaker, sense-giver, and sense-breaker—which capture the diverse ways users interpret, challenge, shape, and resist AI systems (Table 11.1). This framework

provides a structured lens for designing AI systems that recognize and support these roles, fostering a more collaborative and inclusive approach to AI design.

Tab. 11.1.: Expanding user roles in AI sensemaking: main contribution of this dissertation to HCI.

User Role	What It Means	Key Purpose
Sense-maker	The user interprets ambiguous or unexpected AI behavior (e.g., drops in matches).	To construct meaning, cope with uncertainty, and regain a sense of control or understanding.
Sense-unmaker	The user questions or lets go of previous assumptions.	To adapt to new information, break free from misinterpretations, and respond to system changes.
Sense-giver	The user (or system) influences how others (or themselves) interpret the AI.	To shape meanings, guide others, or align system behavior with personal or social expectations.
Sense-breaker	The user disrupts dominant narratives or resists AI system norms.	To challenge systemic power, assert agency, and redefine the user–system relationship.

11.3 Reflections and Closing Thoughts

This dissertation was written with an awareness of the ever-changing ways AI is defined and understood, both by experts and by everyday users. Over the course of my PhD, I often reflected on how "AI" seemed to be associated with different technologies depending on the time, the context, or the person using the term. What has remained constant, though, is the human need to make sense of it. This sensemaking isn't confined to researchers or designers but happens in everyday life, where users encounter AI in its many forms—whether they recognize it as AI or not.

In this dissertation, I treat AI as a socio-technical construct and an umbrella term to investigate how users interact with and understand it across three application domains. At the same time, I acknowledge the critical perspectives

on using the term AI itself, which some scholars argue has been inflated and overhyped in media and social discourse (e.g., Cave and Dihal, 2020). These critiques are valid, as they caution against the dangers of anthropomorphizing or mystifying technologies under the label of "AI." However, since this term exists and shapes public perceptions and interactions, I argue that it is important to study how people make sense of it. My goal is not to fix a singular definition of AI but to explore its evolving meanings and implications as they are experienced in practice.

I want to clarify my use of the term *users' misconceptions*. I realize it might sound contradictory, especially given my own stance on avoiding fixed definitions of AI. However, when I refer to *misconceptions*, I'm drawing on the perspectives of researchers, developers, and designers. These highlight a gap between their understanding of AI and how users perceive it. This gap matters because it can lead to frustration, unmet expectations, and a sense of disempowerment in users' interactions with AI systems.

Reflecting on the limitations of my work, I recognize that although this dissertation has design implications, most of the studies are not design-focused. Apart from one study in Chapter 8, which included a prototype, the others emphasize understanding existing systems and gathering insights through qualitative research. These studies adhere to the earlier stages of the Design Case Study framework, focusing on in-depth exploration rather than prototyping or implementation. While this may limit the direct applicability of the findings for designers working on specific AI systems—be it Tinder or Alexa—it does not diminish the importance of understanding the situated nature of sensemaking. I believe that exploring how users make sense of AI in their contexts, and attending to these dynamics in interaction design,

can inform design processes across different domains. Future research, with a narrower focus on specific design practices, could build on this work to provide actionable recommendations for particular industries.

This work also draws on the EU AI Act, referencing it as an example of ongoing efforts to regulate and define AI while advocating for user rights and empowerment. While I critique its lack of accessible definitions and practical mechanisms for implementing users' rights, I intentionally limited my scope to HCI, acknowledging that regulatory issues require interdisciplinary approaches that extend beyond this work. HCI alone cannot address the full spectrum of challenges posed by AI, but it can contribute by amplifying user voices, surfacing hidden practices, and proposing new ways of thinking about interaction and empowerment.

As I write this concluding thoughts, I'm struck by how much my own perspective has evolved over the course of this journey. Writing this dissertation has been a process of sensemaking in itself—trying to make sense of how AI shapes our world and how we, as researchers, can shape its development. I've come to believe that reflexivity is one of the most valuable tools we have, especially in the context of qualitative research. Reflexivity means not only questioning our assumptions and methods but also acknowledging the ethical dilemmas and power dynamics embedded in our work. For example, as I studied user interactions with AI systems, I became increasingly aware of how my own positionality—my background in AI, my personal encounters with these technologies, and even my biases as a researcher—influenced the questions I asked and the interpretations I made.

Reflexivity also means embracing uncertainty. AI is changing the way we create, share, and understand knowledge, and there are no clear answers

about how researchers should respond to these shifts. But I believe that staying reflexive—continuously questioning, adapting, and learning—will help us navigate these complexities in meaningful ways.

Looking ahead, I hope this dissertation serves as a starting point for others. To my fellow HCI researchers and those studying AI technologies, I hope this work encourages you to push the boundaries of how we understand and design for sensemaking. I hope it inspires you to ask new questions, to take risks, and to bring your own perspectives into the conversation. Most of all, I hope it shows that by involving users as active participants in the design process—by diversifying their roles and amplifying their voices—we can create systems that truly reflect the complexity and richness of everyday life encounters with new emerging AI technologies.

Bibliography

Abdi, N., Ramokapane, K. M., & Such, J. M. (2019). More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. *Proceedings of the Fifteenth Symposium on Usable Privacy and Security*, *15*, 17. https://www.usenix.org/system/files/soups2019-abdi.pdf (cit. on pp. 254, 255, 260).

Abel, C., Pei, L., Larson, I., Olgado, B. S., & Turner, B. (2022). "tinder will know you are a 6": Users' perceptions of algorithms on tinder. *Proceedings of the 55th Hawaii International Conference on System Sciences* (cit. on pp. 45, 199, 200, 229, 232, 331).

Acharya, A. S., Prakash, A., Saxena, P., & Nigam, A. (2013). Sampling: Why and how of it. *Indian journal of medical specialties*, *4*(2), 330–333 (cit. on p. 96).

Acien, A., Morales, A., Fierrez, J., Vera-Rodriguez, R., & Delgado-Mohatar, O. (2020). Becaptcha: Bot detection in smartphone interaction using touchscreen biometrics and mobile sensors. *arXiv preprint arXiv:2005.13655* (cit. on p. 182).

Acker, A., & Beaton, B. (2016). Software update unrest: The recent happenings around tinder and tesla. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1891–1900 (cit. on p. 198).

Adadi, A., & Berrada, M. (2018a). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 (cit. on pp. 38, 130).

Adadi, A., & Berrada, M. (2018b). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160 (cit. on pp. 83, 86).

Adams, A., & Sasse, M. A. (1999). Users are not the enemy. *Communications of the ACM*, *42*(12), 40–46 (cit. on pp. 61, 259).

Adams, B. D., Bruyn, L., Houde, S., & Angelopoulos, P. (2003, June). *Trust in automated systems literature review* (tech. rep.). Defence Research and Development Canada. Toronto, Ontario, Canada. https://cradpdf.drdc-rddc.gc.ca/PDFS/unc17/p520342.pdf (cit. on p. 132).

Agency, U. N. C. (2022, February). *More under 20s sexually assaulted after meeting offenders on dating sites*. National Crime Agency London, England. https://www.nationalcrimeagency.gov.uk/news/more-under-20s-sexually-assaulted-after-meeting-offenders-on-dating-sites (cit. on pp. 190, 193).

Ahmadi, M. (2023). Feminist living labs as research infrastructures for hci: A socio-informatics approach. *OPUS Siegen* (cit. on p. 17).

Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: Predicting and preventing disparate impact. *Available at SSRN*, *2746078*, 29 (cit. on pp. 38, 126, 131).

Akyon, F. C., & Kalfaoglu, M. E. (2019). Instagram fake and automated account detection. *2019 Innovations in intelligent systems and applications conference (ASYU)*, 1–7 (cit. on p. 41).

Albury, K., Byron, P., McCosker, A., Pym, T., Walshe, J., Race, K., Salon, D., Wark, T., Botfield, J., Reeders, D., & Dietzel, C. (2019). *Safety, risk and wellbeing on dating apps: Final report*. Melbourne: Swinburne University of Technology. doi. https://doi.org/10.25916/5dd324c1b33bb (cit. on pp. 193, 195, 236, 240).

Ali, S. J., Christin, A., Smart, A., & Katila, R. (2023). Walking the walk of ai ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 217–226. https://doi.org/10.1145/3593013.3593990 (cit. on p. 135).

Alizadeh, F., Jakobi, T., Boldt, J., & Stevens, G. (2019). GDPR-Reality Check on the Right to Access Data: Claiming and Investigating Personally Identifiable Data from Companies. *Proceedings of Mensch und Computer 2019 on - MuC'19*, 811–814. https://doi.org/10.1145/3340764.3344913 (cit. on pp. 60, 255, 256).

Alizadeh, F., Mniestri, A., & Stevens, G. (2022a). Does anyone dream of invisible ai? a critique of the making invisible of ai policing. *Nordic Human-Computer Interaction Conference*, 1–6 (cit. on p. 294).

Alizadeh, F., Mniestri, A., & Stevens, G. (2022b). The reverse turing test: Being human (is) enough in the age of ai. *Proceedings http://ceur-ws. org ISSN*, *1613*(2022), 0073 (cit. on pp. 42, 43, 179).

Alizadeh, F., Mniestri, A., Uhde, A., & Stevens, G. (2022). On appropriation and nostalgic reminiscence of technology. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–6 (cit. on p. 354).

Alizadeh, F., Pins, D., & Stevens, G. (2023). User-friendly explanatory dialogues (cit. on pp. 294, 311).

Alizadeh, F., Randall, D., Tolmie, P., Lee, M., Xu, Y., Mennicken, S., Woźniak, M. P., Paul, D., & Pins, D. (2024). Future of home-living: Designing smart spaces for modern domestic life. *arXiv preprint arXiv:2407.15956* (cit. on p. 354).

Alizadeh, F., & Stevens, G. (2020). Think like a human, act like a bot: Explaining instagram's automatic ban decisions. *Conference on Human Factors in Computing Systems Extended Abstracts* (cit. on pp. 43, 179).

Alizadeh, F., Stevens, G., & Esau, M. (2021). I don't know, is ai also used in airbags? an empirical study of folk concepts and people's expectations of current and future artificial intelligence. *I-com*, *20*(1), 3–17 (cit. on pp. 37, 40, 197, 294, 359).

Aljasim, H. K., & Zytko, D. (2022). Foregrounding women's safety in mobile social matching and dating apps: A participatory design study. *Proc. ACM Hum.-Comput. Interact.*, *7*(GROUP). https://doi.org/10.1145/3567559 (cit. on pp. 195, 240).

Alm, C. O., Alvarez, A., Font, J., Liapis, A., Pederson, T., & Salo, J. (2020). Invisible ai-driven hci systems–when, why and how. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 1–3 (cit. on p. 177).

Ammitzbøll Flügge, A., Hildebrandt, T., & Møller, N. H. (2021). Street-level algorithms and ai in bureaucratic decision-making: A caseworker perspective. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1). https://doi.org/10.1145/3449114 (cit. on p. 136).

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), 973–989 (cit. on p. 59).

Ancona, D. (2012). Sensemaking: Framing and acting in the unknown. *The handbook for teaching leadership* (cit. on p. 15).

Andalibi, N., Pyle, C., Barta, K., Xian, L., Jacobs, A. Z., & Ackerman, M. S. (2023a). Conceptualizing algorithmic stigmatization. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18 (cit. on pp. 54, 196, 233, 361).

Andalibi, N., Pyle, C., Barta, K., Xian, L., Jacobs, A. Z., & Ackerman, M. S. (2023b). Conceptualizing algorithmic stigmatization. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580970 (cit. on p. 157).

Anderson, M., Vogels, E. A., & Turner, E. (2020). *The virtues and downsides of online dating*. https://www.pewresearch.org/internet/2020/02/06/the-virtues-and-downsides-of-online-dating/ (cit. on pp. 193, 236).

Arioua, A., & Croitoru, M. (2015). Formalizing explanatory dialogues. *International Conference on Scalable Uncertainty Management*, 282–297 (cit. on pp. 309, 311).

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., et al. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (cit. on p. 318).

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12 (cit. on p. 305).

Ashtari, N., Mullins, R., Qian, C., Wexler, J., Tenney, I., & Pushkarna, M. (2023). From discovery to adoption: Understanding the ml practitioners' interpretability journey. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2304–2325. https://doi.org/10.1145/3563657.3596046 (cit. on p. 134).

Augustein, M., Buschek, D., Herder, E., Loepp, B., Yigitbas, E., & Ziegler, J. (2020). Ucai 2020: Workshop on user-centered artificial intelligence. *Bonn: Gesellschaft für Informatik* (cit. on p. 13).

Ayling, J., & Chapman, A. (2022). Putting ai ethics to work: Are the tools fit for purpose? *AI and Ethics*, *2*. https://doi.org/10.1007/s43681-021-00084-x (cit. on pp. 133, 162).

Badillo-Urquiola, K., Razi, A., Edwards, J., & Wisniewski, P. (2020). Children's perspectives on human sex trafficking prevention education. *Companion of the 2020 ACM International Conference on Supporting Group Work*, 123–126. https://doi.org/10.1145/3323994.3369889 (cit. on p. 193).

Bahl, M. (2020). Artificial intelligence: A primer for breast imaging radiologists. *Journal of Breast Imaging*, *2*(4), 304–314 (cit. on p. 28).

Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man–machine systems* (pp. 129–135). Elsevier. (Cit. on p. 345).

Balebako, R., Jung, J., Lu, W., Cranor, L. F., & Nguyen, C. (2013). "Little brothers watching you": Raising awareness of data leaks on smartphones. *Proceedings of the Ninth Symposium on Usable Privacy and Security - SOUPS '13*, 1. https://doi.org/10.1145/2501604.2501616 (cit. on pp. 61, 258).

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-ai teams: Understanding and addressing the performance/-compatibility tradeoff. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. https://doi.org/10.1609/aaai.v33i01.33012429 (cit. on p. 126).

Baran, B. E., & Scott, C. W. (2010). Organizing ambiguity: A grounded theory of leadership and sensemaking within dangerous contexts. *Military Psychology*, *22*(sup1), S42–S69 (cit. on p. 353).

Baratella, R., Amaral, G., Sales, T. P., Guizzardi, R., & Guizzardi, G. (forthcoming). The many facets of trust. In *Formal ontology in information systems*. IOS Press. (Cit. on p. 132).

Bardzell, J., Bardzell, S., & Blythe, M. (2018). *Critical theory and interaction design*. MIT Press. (Cit. on p. 66).

Barnett, J. T. (2020). Opening moves: Composing climate change. (Cit. on p. 74).

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: From allocative to representational harms in machine learning. *SIGCIS conference paper* (cit. on pp. 53, 196).

Bartneck, C. (2013). Robots in the theatre and the media (cit. on p. 330).

Barton, M. A., Sutcliffe, K. M., Vogus, T. J., & DeWitt, T. (2015). Performing under uncertainty: Contextualized engagement in wildland firefighting. *Journal of Contingencies and Crisis Management*, *23*(2), 74–83 (cit. on p. 354).

Battarbee, K. (2003). Defining co-experience. *Proceedings of the 2003 international conference on Designing pleasurable products and interfaces*, 109–113 (cit. on pp. 95, 335).

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl (ai) ned: The impact of explainable artificial intelligence on users' information processing. *Information systems research*, *34*(4), 1582–1602 (cit. on p. 37).

Beirl, D., Rogers, Y., & Yuill, N. (2019). Using voice assistant skills in family life. *Computer-Supported Collaborative Learning Conference, CSCL*, *1*, 96–103 (cit. on p. 295).

Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction*, *16*(2-4), 193–212 (cit. on p. 257).

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., & Hiniker, A. (2019). Communication breakdowns between families and alexa. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–13 (cit. on pp. 288, 291, 305).

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–24 (cit. on pp. 117, 253, 288).

Berger, R. (2015). Now i see it, now i don't: Researcher's position and reflexivity in qualitative research. *Qualitative research*, *15*(2), 219–234 (cit. on p. 21).

Bindewald, J. M., Rusnock, C. F., & Miller, M. E. (2018). Measuring human trust behavior in human-machine teams. In D. N. Cassenti (Ed.), *Advances in human factors in simulation and modeling* (pp. 47–58). Springer International Publishing. (Cit. on p. 132).

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. *Proceedings of the 2018 Chi conference on human factors in computing systems*, 1–14 (cit. on p. 343).

Birnholtz, J., Fitzpatrick, C., Handel, M., & Brubaker, J. R. (2014). Identity, identification and identifiability: The language of self-presentation on a location-based mobile dating app. *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, 3–12 (cit. on p. 235).

Birnholtz, J., Rawat, S., Vashista, R., Baruah, D., Dange, A., & Boyer, A.-M. (2020). Layers of marginality: An exploration of visibility, impressions, and cultural context on geospatial apps for men who have sex with men in mumbai, india. *Social Media+ Society*, *6*, 2056305120913995 (cit. on pp. 194, 237).

Bisantz, A. M., & Seong, Y. (2001). Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics*, *28*(2), 85–97 (cit. on p. 134).

Blackwell, C., Birnholtz, J., & Abbott, C. (2014). Seeing and being seen: Co-situation and impression formation using grindr, a location-aware gay dating app. *New Media & Society*, 1–20. https://doi.org/10.1177/1461444814521595 (cit. on pp. 194, 195, 237).

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, *1*, 1–19. https://doi.org/10.1145/3134659 (cit. on pp. 191, 193, 236).

Bloom, L. (1980). Language development, language disorders, and learning disabilities: Ld 3. *Bulletin of the Orton Society*, *30*, 115–133 (cit. on p. 291).

Blumer, H. (1986). *Symbolic interactionism: Perspective and method*. Univ of California Press. (Cit. on pp. 16, 93, 95, 335).

Board, D. I. (2019, October). *Ai principles: Recommendations on the ethical use of artificial intelligence by the department of defense* (tech. rep.). United States Department of Defense. Virginia, United States. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB$_$AI$_$PRINCIPLES$_$PRIMARY$_$DOCUMENT.PDF (cit. on pp. 38, 131).

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge. (Cit. on pp. 11, 83).

Bogers, S., van Kollenburg, J., Frens, J., & Djajadiningrat, T. (2016). Data-enabled-design: A situated exploration of rich interactions. *Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems*, 45–48 (cit. on p. 55).

Bohus, D., & Rudnicky, A. I. (2008). Sorry, i didn't catch that! an investigation of non-understanding errors and recovery strategies. *Recent trends in discourse and dialogue*, 123–154 (cit. on p. 291).

Bowen, C. W. (1994). Think-aloud methods in chemistry education: Understanding student thinking. (Cit. on p. 73).

Braun, V., & Clarke, V. (2012a). Thematic analysis. In *APA handbook of research methods in psychology, vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association. (Cit. on pp. 128, 141).

Braun, V., & Clarke, V. (2012b). *Thematic analysis*. American Psychological Association. (Cit. on pp. 202, 205).

Brauner, P., Hick, A., Philipsen, R., & Ziefle, M. (2023). What does the public think about artificial intelligence?—a criticality map to understand bias in the public perception of ai. *Frontiers in Computer Science*, *5*, 1113903 (cit. on p. 330).

Bridgwater, T., Giuliani, M., van Maris, A., Baker, G., Winfield, A., & Pipe, T. (2020). Examining profiles for robotic risk assessment: Does a robot's approach to risk affect user trust? *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 23–31. https://doi.org/10.1145/3319502.3374804 (cit. on p. 159).

Brodie, C., Karat, C.-M., Karat, J., & Feng, J. (2005). Usable security and privacy: A case study of developing privacy management tools. *Proceedings of the 2005 symposium on Usable privacy and security - SOUPS '05*, 35–43. https://doi.org/10.1145/1073001.1073005 (cit. on pp. 61, 259).

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300271 (cit. on pp. 135, 154).

Browne, J., Bakker, S., Yu, B., Lloyd, P., & Ben Allouch, S. (2022). Trust in clinical ai: Expanding the unit of analysis. *HHAI2022: Augmenting Human Intellect*, *354*, 96–113 (cit. on pp. 133, 153).

Brush, A. B., Lee, B., Mahajan, R., Agarwal, S., Saroiu, S., & Dixon, C. (2011). Home automation in the wild: Challenges and opportunities. *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 2115. https://doi.org/10.1145/1978942.1979249 (cit. on p. 256).

Bucher, T. (2019). The algorithmic imaginary: Exploring the ordinary affects of facebook algorithms. In *The social power of algorithms* (pp. 30–44). Routledge. (Cit. on p. 178).

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1). https://doi.org/10.1145/3449287 (cit. on p. 160).

Buolamwini, J. (2022). *Facing the coded gaze with evocative audits and algorithmic audits* [Doctoral dissertation, Massachusetts Institute of Technology]. (Cit. on pp. 53, 196).

Buolamwini, J., & Gebru, T. (2018, 23–24 Feb). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 77–91, Vol. 81). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html (cit. on pp. 156, 334).

Burgess, E. R., Jankovic, I., Austin, M., Cai, N., Kapuścińska, A., Currie, S., Overhage, J. M., Poole, E. S., & Kaye, J. (2023). Healthcare ai treatment decision support: Design principles to enhance clinician adoption and trust. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581251 (cit. on p. 153).

Busciantella-Ricci, D., & Scataglini, S. (2024). Research through co-design. *Design Science*, *10*, e3 (cit. on pp. 67, 70).

Cabitza, F., Campagner, A., Angius, R., Natali, C., & Reverberi, C. (2023). Ai shall have no dominion: On how to measure technology dominance in ai-supported human decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581095 (cit. on p. 136).

Cadogan, R. A., et al. (2004). An imbalance of power: The readability of internet privacy policies. *Journal of Business & Economics Research (JBER)*, *2*(3) (cit. on p. 253).

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., Stumpe, M. C., & Terry, M. (2019). Human-centered tools for coping with imperfect algorithms during medical decision-making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300234 (cit. on pp. 158, 159).

Calvo, R. A., & Peters, D. (2014). *Positive computing: Technology for wellbeing and human potential*. MIT press. (Cit. on p. 343).

Caramiaux, B., & Fdili Alaoui, S. (2022). " explorers of unknown planets" practices and politics of artificial intelligence in visual arts. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–24 (cit. on p. 17).

Carr, A., Day, M., Frier, S., & Gurman, M. (2019). Silicon valley is listening to your most intimate moments. *Bloomberg Businessweek*, *11* (cit. on p. 255).

Castelli, N., Ogonowski, C., Jakobi, T., Stein, M., Stevens, G., & Wulf, V. (2017). What Happened in my Home?: An End-User Development Approach for Smart Home Data Visualization. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 853–866. https://doi.org/10.1145/3025453.3025485 (cit. on pp. 55, 60, 61, 256, 258, 259, 281, 286).

Castelluccia, C., & Le Métaye, D. (2019). *Understanding algorithmic decision-making: Opportunities and challenges*. EU Publications. https://doi.org/10.2861/536131 (cit. on pp. 37, 130).

Cave, S., & Dihal, K. (2020). The whiteness of ai. *Philosophy & Technology*, *33*(4), 685–703 (cit. on p. 366).

Cha, I., Kim, S.-I., Hong, H., Yoo, H., & Lim, Y.-k. (2021). Exploring the use of a voice-based conversational agent to empower adolescents with autism spectrum disorder. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–15 (cit. on p. 50).

Chalmers, M., MacColl, I., & Bell, M. (2003). Seamful design: Showing the seams in wearable computing. *2003 IEE Eurowearable*, 11–16 (cit. on pp. 58, 59, 363).

Chen, Y., Li, H., Teng, S.-Y., Nagels, S., Li, Z., Lopes, P., Zhao, B. Y., & Zheng, H. (2020). Wearable microphone jamming. *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–12 (cit. on p. 254).

Cheng, Y., Yen, K., Chen, Y., Chen, S., & Hiniker, A. (2018). Why doesn't it work? voice-driven interfaces and young children's communication repair strategies. *Proceedings of the 17th ACM conference on interaction design and children*, 337–348 (cit. on pp. 288, 292, 305).

Cheng, Y.-T., Funk, M., Liang, R.-H., & Chen, L.-L. (2023). Seeing through things: Exploring the design space of privacy-aware data-enabled objects. *ACM Transactions on Computer-Human Interaction*, *30*(4), 1–43 (cit. on p. 61).

Cheng, Y.-T., Funk, M., Tsai, W.-C., & Chen, L.-L. (2019). Peekaboo cam: Designing an observational camera for home ecologies concerning privacy. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 823–836 (cit. on p. 61).

Chhetri, C. (2019). Towards a Smart Home Usable Privacy Framework. *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing - CSCW '19*, 43–46. https://doi.org/10.1145/3311957.3361849 (cit. on pp. 258, 279).

Cho, E., Motalebi, N., Sundar, S. S., & Abdullah, S. (2022). Alexa as an active listener: How backchanneling can elicit self-disclosure and promote user experience. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–23 (cit. on p. 312).

Cho, J., & Rader, E. (2020). The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–28 (cit. on pp. 56, 57, 289, 292, 294, 306, 310, 312, 316).

Cho, M., Lee, S.-s., & Lee, K.-P. (2019). Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten. *Proceedings of the 2019 on designing interactive systems conference*, 1557–1569 (cit. on pp. 253, 254).

Chung, H., Iorga, M., Voas, J., & Lee, S. (2017). Alexa, can i trust you? *Computer*, *50*(9), 100–104 (cit. on p. 254).

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Wash. L. Rev.*, *89*, 1 (cit. on pp. 53, 196).

Clark, H. H. (1996). *Using language*. Cambridge University Press. (Cit. on p. 302).

Clarke, V., & Braun, V. (2013). Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist*, *26*, 120–123 (cit. on pp. 128, 141).

Clarke, V., & Braun, V. (2017). Thematic analysis. *The journal of positive psychology*, *12*(3), 297–298 (cit. on p. 300).

Clausen, M., Kyhn, M. P., Papachristos, E., & Merritt, T. (2023). Exploring humor as a repair strategy during communication breakdowns with voice assistants. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–9 (cit. on pp. 291, 305).

Clegg, S. R. (2023). Frameworks of power (cit. on p. 49).

Clegg, S. R., & Haugaard, M. (2009). *The sage handbook of power*. Sage. (Cit. on p. 49).

Cobbe, J., Veale, M., & Singh, J. (2023). Understanding accountability in algorithmic supply chains. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1186–1197. https://doi.org/10.1145/3593013.3594073 (cit. on p. 163).

Cohen, A. E. (2024). Challenging transhumanist apocalyptic ai narratives through speculative fabulation. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8 (cit. on pp. 29, 327).

Colville, I., Brown, A. D., & Pye, A. (2012). Simplexity: Sensemaking, organizing and storytelling for our time. *Human relations*, *65*(1), 5–15 (cit. on p. 354).

Colville, I., Pye, A., & Brown, A. D. (2016). Sensemaking processes and weickarious learning. *Management Learning*, *47*(1), 3–13 (cit. on p. 355).

Commission, E. (2020, February). *On artificial intelligence - a european approach to excellence and trust* (tech. rep.). European Commission. Brussels, Belgium. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020$_$en.pdf (cit. on pp. 38, 131).

Commons, M. L., & Miller, P. M. (2002). A complete theory of human evolution of intelligence must consider stage changes. *Behavioral and Brain Sciences*, *25*(3), 404–405. https://doi.org/10.1017/S0140525X02240078 (cit. on p. 333).

Comunello, F., Parisi, L., & Ieracitano, F. (2020). Negotiating gender scripts in mobile dating apps: Between affordances, usage norms and practices. *Information, Communication & Society*, 1–17 (cit. on pp. 195, 236).

Corriero, E. F., & Tong, S. T. (2016). Managing uncertainty in mobile dating applications: Goals, concerns of use, and information seeking in grindr. *Mobile Media & Communication*, *4*, 121–141 (cit. on pp. 195, 237).

Courtois, C., & Timmermans, E. (2018). Cracking the tinder code: An experience sampling approach to the dynamics and impact of platform governing algorithms. *Journal of Computer-Mediated Communication*, *23*(1), 1–16 (cit. on pp. 40, 45, 190, 199, 200, 238).

Cranor, L. F. (2008). A Framework for Reasoning About the Human in the Loop. *Proceedings of the 1st Conference on Usability, Psychology, and Security*, 1–15 (cit. on pp. 61, 253, 259).

Crowley, J., OrSullivan, A., Nowak, A., Jonker, C., Pedreschi, D., Giannotti, F., & Rogers, Y. (2019). Toward ai systems that augment and empower humans by understanding us, our society and the world around us. *Report of*, *761758*, 1–32 (cit. on p. 13).

Cuadra, A., Li, S., Lee, H., Cho, J., & Ju, W. (2021). My bad! repairing intelligent voice assistant errors improves interaction. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–24 (cit. on pp. 289, 293).

Daim, T. U., Ha, A., Reutiman, S., Hughes, B., Pathak, U., Bynum, W., & Bhatla, A. (2012). Exploring the communication breakdown in global virtual teams. *International journal of project management*, *30*(2), 199–212 (cit. on p. 291).

Das Swain, V., Gao, L., Wood, W. A., Matli, S. C., Abowd, G. D., & De Choudhury, M. (2023). Algorithmic power or punishment: Information worker perspectives on passive sensing enabled ai phenotyping of performance and wellbeing. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581376 (cit. on pp. 154, 156).

Davidoff, S., Zimmerman, J., & Dey, A. K. (2010). How routine learners can support family coordination. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2461–2470 (cit. on pp. 251, 261, 281).

De Marsico, M., Marchionni, L., Novelli, A., & Oertel, M. (2017). Fatcha: Biometrics lends tools for captchas. *Multimedia Tools and Applications*, *76*, 5117–5140 (cit. on p. 181).

De Tona, C., et al. (2006). But what is interesting is the story of why and how migration happened. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, *7*(3) (cit. on p. 21).

Deng, W. H., Yildirim, N., Chang, M., Eslami, M., Holstein, K., & Madaio, M. (2023). Investigating practices and opportunities for cross-functional collaboration around ai fairness in industry practice. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 705–716. https://doi.org/10.1145/3593013.3594037 (cit. on p. 134).

Dervin, B. (2015). Dervin's sense-making theory. In *Information seeking behavior and technology adoption: Theories and trends* (pp. 59–80). IGI Global. (Cit. on pp. 33, 36, 350).

Dervin, B., & Naumer, C. M. (2009). Sense-making. *Encyclopedia of communication theory*, *2*, 876–880 (cit. on pp. 33, 50, 341).

Deshpande, A., & Sharp, H. (2022). Responsible ai systems: Who are the stakeholders? *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 227–236. https://doi.org/10.1145/3514094.3534187 (cit. on p. 133).

Deshpande, A., Rajpurohit, T., Narasimhan, K., & Kalyan, A. (2023). Anthropomorphization of ai: Opportunities and risks. *arXiv preprint arXiv:2305.14784* (cit. on p. 330).

DeVito, M. A., Birnholtz, J., Hancock, J. T., French, M., & Liu, S. (2018). How people form folk theories of social media feeds and what it means for how we study self-presentation. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12 (cit. on pp. 40, 41, 197, 198, 238).

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017). " algorithms ruin everything" # riptwitter, folk theories, and resistance to algorithmic change in social media. *Proceedings of the 2017 CHI conference on human factors in computing systems*, 3163–3174 (cit. on pp. 18, 37, 40, 41, 87, 88, 191, 197, 198, 230, 329, 331, 359).

DeVito, M. A., Walker, A. M., & Birnholtz, J. (2018). 'too gay for facebook' presenting lgbtq+ identity throughout the personal social media ecosystem. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–23 (cit. on p. 234).

DeVito, M. A. (2021). Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *Proc. ACM Hum. Comput. Interact.*, *5*(CSCW2), 1–38 (cit. on pp. 41, 198, 230–232).

DeVito, M. A. (2022). How transfeminine tiktok creators navigate the algorithmic trap of visibility via folk theorization. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–31 (cit. on pp. 227, 230, 231, 233, 234, 238).

DeVito, M. A., Walker, A. M., & Fernandez, J. R. (2021). Values (mis) alignment: Exploring tensions between platform and lgbtq+ community design values. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–27 (cit. on p. 234).

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62 (cit. on p. 345).

Diao, W., Liu, X., Zhou, Z., & Zhang, K. (2014). Your voice assistant is mine: How to abuse speakers to steal information and control your phone. *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, 63–74 (cit. on pp. 254, 279).

Dietrich, T., Trischler, J., Schuster, L., & Rundle-Thiele, S. (2017). Co-designing services with vulnerable consumers. *Journal of Service Theory and Practice*, *27*(3), 663–688 (cit. on pp. 67, 68).

D'ignazio, C., & Klein, L. F. (2023). *Data feminism*. MIT press. (Cit. on pp. 55, 182).

Doerfler, P., Forte, A., Cristofaro, E. D., Stringhini, G., Blackburn, J., & McCoy, D. (2021). "i'm a professor, which isn't usually a dangerous job": Internet-facilitated harassment and its impact on researchers. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–32. https://doi.org/10.1145/3476082 (cit. on pp. 191, 193, 236).

Dörrenbächer, J., Löffler, D., & Hassenzahl, M. (2020). Becoming a robot-overcoming anthropomorphism with techno-mimesis. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–12 (cit. on pp. 298, 299).

Dragoni, A. F., Sernani, P., Calvaresi, D., et al. (2018). When rationality entered time and became a real agent in a cyber-society. *RTA-CSIT*, 167–171 (cit. on p. 46).

Draxler, S., Stevens, G., Stein, M., Boden, A., & Randall, D. (2012). Supporting the social context of technology appropriation: On a synthesis of sharing tools and tool knowledge. *Proceedings of the SIGCHI conference on human factors in computing systems*, 2835–2844 (cit. on p. 16).

Drever, E. (1995). *Using semi-structured interviews in small-scale research. a teacher's guide.* ERIC. (Cit. on p. 205).

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). " hey google is it ok if i eat you?" initial explorations in child-agent interaction. *Proceedings of the 2017 conference on interaction design and children*, 595–600 (cit. on pp. 36, 47, 48, 91, 92, 359).

Du, Y., Zhang, K., Ramabadran, S., & Liu, Y. (2021). "alexa, what is that sound?" a video analysis of child-agent communication from two amazon alexa games. *Proceedings of the 20th Annual ACM Interaction Design and Children Conference*, 513–520 (cit. on p. 291).

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63–71. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2019.01.021 (cit. on p. 130).

Duguay, S. (2017). *Identity modulation in networked publics: Queer women's participation and representation on tinder, instagram, and vine* [Doctoral dissertation, Queensland University of Technology]. (Cit. on pp. 194, 234).

Duguay, S. (2016). Lesbian, gay, bisexual, trans, and queer visibility through selfies: Comparing platform mediators across ruby rose's instagram and vine presence. *Social Media+ Society*, *2*(2), 2056305116641975 (cit. on p. 234).

Duguay, S., Burgess, J., & Suzor, N. (2020). Queer women's experiences of patchwork platform governance on tinder, instagram, and vine. *Convergence*, *26*(2), 237–252 (cit. on p. 194).

Duguay, S., Dietzel, C., & Myles, D. (2022). The year of the "virtual date": Reimagining dating app affordances during the covid-19 pandemic. *new media & society*, 14614448211072257 (cit. on pp. 194, 198).

Dupree, J. L., Devries, R., Berry, D. M., & Lank, E. (2016). Privacy personas: Clustering users via attitudes and behaviors toward security practices. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5228–5239 (cit. on p. 280).

Durey, A., Ward, P., Haynes, E., Baker, S., Calache, H., & Slack-Smith, L. (2024). Applying social practice theory to explore australian preschool children's oral health. *JDR Clinical & Translational Research*, 23800844241235615 (cit. on p. 349).

Dybkjær, L., Bernsen, N. O., & Dybkjær, H. (1996). Grice incorporated: Cooperativity in spoken dialogue. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics* (cit. on p. 312).

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021a). Expanding explainability: Towards social transparency in ai systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445188 (cit. on pp. 153–155).

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021b). Expanding explainability: Towards social transparency in ai systems. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–19 (cit. on p. 336).

Ehsan, U., Liao, Q. V., Passi, S., Riedl, M. O., & Daumé III, H. (2024). Seamful xai: Operationalizing seamful design in explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–29 (cit. on p. 58).

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable ai: Towards a reflective sociotechnical approach. *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, 449–466 (cit. on pp. 13, 294, 363).

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: A technique for explainable ai and its effects on human perceptions. *Proceedings of the 24th international conference on intelligent user interfaces*, 263–274 (cit. on p. 294).

Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A., & Riedl, M. O. (2021). Operationalizing human-centered perspectives in explainable ai. *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 1–6 (cit. on p. 58).

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of computer-mediated communication*, *11*(2), 415–441 (cit. on p. 195).

Elsden, C., Chatting, D., Durrant, A. C., Garbett, A., Nissen, B., Vines, J., & Kirk, D. S. (2017). On speculative enactments. *Proceedings of the 2017 CHI conference on human factors in computing systems*, 5386–5399 (cit. on p. 298).

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First i" like" it, then i hide it: Folk theories of social feeds. *Proceedings of the 2016 cHI conference on human factors in computing systems*, 2371–2382 (cit. on pp. 17, 40, 57, 83, 87, 93, 115, 191, 197, 204, 329, 359, 363).

Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating algorithmic process in online behavioral advertising. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–13 (cit. on p. 313).

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). " i always assumed that i wasn't really that close to [her]" reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 153–162 (cit. on pp. 17, 30, 39, 55, 59, 83, 87, 88, 93, 115, 178, 190, 329).

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. *Proceedings of the international AAAI conference on web and social media*, *11*(1), 62–71 (cit. on pp. 59, 87, 88, 178, 359).

Eslami, M., Vaccaro, K., Lee, M. K., Elazari Bar On, A., Gilbert, E., & Karahalios, K. (2019). User attitudes towards algorithmic opacity and transparency in online reviewing platforms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on pp. 30, 40, 59, 83, 87, 88, 93, 197).

Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. *Proceedings of the AAAI conference on artificial intelligence*, *31*(1) (cit. on pp. 11, 83, 359).

Feldman, S. P. (1994). Cultural knowledge in organizations: Exploring the collective mind. (Cit. on p. 15).

Fernandez, J. R., & Birnholtz, J. (2019). " i don't want them to not know" investigating decisions to disclose transgender identity on dating platforms. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–21. https://doi.org/https://doi.org/10.1145/3359328 (cit. on pp. 195, 227, 240).

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104 (cit. on p. 41).

Ferrario, A., & Loi, M. (2022). How explainability contributes to trust in ai. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1457–1466. https://doi.org/10.1145/3531146.3533202 (cit. on pp. 135, 136).

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair ai: Challenges and opportunities. *Business & information systems engineering*, *62*, 379–384 (cit. on p. 14).

Fiesler, C., & Proferes, N. (2018). "participant" perceptions of twitter research ethics. *Social Media+ Society*, *4*(1), 2056305118763366 (cit. on p. 202).

Filina, E. (2024). Decoding intelligence in artificial intelligence: Tracing historical evolution, analysing definitions, and regulatory challenges. *International Journal of Research and Innovation in Social Science*, *8*(1), 1880–1893 (cit. on p. 28).

Fischer, F., & Kleen, S. (2021). Possibilities, problems, and perspectives of data collection by mobile apps in longitudinal epidemiological studies: Scoping review. *Journal of medical Internet research*, *23*(1), e17691 (cit. on p. 73).

Fischer, J. E., Crabtree, A., Rodden, T., Colley, J. A., Costanza, E., Jewell, M. O., & Ramchurn, S. D. (2016). "Just whack it on until it gets hot": Working with IoT Data in the Home. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5933–5944. https://doi.org/10.1145/2858036.2858518 (cit. on pp. 60, 61, 258, 259).

Fischer, J. E., Reeves, S., Porcheron, M., & Sikveland, R. O. (2019). Progressivity for voice interface design. *Proceedings of the 1st international conference on conversational user interfaces*, 1–8 (cit. on pp. 307, 312–314).

Fitria, R. (2024). Integrating heuristic evaluation and cognitive walkthrough in usability evaluation of mobile application. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *3*(3), 667–670 (cit. on p. 73).

Fitzgerald, M. S., & Palincsar, A. S. (2019). Teaching practices that support student sensemaking across grades and disciplines: A conceptual review. *Review of Research in Education*, *43*(1), 227–248 (cit. on p. 65).

Fleming, P., & Spicer, A. (2014). Power in management and organization science. *Academy of Management Annals*, *8*(1), 237–298 (cit. on p. 49).

Flick, U. (2013). *The sage handbook of qualitative data analysis*. Sage. (Cit. on p. 97).

Fogg, B., & Nass, C. (1997). How users reciprocate to computers: An experiment that demonstrates behavior change. In *Chi'97 extended abstracts on human factors in computing systems* (pp. 331–332). (Cit. on p. 332).

Fogliato, R., Chouldechova, A., & Lipton, Z. (2021). The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2). https://doi.org/10.1145/3479572 (cit. on p. 160).

Fortunati, L., Manganelli, A. M., Cavallo, F., & Honsell, F. (2019). You need to show that you are not a robot. *New Media & Society*, *21*(8), 1859–1876 (cit. on pp. 42, 179, 183).

Freeman, G., Zamanifard, S., Maloney, D., & Acena, D. (2022). Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, *6*, 1–30. https://doi.org/10.1145/3512932 (cit. on pp. 191, 193, 236).

Frohlich, D., & Kraut, R. (2002). The social context of home computing. inside the smart home||. hci institute. (Cit. on pp. 279, 285).

Frost, J. H., Chance, Z., Norton, M. I., & Ariely, D. (2008). People are experience goods: Improving online dating with virtual dates. *Journal of Interactive Marketing*, *22*, 51–61. https://doi.org/https://doi.org/10.1002/dir.20107 (cit. on p. 195).

Fruchter, N., & Liccardi, I. (2018). Consumer attitudes towards privacy and security in home assistants. *Extended Abstracts of the 2018 CHI conference on human factors in computing systems*, 1–6 (cit. on p. 250).

Fund, A. R. (2019, March). *Artificial intelligence: Fostering trust* (tech. rep.). AXA. https://www.axa-research.org/en/news/AI-research-guide (cit. on pp. 38, 131).

G20. (2019, June). *G20 ministerial statement on trade and digital economy* (tech. rep.). G20. Brussels, Belgium. http://trade.ec.europa.eu/doclib/press/index.cfm?id=2027 (cit. on pp. 38, 131).

Gafni, R., & Nagar, I. (2016). The effect of captcha on user experience among users with and without learning disabilities. *Proceedings of the 11th Chais Conference for the Study of Innovation and Learning Technologies: Learning in the Technological Era*, 25–33 (cit. on p. 42).

Gammelgård-Larsen, A., van Berkel, N., Skov, M. B., & Kjeldskov, J. (2024). Designing for human-ai interaction: Comparing intermittent, continuous, and proactive interactions for a music application. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8 (cit. on p. 56).

Gao, C., Zeng, J., Lyu, M. R., & King, I. (2018). Online app review analysis for identifying emerging issues. *Proceedings of the 40th International Conference on Software Engineering*, 48–58 (cit. on p. 72).

Gardin, F. A. (2010). The" think-aloud" method to promote student modeling of expert thinking. *Athletic Therapy Today*, *15*(4) (cit. on p. 73).

Garfinkel, H. (2023). Studies in ethnomethodology. In *Social theory re-wired* (pp. 58–66). Routledge. (Cit. on p. 95).

Garfinkel, S., & Lipford, H. R. (2014). *Usable security: History, themes, and challenges*. Morgan & Claypool Publishers. (Cit. on pp. 61, 258, 259).

Garg, R., & Sengupta, S. (2020). Conversational technologies for in-home learning: Using co-design to understand children's and parents' perspectives. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13 (cit. on pp. 47, 48, 91, 92).

Gáti, D. (2023). Theorizing mathematical narrative through machine learning. *Journal of Narrative Theory*, *53*(1), 139–165 (cit. on pp. 29, 327).

Geeng, C., & Roesner, F. (2019). Who's in control? interactions in multi-user smart homes. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–13 (cit. on p. 285).

Gemalmaz, M. A., & Yin, M. (2022). Understanding decision subjects' fairness perceptions and retention in repeated interactions with ai-based decision systems. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 295–306. https://doi.org/10.1145/3514094.3534201 (cit. on pp. 128, 134).

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *new media & society*, *22*(7), 1266–1286 (cit. on pp. 53, 196).

Gerring, J. (1999). What makes a concept good? a criterial framework for understanding concept formation in the social sciences. *Polity*, *31*(3), 357–393 (cit. on p. 329).

Ghallab, M. (2019). Responsible ai: Requirements and challenges. *AI Perspectives*, *1*(1), 1–7 (cit. on p. 14).

Giaccardi, E. (2019). Histories and futures of research through design: From prototypes to connected things. *International Journal of Design*, *13*(3), 139–155 (cit. on p. 64).

Gibbs, J. L., Ellison, N. B., & Lai, C.-H. (2011). First comes love, then comes google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. *Communication Research*, *38*, 70–100. https://doi.org/10.1177/0093650210377091 (cit. on pp. 195, 231).

Gil, Y., & Selman, B. (2019). A 20-year community roadmap for artificial intelligence research in the us. *arXiv preprint arXiv:1908.02624* (cit. on p. 82).

Gilbert, L., Sarvet, A. L., Wall, M., Walsh, K., Reardon, L., Wilson, P., Santelli, J., Khan, S., Thompson, M., Hirsch, J. S., & Mellins, C. A. (2019). Situational contexts and risk factors associated with incapacitated and nonincapacitated sexual assaults among college women. *Journal of Women's Health*, *28*, 185–193. https://doi.org/10.1089/jwh.2018.7191 (cit. on p. 194).

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. (Cit. on p. 332).

Gioia, D. A., & Chittipeddi, K. (1991). Sensemaking and sensegiving in strategic change initiation. *Strategic management journal*, *12*(6), 433–448 (cit. on p. 51).

Glikson, E., & Woolley, A. (2020). Human trust in artificial intelligence: Review of empirical research (in press). *The Academy of Management Annals*, *14*(2). https://doi.org/https://doi.org/10.5465/annals.2018.0057 (cit. on p. 132).

Greenwald, E., Krakowski, A., Hurt, T., Grindstaff, K., & Wang, N. (2024). It's like i'm the ai: Youth sensemaking about ai through metacognitive embodiment. *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, 789–793 (cit. on p. 55).

Grice, H. (1975). Logic and conversation. *Syntax and semantics*, *3* (cit. on pp. 311, 312).

Guerar, M., Merlo, A., Migliardi, M., & Palmieri, F. (2018). Invisible cappcha: A usable mechanism to distinguish between malware and humans on the mobile iot. *computers & security*, *78*, 255–266 (cit. on p. 41).

Güngör, H. (2020). Creating value with artificial intelligence: A multi-stakeholder perspective. *Journal of Creating Value*, *6*(1), 72–85. https://doi.org/10.1177/2394964320921071 (cit. on p. 133).

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120 (cit. on p. 36).

Gupta, A., Basu, D., Ghantasala, R., Qiu, S., & Gadiraju, U. (2022). To trust or not to trust: How a conversational interface affects trust in a decision support system. *Proceedings of the ACM Web Conference 2022*, 3531–3540. https://doi.org/10.1145/3485447.3512248 (cit. on p. 159).

Habermas, J. (2015). *Strukturwandel der öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft ; mit einem Vorwort zur Neuauflage 1990* (14. Aufl) [OCLC: 953254875]. Suhrkamp. (Cit. on p. 256).

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and machines*, *30*(1), 99–120 (cit. on p. 14).

Hagras, H. (2018). Toward human-understandable, explainable ai. *Computer*, *51*(9), 28–36 (cit. on pp. 13, 57, 86).

Haimson, O. L., Brubaker, J. R., Dombrowski, L., & Hayes, G. R. (2015). Disclosure, stress, and support during gender transition on facebook. *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 1176–1190 (cit. on p. 234).

Halpin, M. (2022). Weaponized subordination: How incels discredit themselves to degrade women. *Gender & Society*, *36*(6), 813–837. https://doi.org/10.1177/08912432221128545 (cit. on p. 237).

Hamilton, M. (2019). The sexist algorithm. *Behavioral sciences & the law*, *37*(2), 145–157 (cit. on pp. 53, 196).

Hancock, J. T., Naaman, M., & Levy, K. (2020). Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100 (cit. on pp. 39, 190).

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254 (cit. on p. 132).

Hardy, J., & Lindtner, S. (2017). Constructing a desiring user: Discourse, rurality, and design in location-based social networks. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 13–25 (cit. on p. 235).

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing science*, *19*(1), 4–21 (cit. on p. 190).

Heath, R. L., & Johansen, W. (2018). *The international encyclopedia of strategic communication*. Wiley Online Library. (Cit. on pp. 32, 33, 35).

Hektner, J. (2007). *Experience sampling method: Measuring the quality of everyday life*. Sage Publications. (Cit. on p. 297).

Higashinaka, R., D'Haro, L. F., Abu Shawar, B., Banchs, R. E., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., & Sedoc, J. (2021). Overview of the dialogue breakdown detection challenge 4. *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, 403–417 (cit. on p. 292).

Hobbes, T., & Missner, M. (2016). *Thomas hobbes: Leviathan (longman library of primary sources in philosophy)*. Routledge. (Cit. on p. 51).

Hocutt, D. (2021). Interrogating alexa: Holding voice assistants accountable for their answers. *Proceedings of the 39th ACM International Conference on Design of Communication*, 142–150 (cit. on pp. 56, 294).

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570 (cit. on pp. 126, 132).

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (cit. on pp. 57, 83, 85, 363).

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300830 (cit. on p. 135).

Honer, A. (2004). Life-world analysis in ethnography. *A companion to qualitative research*, 113–17 (cit. on p. 69).

Hong, J., & Cross Francis, D. (2020). Unpacking complex phenomena through qualitative inquiry: The case of teacher identity research. *Educational psychologist*, *55*(4), 208–219 (cit. on p. 68).

Hou, Q., Han, M., & Cai, Z. (2020). Survey on data analysis in social media: A practical application aspect. *Big Data Mining and Analytics*, *3*(4), 259–279 (cit. on p. 72).

Hou, Y. T.-Y., Lee, W.-Y., & Jung, M. (2023). "should i follow the human, or follow the robot?" — robots in power can have more influence than humans on decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581066 (cit. on p. 136).

Huang, S. A., Hancock, J., & Tong, S. T. (2022). Folk theories of online dating: Exploring people's beliefs about the online dating process and online dating algorithms. *Social Media+ Society*, *8*(2), 20563051221089561 (cit. on pp. 44, 46, 191, 200, 201, 229, 331).

Huang, Y. (2024). Levels of ai agents: From rules to large language models. *arXiv preprint arXiv:2405.06643* (cit. on p. 46).

Huang, Y., Obada-Obieh, B., & Beznosov, K. (2020). Amazon vs. my brother: How users of shared smart speakers perceive and cope with privacy risks. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13 (cit. on p. 285).

Hutson, J. A., Taft, J. G., Barocas, S., & Levy, K. (2018). Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–18 (cit. on p. 190).

Iacucci, G., Kuutti, K., & Ranta, M. (2000). On the move with a magic thing: Role playing in concept design of mobile services and devices. *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, 193–202 (cit. on p. 74).

Inman, S., & Ribes, D. (2019). " beautiful seams" strategic revelations and concealments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on p. 58).

Jacobs, M., He, J., F. Pradier, M., Lam, B., Ahn, A. C., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). Designing ai for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445385 (cit. on p. 154).

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923 (cit. on pp. 131, 132, 144, 156).

Jain, A. K., & Kumar, A. (2012). Biometric recognition: An overview. *Second generation biometrics: The ethical, legal and social context*, 49–79 (cit. on pp. 54, 180).

Jain, N., & Brockova, K. (2022). Adapting data collection tools for qualitative research in times of covid-19 pandemic. *MAP Social Sciences*, *2*(1), 18–27 (cit. on p. 72).

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How different groups prioritize ethical values for responsible ai. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 310–323. https://doi.org/10.1145/3531146.3533097 (cit. on pp. 127, 133, 159, 162).

Jakobi, T., Stevens, G., Castelli, N., Ogonowski, C., Schaub, F., Vindice, N., Randall, D., Tolmie, P., & Wulf, V. (2018). Evolving Needs in IoT Control and Accountability: A Longitudinal Study on Smart Home Intelligibility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(4), 171:1–171:28. https://doi.org/10.1145/3287049 (cit. on pp. 61, 257, 261, 281, 286).

Jakobi, T., Stevens, G., von Grafenstein, M., Pins, D., & Boden, A. (2020). User-friendly formulation of data processing purposes of voice assistants: A user perspective on the principle of purpose limitation. *Proceedings of the Conference on Mensch und Computer*, 361–372. https://doi.org/10.1145/3404983.3405588 (cit. on pp. 60, 253, 255).

Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. *Journal of computer and system sciences*, *80*(5), 973–993 (cit. on p. 181).

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction*, *25*, 1–33. https://doi.org/10.1145/3185593 (cit. on pp. 191, 193, 236).

Jhaver, S., Karpfen, Y., & Antin, J. (2018). Algorithmic anxiety and coping strategies of airbnb hosts. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12 (cit. on pp. 40, 190, 197).

Jiang, J., Jeng, W., & He, D. (2013). How do users respond to voice input errors? lexical and phonetic query reformulation in voice search. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 143–152 (cit. on p. 291).

Kacen, L., & Chaitin, J. (2006). 'the times they are a changing': Undertaking qualitative research in ambiguous, conflictual, and changing contexts. *The Qualitative Report*, *11*(2), 209–228 (cit. on p. 21).

Kalhan, A. (2013). Immigration policing and federalism through the lens of technology, surveillance, and privacy. *Ohio St. LJ*, *74*, 1105 (cit. on pp. 38, 126, 131).

Kang, R., Dabbish, L., Fruchter, N., & Kiesler, S. (2015a). "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. *Proceedings of the Eleventh USENIX Conference on Usable Privacy and Security*, 39–52. https://www.usenix.org/system/files/conference/soups2015/soups15-paper-kang.pdf (cit. on p. 255).

Kang, R., Dabbish, L., Fruchter, N., & Kiesler, S. (2015b). {"my} data just goes {everywhere:"} user mental models of the internet and implications for privacy and security. *Eleventh symposium on usable privacy and security (SOUPS 2015)*, 39–52 (cit. on p. 250).

Kapania, S., Siy, O., Clapper, G., SP, A. M., & Sambasivan, N. (2022). "because ai is 100% right and safe": User attitudes and sources of ai authority in india. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517533 (cit. on p. 160).

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on human-computer interaction*, *5*(CSCW2), 1–44 (cit. on pp. 17, 40, 41, 54, 196–198, 234, 339, 361).

Katzman, J., Wang, A., Scheuerman, M., Blodgett, S. L., Laird, K., Wallach, H., & Barocas, S. (2023). Taxonomizing and measuring representational harms: A look at image tagging. *arXiv preprint arXiv:2305.01776* (cit. on pp. 54, 196).

Kaur, E., Haghighi, P. D., Burstein, F., Urquhart, D., & Cicuttini, F. (2020). Challenges and opportunities of mobile data collection in clinical studies. *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, 129–137 (cit. on p. 74).

Kaur, H., Adar, E., Gilbert, E., & Lampe, C. (2022). Sensible ai: Re-imagining interpretability and explainability using sensemaking theory. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 702–714 (cit. on p. 34).

Keedle, H., Schmied, V., Burns, E., & Dahlen, H. (2018). The design, development, and evaluation of a qualitative data collection application for pregnant women. *Journal of Nursing Scholarship*, *50*(1), 47–55 (cit. on p. 73).

Kennedy, A., Wilkes, A., Elder, L., & Murray, W. S. (1988). Dialogue with machines. *Cognition*, *30*(1), 37–72 (cit. on p. 291).

Kianzad, S., Kim, Y., Lindsay, J. A. B., Huang, Y., Benavides, J. B., Leung, R., & MacLean, K. E. (n.d.). Accountability-aware design of voice user interfaces for home appliances. *Graphics Interface 2021* (cit. on pp. 57, 289, 295, 308, 314, 316, 363).

Kiesler, S., & Goetz, J. (2002). Mental models of robotic assistants. *CHI'02 extended abstracts on Human Factors in Computing Systems*, 576–577 (cit. on pp. 47, 90).

Kim, J. Y., Boag, W., Gulamali, F., Hasan, A., Hogg, H. D. J., Lifson, M., Mulligan, D., Patel, M., Raji, I. D., Sehgal, A., Shaw, K., Tobey, D., Valladares, A., Vidal, D., Balu, S., & Sendak, M. (2023). Organizational governance of emerging technologies: Ai adoption in healthcare. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1396–1417. https://doi.org/10.1145/3593013.3594089 (cit. on p. 134).

Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., & Choudhury, M. D. (2021). A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–34. https://doi.org/10.1145/3476066 (cit. on pp. 193, 236).

Kim, T., Molina, M. D., Rheu, M. (, Zhan, E. S., & Peng, W. (2023). One ai does not fit all: A cluster analysis of the laypeople's perception of ai roles. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581340 (cit. on p. 136).

Kizilcec, R. F. (2016). How much information? effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2390–2395 (cit. on p. 14).

Klein, G. (2000). Cognitive task analysis of teams. In *Cognitive task analysis* (pp. 431–444). Psychology Press. (Cit. on p. 34).

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press. (Cit. on p. 355).

Kolkman, D., Bex, F., Narayan, N., & van der Put, M. (2024). Justitia ex machina: The impact of an ai system on legal decision-making and discretionary authority. *Big Data & Society*, *11*(2), 20539517241255101 (cit. on p. 38).

Koskinen, I., Zimmerman, J., Binder, T., Redstrom, J., & Wensveen, S. (2013). Design research through practice: From the lab, field, and showroom. *IEEE Transactions on Professional Communication*, *56*(3), 262–263 (cit. on p. 67).

KPMG. (2019, June). *Controlling ai: The imperative for transparency and explainability* (tech. rep.). KPMG. https://advisory.kpmg.us/articles/2019/controlling-ai.html (cit. on pp. 38, 131).

Kraus, M., Wagner, N., & Minker, W. (2020). Effects of proactive dialogue strategies on human-computer trust. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 107–116. https://doi.org/10.1145/3340631.3394840 (cit. on p. 160).

Krings, K., Bohn, N. S., Hille, N. A. L., & Ludwig, T. (2023). "what if everyone is able to program?"–exploring the role of software development in science fiction. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on p. 71).

Krol, K., Parkin, S., & Sasse, M. A. (2016a). " i don't like putting my face on the internet!": An acceptance study of face biometrics as a captcha replacement. *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 1–7 (cit. on p. 185).

Krol, K., Parkin, S., & Sasse, M. A. (2016b). Better the devil you know: A user study of two captchas and a possible replacement technology. *NDSS Workshop on Usable Security (USEC)*, *10* (cit. on pp. 183, 185, 186).

Kuhl, N., Lobana, J., & Meske, C. (2020). Do you comply with ai?–personalized explanations of learning algorithms and their impact on employees' compliance behavior. *arXiv preprint arXiv:2002.08777* (cit. on pp. 48, 89).

Kuijer, L., & Giaccardi, E. (2018). Co-performance: Conceptualizing the role of artificial agency in the design of everyday life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13 (cit. on pp. 308, 316).

Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more? the effects of mental model soundness on personalizing an intelligent agent. *Proceedings of the sigchi conference on human factors in computing systems*, 1–10 (cit. on pp. 83, 88, 89, 93, 115).

Kurze, A., Bischof, A., Totzauer, S., Storz, M., Eibl, M., Brereton, M., & Berger, A. (2020). Guess the Data: Data Work to Understand How People Make Sense of and Use Simple Sensor Data from Homes. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376273 (cit. on pp. 61, 258, 259, 279).

Kuzminykh, A., Sun, J., Govindaraju, N., Avery, J., & Lank, E. (2020). Genie in the bottle: Anthropomorphized perceptions of conversational agents. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13 (cit. on pp. 47, 91).

Kvale, S. (2009). *Interviews: Learning the craft of qualitative research interviewing*. Sage. (Cit. on p. 68).

Kwak, I.-Y., Huh, J. H., Han, S. T., Kim, I., & Yoon, J. (2019). Voice presentation attack detection through text-converted voice command analysis. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12 (cit. on p. 254).

Ladner, R. E. (2015). Design for user empowerment. *interactions*, *22*(2), 24–29 (cit. on p. 50).

Lagioia, F., & Contissa, G. (2020). The strange case of Dr Watson : Liability implications of AI evidence-based decision support systems in health care. *Eur. J. Leg. Stud.*, *12*(2), 241–289 (cit. on pp. 38, 130).

Lahoual, D., & Frejus, M. (2019). When users assist the voice assistants: From supervision to failure resolution. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–8 (cit. on pp. 292, 310).

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1369–1385. https://doi.org/10.1145/3593013.3594087 (cit. on pp. 39, 126, 134, 162, 334).

Langer, E. J., & Ngnoumen, C. T. (2017). Mindfulness. In *Positive psychology* (pp. 95–111). Routledge. (Cit. on p. 355).

Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on human-computer interaction*, *2*(CSCW), 1–31 (cit. on pp. 252–254, 286).

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. *Proceedings of the SIGCHI conference on human factors in computing systems*, 719–728 (cit. on p. 117).

Lee, J., & See, K. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*, 50–80. https://doi.org/10.1518/hfes.46.1.50.30392 (cit. on pp. 123, 132).

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684 (cit. on p. 115).

Lee, M. (2021). Interactional morality: Technology as our moral mirror (cit. on p. 18).

Lee, M. (2020). Speech acts redux: Beyond request-response interactions. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–10 (cit. on p. 302).

Lee, S., Kim, K. J., & Sundar, S. S. (2015). Customization in location-based advertising: Effects of tailoring source, locational congruity, and product involvement on ad attitudes. *Computers in Human Behavior*, *51*, 336–343 (cit. on p. 347).

Lee, S., Kim, S., & Lee, S. (2019). " what does your agent look like?" a drawing study to understand users' perceived persona of conversational agent. *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, 1–6 (cit. on pp. 47, 48, 89, 91, 92).

Lewicki, R., & Brinsfield, C. (2011, January). Measuring trust beliefs and behaviours. In F. Lyon, G. Möllering, & M. Saunders (Eds.), *Handbook of research methods on trust* (pp. 29–39). Edward Elgar. https://doi.org/10.4337/9781781009246.00013 (cit. on p. 126).

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces*, *63*(4), 967–985. http://www.jstor.org/stable/2578601 (cit. on p. 132).

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (cit. on p. 294).

Liao, Q. V., Wang, Y.-C., Bickmore, T., Fung, P., Grudin, J., Yu, Z., & Zhou, M. (2019). Human-agent communication: Connecting research and development in hci and ai. *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 122–126 (cit. on p. 352).

Liao, Q., & Sundar, S. S. (2022). Designing for responsible trust in ai systems: A communication perspective. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1257–1268. https://doi.org/10.1145/3531146.3533182 (cit. on pp. 145, 156).

Liao, T., & Tyson, O. (2021). "crystal is creepy, but cool": Mapping folk theories and responses to automated personality recognition algorithms. *Social Media+ Society*, *7*(2), 20563051211010170 (cit. on pp. 40, 197).

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the SIGCHI conference on human factors in computing systems*, 2119–2128 (cit. on pp. 57, 83, 84).

Lima, G., Grgić-Hlača, N., Jeong, J. K., & Cha, M. (2022). The conflict between explainable and accountable decision-making algorithms. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2103–2113. https://doi.org/10.1145/3531146.3534628 (cit. on p. 134).

Lind, M., & Dickel, S. (2024). Speaking, but having no voice. negotiating agency in advertisements for intelligent personal assistants. *Convergence*, *30*(3), 1008–1024 (cit. on pp. 311, 315, 352).

Lockey, S., Gillespie, N., Holm, D., & Asadi Someh, I. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. *HICSS-54*. https://doi.org/10.24251/HICSS.2021.664 (cit. on p. 127).

Long, D., & Magerko, B. (2020). What is ai literacy? competencies and design considerations. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16 (cit. on pp. 11, 30, 36, 83, 87, 359).

Lopatovska, I., & Williams, H. (2018). Personification of the amazon alexa: Bff or a mindless companion. *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 265–268 (cit. on p. 253).

Lopes, M. R., & Vogel, C. (2018). Gender bias on tinder: Transforming an exploratory qualitative survey into statistical data for contextualized interpretation. *Computer Supported Qualitative Research: Second International Symposium on Qualitative Research (ISQR 2017)*, 225–236 (cit. on p. 199).

Lorenzo-Dus, N., Kinzel, A., & Cristofaro, M. D. (2020). The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. *Journal of Pragmatics*, *155*, 15–27. https://doi.org/10.1016/j.pragma.2019.09.010 (cit. on p. 193).

Lowe, S., & Rod, M. (2018). Weathering contextual activities and situated sensemaking. *Journal of Business & Industrial Marketing*, *33*(8), 1141–1152 (cit. on p. 65).

Lucero, A. (2015). Using affinity diagrams to evaluate interactive prototypes. *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15*, 231–248 (cit. on p. 298).

Luger, E., & Sellen, A. (2016). " like having a really bad pa" the gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5286–5297 (cit. on pp. 117, 252).

Lyon, F., Möllering, G., & Saunders, M. (2015, January). *Handbook of research methods on trust: Second edition*. Edward Elgar Publishing. https://doi.org/10.4337/9781782547419 (cit. on p. 126).

Lyons, H., Miller, T., & Velloso, E. (2023). Algorithmic decisions, desire for control, and the preference for human review over algorithmic review. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 764–774. https://doi.org/10.1145/3593013.3594041 (cit. on p. 134).

Lyons, H., Wijenayake, S., Miller, T., & Velloso, E. (2022). What's the appeal? perceptions of review processes for algorithmic decisions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517606 (cit. on p. 134).

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581058 (cit. on pp. 38, 126).

Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proc. ACM Hum.-Comput. Interact.*, *6*(CSCW1). https://doi.org/10.1145/3512899 (cit. on p. 135).

Magaldi, D., & Berler, M. (2020). Semi-structured interviews. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 4825–4830). Springer International Publishing. https://doi.org/10.1007/978-3-319-24612-3_857 (cit. on p. 139).

Magnet, S. A. (2020). *When biometrics fail: Gender, race, and the technology of identity*. Duke University Press. (Cit. on p. 171).

Maitlis, S. (2005). The social processes of organizational sensemaking. *Academy of management journal*, *48*(1), 21–49 (cit. on p. 51).

Maitlis, S., & Christianson, M. (2014). Sensemaking in organizations: Taking stock and moving forward. *Academy of management annals*, *8*(1), 57–125 (cit. on pp. 32, 34, 36).

Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., & Wagner, D. (2019). Privacy Attitudes of Smart Speaker Users. *Proceedings on Privacy Enhancing Technologies*, *2019*(4), 250–271. https://doi.org/10.2478/popets-2019-0068 (cit. on p. 255).

Maller, C. J. (2015). Understanding health through social practices: Performance and materiality in everyday life. *Sociology of health & illness*, *37*(1), 52–66 (cit. on p. 349).

Malz, E. (2020). *The relationship between online dating, self-esteem and body image* [B.S. thesis]. University of Twente. (Cit. on p. 233).

Marabelli, M., & Newell, S. (2019). Algorithmic decision-making in the us healthcare industry: Good for whom? *Academy of Management Proceedings*, *2019*, 15581. https://doi.org/10.5465/AMBPP.2019.15581 (cit. on p. 135).

March–Russell, P. (2020). And the question of the robot. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*, 165 (cit. on p. 29).

Marcus, A. (2023). Rockets, robots, and ai: Lessons learned from science-fiction movies and tv for hci/ux. In *Acm siggraph 2023 courses* (pp. 1–34). (Cit. on p. 71).

Marian, A. (2023). Algorithmic transparency and accountability through crowdsourcing: A study of the nyc school admission lottery. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 434–443. https://doi.org/10.1145/3593013.3594009 (cit. on pp. 134, 155).

Maurer, S., Erbach, R., Kraiem, I., Kuhnert, S., Grimm, P., & Rukzio, E. (2018). Designing a guardian angel: Giving an automated vehicle the possibility to override its driver. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 341–350. https://doi.org/10.1145/3239060.3239078 (cit. on p. 160).

Mavrina, L., Szczuka, J., Strathmann, C., Bohnenkamp, L. M., Krämer, N., & Kopp, S. (2022). "alexa, you're really stupid": A longitudinal field study on communication breakdowns between family members and a voice assistant. *Frontiers in Computer Science*, *4*, 791704 (cit. on pp. 291, 293, 305).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734 (cit. on p. 333).

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal*, *38*(1), 24–59. http://www.jstor.org/stable/256727 (cit. on p. 132).

McCarthy, J. (2007). From here to human-level ai. *Artificial Intelligence*, *171*(18), 1174–1182 (cit. on p. 10).

McCarthy, J., et al. (2007). What is artificial intelligence (cit. on p. 28).

Mennicken, S., Hofer, J., Dey, A., & Huang, E. M. (2014). Casalendar: A temporal interface for automated homes. In *Chi'14 extended abstracts on human factors in computing systems* (pp. 2161–2166). (Cit. on pp. 261, 281).

Mennicken, S., Hwang, A., Yang, R., Hoey, J., Mihailidis, A., & Huang, E. M. (2015). Smart for life: Designing smart home technologies that evolve with users. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2377–2380 (cit. on p. 354).

Merrill, N., Chuang, J., & Cheshire, C. (2019). Sensing is Believing: What People Think Biosensors Can Reveal About Thoughts and Feelings. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 413–420. https://doi.org/10.1145/3322276.3322286 (cit. on pp. 60, 255, 256, 278).

Meuser, M., & Nagel, U. (2009). Das experteninterview—konzeptionelle grundlagen und methodische anlage. *Methoden der vergleichenden Politik-und Sozialwissenschaft: neue Entwicklungen und Anwendungen*, 465–479 (cit. on p. 70).

Mezias, J. M., & Starbuck, W. H. (2003). Managers and their inaccurate perceptions: Good, bad or inconsequential. *British Journal of Management*, *14*(1), 3–17 (cit. on p. 339).

Mhaidli, A., Venkatesh, M. K., Zou, Y., & Schaub, F. (2020). Listen only when spoken to: Interpersonal communication cues as smart speaker privacy controls. *Proceedings on Privacy Enhancing Technologies* (cit. on p. 255).

Miah, M. W. R., Yearwood, J., & Kulkarni, S. (2011). Detection of child exploiting chats from a mixed chat dataset as a text classification task. *Proceedings of the Australasian Language Technology Association Workshop 2011*, 157–165 (cit. on p. 193).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679 (cit. on p. 346).

Mladenović, M., Ošmjanski, V., & Stanković, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming. *ACM Computing Surveys*, *54*, 1–42. https://doi.org/10.1145/3424246 (cit. on pp. 190, 193, 236).

Mohammadzadeh, B., Françoise, J., Gouiffès, M., & Caramiaux, B. (2024). Studying collaborative interactive machine teaching in image classification. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 195–208 (cit. on p. 17).

Mohammed, M. O. M., Abdallah, A. S. R., Ayele, D. G., Wacho, G. A., & Mohamed, A. A. A. (2024). Multiple correspondence analysis for assessment of the socioeconomic and social impact of the covid-19 pandemic in gezira state, sudan. *The Open Public Health Journal*, *17*(1) (cit. on p. 70).

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. (Cit. on pp. 57, 58, 83, 86, 87).

Moon, Y., & Nass, C. (1996). How "real" are computer personalities? psychological responses to personality types in human-computer interaction. *Communication research*, *23*(6), 651–674 (cit. on p. 332).

Morrison, C., Cutrell, E., Grayson, M., Thieme, A., Taylor, A., Roumen, G., Longden, C., Tschiatschek, S., Faia Marques, R., & Sellen, A. (2021). Social sensemaking with ai: Designing an open-ended ai experience with a blind child. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on pp. 55, 58).

Mubin, O., Obaid, M., Jordan, P., Alves-Oliveria, P., Eriksson, T., Barendregt, W., Sjolle, D., Fjeld, M., Simoff, S., & Billinghurst, M. (2016). Towards an agenda for sci-fi inspired hci research. *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*, 1–6 (cit. on p. 71).

Muller, D. A., & Lévy, P. (2019). A design approach towards affording the trend of privacy. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 435–446 (cit. on p. 61).

Muller, Z. (2019). Algorithmic harms to workers in the platform economy: The case of uber. *Colum. JL & Soc. Probs.*, *53*, 167 (cit. on p. 234).

Murad, C. (2019). Tools to support voice user interface design. *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–5 (cit. on p. 313).

Murad, C., Candello, H., & Munteanu, C. (2023). What's the talk on vui guidelines? a meta-analysis of guidelines for voice user interface design. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–16 (cit. on pp. 306, 312, 313).

Murad, C., Munteanu, C., R. Cowan, B., & Clark, L. (2021). Finding a new voice: Transitioning designers from gui to vui design. *Proceedings of the 3rd Conference on Conversational User Interfaces*, 1–12 (cit. on p. 313).

Musan, E. (2020). *"does online dating harm your mental wellbeing?": The relationship between online dating rejection and mental wellbeing and the moderating role of self-compassion* [B.S. thesis]. University of Twente. (Cit. on p. 233).

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383 (cit. on pp. 167, 346).

Nader, K., & Lee, M. K. (2022). Folk theories and user strategies on dating apps: How users understand and manage their experience with algorithmic matchmaking. *Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I*, 445–458 (cit. on pp. 44, 45, 191, 199, 200, 229, 232, 235, 331, 340).

Nader, K., Toprac, P., Scott, S., & Baker, S. (2024). Public understanding of artificial intelligence through entertainment media. *AI & society*, *39*(2), 713–726 (cit. on p. 29).

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78 (cit. on p. 332).

Naumer, C., Fisher, K., & Dervin, B. (2008). Sense-making: A methodological perspective. *Sensemaking Workshop, CHI*, *8*, 506–513 (cit. on pp. 31–33, 35, 347, 350).

Ngo, T., Kunkel, J., & Ziegler, J. (2020). Exploring mental models for transparent and controllable recommender systems: A qualitative study. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 183–191 (cit. on p. 94).

Nilsson, N. J. (1971). Problem-solving methods in. *Artificial Intelligence*, *5* (cit. on p. 28).

Noble, S. U. (2018). Algorithms of oppression (cit. on pp. 53, 196).

Norwich, K. H. (1982). Perception as an active process. *Mathematics and Computers in Simulation*, *24*(6), 535–539 (cit. on p. 326).

O'Dwyer, R. (2018). Algorithms are making the same mistakes as humans assessing credit scores. *Retrieved April*, *17*, 2019 (cit. on pp. 38, 126, 131).

of Science, W. H. O., & Policy, T. (2020, February). *American ai initiative: Year one annual report* (tech. rep.). White House Office of Science and Technology Policy. Brussels, Belgium. https://www.whitehouse.gov/ai/ (cit. on pp. 38, 131).

Ofe-Grant, B. (2022). Rethinking oceanic-pacific methods of data collection during covid-19: Insights from the field. *International Journal of Qualitative Methods*, *21*, 16094069221111119 (cit. on p. 70).

Ogonowski, C. (2021). The living lab as a practical approach for ict design: Setup and maintenance work in domestic environments. *OPUS Siegen* (cit. on p. 17).

Okolo, C. T., Kamath, S., Dell, N., & Vashistha, A. (2021). "it cannot do all of my work": Community health worker perceptions of ai-enabled mobile health applications in rural india. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445420 (cit. on pp. 136, 154, 158, 159).

on AI (AI HLEG), H.-L. E. G. (2019, April). *Building trust in human-centric artificial intelligenc* (tech. rep.). European Commission. Brussels, Belgium. https://ec.europa.eu/jrc/communities/en/community/digitranscope/document/building-trust-human-centric-artificial-intelligence (cit. on p. 126).

Ooge, J., Kato, S., & Verbert, K. (2022). Explaining recommendations in e-learning: Effects on adolescents' trust. *27th International Conference on Intelligent User Interfaces*, 93–105. https://doi.org/10.1145/3490099.3511140 (cit. on pp. 38, 126).

Oppy, G., & Dowe, D. (2003). The turing test (cit. on p. 28).

Orosz, G., Tóth-Király, I., Bőthe, B., & Melher, D. (2016). Too many swipes for today: The development of the problematic tinder use scale (ptus). *Journal of Behavioral Addictions*, *5*(3), 518–523 (cit. on p. 199).

Orr, D. A., & Sanchez, L. (2018). Alexa, did you get that? determining the evidentiary value of data stored by the amazon® echo. *Digit. Investig.*, *24*, 72–78 (cit. on p. 254).

Palmiotto, F. (2020). Regulating algorithmic opacity in criminal proceedings: An opportunity for the eu legislator? *Maastricht Law, Faculty of Law WP*, *1* (cit. on p. 59).

Park, D., & Faraway, J. J. (1998). Sequential design for response curve estimation. *Journal of Nonparametric Statistics*, *9*(2), 155–164 (cit. on p. 74).

Parry, W. T., & Hacker, E. A. (1991). *Aristotelian logic*. Suny Press. (Cit. on pp. 79, 94).

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. (Cit. on p. 178).

Passi, S., & Jackson, S. J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW). https://doi.org/10.1145/3274405 (cit. on p. 136).

Pater, J., Fiesler, C., & Zimmer, M. (2022). No humans here: Ethical speculation on public data, unintended consequences, and the limits of institutional review. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1–13 (cit. on p. 202).

Persson, H., Åhman, H., Yngling, A. A., & Gulliksen, J. (2015). Universal design, inclusive design, accessible design, design for all: Different concepts—one goal? on the concept of accessibility—historical, methodological and philosophical aspects. *Universal access in the information society*, *14*, 505–526 (cit. on p. 281).

Pfeifle, A. (2018). Alexa, what should we do about privacy: Protecting privacy for users of voice-activated devices. *Wash. L. Rev.*, *93*, 421 (cit. on p. 254).

Phan, A., Seigfried-Spellar, K., & Choo, K.-K. R. (2021). Threaten me softly: A review of potential dating app risks. *Computers in human behavior reports*, *3*, 100055 (cit. on p. 234).

Phillips, E., Ososky, S., Grove, J., & Jentsch, F. (2011). From tools to teammates: Toward the development of appropriate mental models for intelligent robots. *Proceedings of the human factors and ergonomics society annual meeting*, *55*(1), 1491–1495 (cit. on pp. 46, 47, 90, 91).

Picard, R. W. (2000). *Affective computing*. MIT press. (Cit. on p. 343).

Pins, D. (2019). Appropriation and practices of working with voice assistants in the kitchen (cit. on p. 295).

Pins, D., & Alizadeh, F. (2021). „im wohnzimmer kriegt die schon alles mit "– sprachassistentendaten im alltag. *Alexander Boden, Timo Jakobi, Gunnar Stevens, Christian Bala (Hgg.): Verbraucherdatenschutz-Technik und Regulation zur Unterstützung des Individuums*, 1–20 (cit. on p. 291).

Pins, D., Boden, A., Essing, B., & Stevens, G. (2020). " miss understandable" a study on how users appropriate voice assistants and deal with misunderstandings. In *Proceedings of mensch und computer 2020* (pp. 349–359). (Cit. on pp. 250, 253, 258, 260, 279).

Pins, D., Jakobi, T., Boden, A., Alizadeh, F., & Wulf, V. (2021). Alexa, we need to talk: A data literacy approach on voice assistants. *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 495–507 (cit. on p. 291).

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12 (cit. on pp. 56, 117, 253, 289, 292, 294, 306, 310, 312–314, 316, 363).

Powell, A., & Henry, N. (2019). Technology-facilitated sexual violence victimization: Results from an online survey of australian adults [PMID: 27697966]. *Journal of Interpersonal Violence*, *34*, 3637–3665. https://doi.org/10.1177/0886260516672055 (cit. on p. 194).

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 218–225 (cit. on p. 91).

Powers, W. (1973). Behavior: The control of perception. (Cit. on p. 340).

Pugliese, J. (2012). *Biometrics: Bodies, technologies, biopolitics*. Routledge. (Cit. on pp. 54, 181).

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). " alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo. *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, 2853–2859 (cit. on pp. 253, 316).

Race, K. (2015). Speculative pragmatism and intimate arrangements: Online hook-up devices in gay life. *Culture, Health & Sexuality*, *17*, 496–511. https://doi.org/10.1080/13691058.2014.930181 (cit. on pp. 195, 236).

Rader, E., & Slaker, J. (2017). The importance of visibility for folk theories of sensor data. *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, 257–270 (cit. on p. 89).

Rajaonah, B., Anceaux, F., Tricot, N., & Pacaux-Lemoine, M.-P. (2006). Trust, cognitive control, and control: The case of drivers using an auto-adaptive cruise control. *Proceedings of the 13th Eurpoean Conference on Cognitive Ergonomics: Trust and Control in Complex Socio-Technical Systems*, 17–24. https://doi.org/10.1145/1274892.1274896 (cit. on p. 160).

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1). https://doi.org/10.1145/3449081 (cit. on p. 135).

Ramesh, D., Kameswaran, V., Wang, D., & Sambasivan, N. (2022). How platform-user power relations shape algorithmic accountability: A case study of instant loan platforms and financially stressed users in india. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1917–1928. https://doi.org/10.1145/3531146.3533237 (cit. on p. 136).

Randall, D. (2003). Living inside a smart home: A case study. In *Inside the smart home* (pp. 227–246). Springer. (Cit. on p. 285).

Rantatalo, O. (2013). *Sensemaking and organising in the policing of high risk situations: Focusing the swedish police national counter-terrorist unit* [Doctoral dissertation, Umeå Universitet]. (Cit. on p. 353).

Ranzini, G., & Lutz, C. (2017). Love at first swipe? explaining tinder self-presentation and motives. *Mobile Media & Communication*, *5*(1), 80–101 (cit. on p. 227).

Rao, A., Schaub, F., & Sadeh, N. (2015). What do they know about me? contents and concerns of online behavioral profiles. *arXiv preprint arXiv:1506.01675* (cit. on p. 284).

Rappaport, J. (1985). The power of empowerment language. *Social policy*, *16*(2), 15–21 (cit. on p. 49).

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. *Explainable and interpretable models in computer vision and machine learning*, 19–36 (cit. on p. 294).

Rauchfleisch, A., & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PloS one*, *15*(10), e0241045 (cit. on p. 42).

Raval, N., & Dourish, P. (2016). Standing out from the crowd: Emotional labor, body labor, and temporal labor in ridesharing. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 97–107 (cit. on pp. 40, 197).

Reckwitz, A. (2003). Grundelemente einer theorie sozialer praktiken/basic elements of a theory of social practices: Eine sozialtheoretische perspektive/a perspective in social theory. *Zeitschrift für Soziologie*, *32*(4), 282–301 (cit. on p. 349).

Reddy, A. (2020). *Researching iot through design: An inquiry into being-at-home* [Doctoral dissertation, Malmö universitet]. (Cit. on pp. 18, 55, 60, 354).

Reeves, S., Porcheron, M., Fischer, J. E., Candello, H., McMillan, D., McGregor, M., Moore, R. J., Sikveland, R., Taylor, A. S., Velkovska, J., et al. (2018). Voice-based conversational ux studies and design. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–8 (cit. on pp. 289, 312).

Register, Y., Qin, L., Baughan, A., & Spiro, E. S. (2023). Attached to "the algorithm": Making sense of algorithmic precarity on instagram. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15 (cit. on pp. 43, 343).

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J. T., Liu, F., McDonald, A., Norton, T. B., Ramanath, R., et al. (2015). Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ*, *30*, 39 (cit. on p. 255).

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? a proposal of user-centered explainable ai (cit. on pp. 57, 83, 85, 86, 119, 318).

Riveiro, M., & Thill, S. (2021). "that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems. *Artificial Intelligence*, *298*, 103507 (cit. on p. 318).

Rohde, M., Brödner, P., Stevens, G., Betz, M., & Wulf, V. (2017). Grounded design-a praxeological is research perspective. *Journal of Information Technology*, *32*(2), 163–179 (cit. on p. 16).

Rubin, J. D., Blackwell, L., & Conley, T. D. (2020). Fragile masculinity: Men, gender, and online harassment. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on pp. 191, 193, 236).

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 269–276 (cit. on p. 36).

Rwafa-Ponela, T., Price, J., Nyatela, A., Nqakala, S., Mosam, A., Erzse, A., Lalla-Edward, S. T., Hove, J., Kahn, K., Tollman, S., et al. (2022). "we were afraid": Mental health effects of the covid-19 pandemic in two south african districts. *International journal of environmental research and public health*, *19*(15), 9217 (cit. on p. 70).

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics*, *26*(5), 2749–2767 (cit. on p. 132).

Samoili, S., Lopez, C. M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F., & Delipetrev, B. (2020). *AI watch. defining artificial intelligence* (tech. rep. No. KJ-NA-30117-EN-N (online)). European Commission. Luxembourg (Luxembourg), Publications Office of the European Union. https://doi.org/10.2760/382730(online) (cit. on pp. 12, 28, 130).

Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co-design*, *4*(1), 5–18 (cit. on pp. 64, 66).

Saurwein, F., & Spencer-Smith, C. (2021). Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, *9*(4), 222–233 (cit. on pp. 54, 196).

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2021a). A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2). https://doi.org/10.1145/3476089 (cit. on pp. 154, 158, 159).

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2021b). A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–41 (cit. on p. 38).

Saxena, D., Moon, E. S.-Y., Chaurasia, A., Guan, Y., & Guha, S. (2023). Rethinking "risk" in algorithmic systems through a computational narrative analysis of casenotes in child-welfare. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581308 (cit. on pp. 136, 157).

Sayago, S. (2020). Voice assistants as learning companions: An initial exploration with computer science students. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–3 (cit. on p. 352).

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T., Chen, J. Y. C., & Hancock, P. A. (2014, July). *A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction* (tech. rep.). U.S. Army Research Laboratory. (Cit. on p. 132).

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems [PMID: 27005902]. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228 (cit. on p. 132).

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your ai: Expertise and explanations. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 240–251. https://doi.org/10.1145/3301275.3302308 (cit. on p. 160).

Schafheitle, S., Weibel, A., & Möllering, G. (2023). Inviting submissions to the special issue on trust and vulnerability (deadline 31 august 2024). (Cit. on p. 333).

Schank, R. C. (1987). What is ai, anyway? *AI magazine*, *8*(4), 59–59 (cit. on pp. 82, 83, 114, 327).

Schank, R. C. (1991). Where's the ai? *AI magazine*, *12*(4), 38–38 (cit. on p. 11).

Schegloff, E. A. (1968). Sequencing in conversational openings 1. *American anthropologist*, *70*(6), 1075–1095 (cit. on p. 97).

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *SSRN Electron. J.*, *29*, 48 (cit. on pp. 38, 131).

Schildt, H., Mantere, S., & Cornelissen, J. (2020). Power in sensemaking processes. *Organization Studies*, *41*(2), 241–265 (cit. on pp. 50, 51, 341).

Schmude, T., Koesten, L., Möller, T., & Tschiatschek, S. (2023). On the impact of explanations on understanding of algorithmic decision-making. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 959–970. https://doi.org/10.1145/3593013.3594054 (cit. on pp. 37, 130, 135).

Schneider, H., Eiband, M., Ullrich, D., & Butz, A. (2018). Empowerment in hci-a survey and framework. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on pp. 49, 52, 53, 344, 345).

Schöffer, J., Machowski, Y., & Kühl, N. (2021). A study on fairness and trust perceptions in automated decision making. *Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA*, *2903*, 170005 (cit. on p. 134).

Schön, D. A. (2017). *The reflective practitioner: How professionals think in action*. Routledge. (Cit. on p. 67).

Schönherr, L., Golla, M., Eisenhofer, T., Wiele, J., Kolossa, D., & Holz, T. (2020). Unacceptable, where is my privacy? exploring accidental triggers of smart speakers. *arXiv preprint arXiv:2008.00508* (cit. on pp. 250, 254, 257, 279, 291).

Schutz, A. (1970). Concept and theory formation in the social sciences. *Sociological theory and philosophical analysis*, 1–19 (cit. on p. 329).

Schwarz, A., & Unselt, J. J. (2024). Rage against the machine? framing societal threat and efficacy in youtube videos about artificial intelligence. *Risk Analysis* (cit. on p. 29).

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). " hey alexa, what's up?" a mixed-methods studies of in-home conversational agent usage. *Proceedings of the 2018 designing interactive systems conference*, 857–868 (cit. on pp. 117, 250, 253, 254).

Scott, I. A., Carter, S. M., & Coiera, E. (2021). Exploring stakeholder attitudes towards ai in clinical practice. *BMJ Health & Care Informatics*, *28*(1), 7. https://doi.org/10.1136/bmjhci-2021-100450 (cit. on p. 133).

Seberger, J. S., Shklovski, I., Swiatek, E., & Patil, S. (2022). Still creepy after all these years: The normalization of affective discomfort in app use. *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–19 (cit. on pp. 308, 312).

Seong, Y., Bisantz, A. M., & Bisantz, A. M. (2002). Judgment and trust in conjunction with automated decision aids: A theoretical model and empirical investigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(3), 423–427. https://doi.org/10.1177/154193120204600344 (cit. on p. 134).

Services, A. F. (2019, January). *Responsible ai: A framework for building trust in your ai solutions* (tech. rep.). Accenture. https://www.accenture.com/us-en/insights/us-federal-government/ai-is-ready-are-we (cit. on pp. 38, 131).

Seymour, W. (2019). Privacy Therapy with Aretha: What If Your Firewall Could Talk? *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*, 1–6. https://doi.org/10.1145/3290607.3308449 (cit. on pp. 61, 258, 259).

Seymour, W., & Such, J. (2023). Ignorance is bliss? the effect of explanations on perceptions of voice assistants. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–24 (cit. on p. 312).

Seymour, W., & Van Kleek, M. (2021). Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–16 (cit. on pp. 57, 289, 295, 309).

Sharabi, L. L. (2021). Exploring how beliefs about algorithms shape (offline) success in online dating: A two-wave longitudinal investigation. *Communication Research*, *48*(7), 931–952 (cit. on pp. 44, 191, 199).

Shaw, F. (2016). "bitch i said hi": The bye felipe campaign and discursive activism in mobile dating apps. *Social Media+ Society*, *2*, 2056305116672889 (cit. on pp. 193, 194, 236, 237).

Shin, J., Park, Y., & Lee, D. (2018). Who will be smart home users? an analysis of adoption and diffusion of smart homes. *Technological Forecasting and Social Change*, *134*, 246–253 (cit. on p. 266).

Shneiderman, B. (2022). *Human-centered ai*. Oxford University Press. (Cit. on pp. 13, 50, 361).

Siegert, I. (2020). "alexa in the wild"–collecting unconstrained conversations with a modern voice assistant in a public environment. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 615–619 (cit. on p. 295).

Siegert, I. (2021). Effects of prosodic variations on accidental triggers of a commercial voice assistant. *Interspeech*, 1674–1678 (cit. on p. 291).

Siles, I., Segura-Castillo, A., Solís, R., & Sancho, M. (2020). Folk theories of algorithmic recommendations on spotify: Enacting data assemblages in the global south. *Big Data & Society*, *7*(1), 2053951720923377 (cit. on pp. 40, 197).

Silverstone, R., & Haddon, L. (1996). Design and the Domestication of Information and Communication Technologies: Technical Change and Everyday Life (cit. on pp. 257, 278).

Simon, G. (2021, July). Dating app insiders remain 'highly concerned' about user security, according to a recent survey. https://www.forbes.com/sites/traversmark/2021/07/15/dating-app-insiders-remain-highly-concerned-about-user-security-according-to-a-recent-survey/?sh=2841415f4ed7 (cit. on pp. 195, 198, 240).

Simpson, E., Hamann, A., & Semaan, B. (2022). How to tame" your" algorithm: Lgbtq+ users' domestication of tiktok. *Proceedings of the ACM on Human-computer Interaction*, *6*(GROUP), 1–27 (cit. on pp. 234, 238).

Simpson, E., & Semaan, B. (2021). For you, or for" you"? everyday lgbtq+ encounters with tiktok. *Proceedings of the ACM on human-computer interaction*, *4*(CSCW3), 1–34 (cit. on p. 234).

Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., & Perer, A. (2023). Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581075 (cit. on p. 153).

Smith, J. J., Buhayh, A., Kathait, A., Ragothaman, P., Mattei, N., Burke, R., & Voida, A. (2023). The many faces of fairness: Exploring the institutional logics of multistakeholder microlending recommendation. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1652–1663. https://doi.org/10.1145/3593013.3594106 (cit. on p. 135).

Smith, J. A. (2024). Qualitative psychology: A practical guide to research methods (cit. on p. 97).

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. https://doi.org/https://doi.org/10.1006/obhd.2000.2926 (cit. on pp. 38, 131).

Sokol, K., & Flach, P. (2020). One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, *34*(2), 235–250 (cit. on p. 318).

Solomonoff, R. J. (1985). The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, *5*(2), 149–153 (cit. on p. 10).

Søndergaard, M. L. J., & Hansen, L. K. (2018). Intimate futures: Staying with the trouble of digital personal assistants through design fiction. *Proceedings of the 2018 designing interactive systems conference*, 869–880 (cit. on pp. 254, 265).

Spitale, M., & Garzotto, F. (2020). Towards empathic conversational interaction. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–4 (cit. on p. 313).

Starbuck, W. H., & Milliken, F. J. (1988). Executives' perceptual filters: What they notice and how they make sense. *JAI Press* (cit. on p. 15).

Stevens, G. (2009). Understanding and designing appropriation infrastructures. *Institute for Information Systems* (cit. on p. 350).

Stevens, G., & Bossauer, P. (2017). Dealing with personal data in the age of big data economies. *Zeitschrift für Geistiges Eigentum= Intellectual Property Journal*, *9*(3), 266–278 (cit. on p. 250).

Stevens, G., & Pipek, V. (2018). Making use: Understanding, studying, and supporting appropriation. In *Wulf, pipek et al.(eds.): Socio-informatics: A practice-based perspective on the design and use of it artifacts* (pp. 139–176). Oxford University Press. (Cit. on pp. 337, 356).

Stevens, G., Pipek, V., & Wulf, V. (2009). Appropriation infrastructure: Supporting the design of usages. *International Symposium on End User Development*, 50–69 (cit. on pp. 16, 356).

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., et al. (2022). Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence. *arXiv preprint arXiv:2211.06318* (cit. on p. 82).

Stonier, T. (1992). The evolution of machine intelligence. In *Beyond information: The natural history of intelligence* (pp. 107–133). Springer. (Cit. on p. 28).

Strauss, A. L., Corbin, J. M., Niewiarra, S., & Legewie, H. (1996). *Grounded theory: Grundlagen qualitativer sozialforschung*. Beltz, Psychologie-Verlag-Union Weinheim. (Cit. on p. 263).

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge university press. (Cit. on pp. 288, 319).

Sumter, S. R., & Vandenbosch, L. (2019). Dating gone mobile: Demographic and personality-based correlates of using smartphone-based dating applications among emerging adults. *New media & society*, *21*(3), 655–673 (cit. on p. 204).

Sundar, S. S., Kim, J., & Cho, E. (2019). Where in the cloud is my data? location and brand effects on trust in cloud services. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6 (cit. on p. 250).

Sutherland, S. C., Harteveld, C., & Young, M. E. (2015). The role of environmental predictability and costs in relying on automation. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2535–2544. https://doi.org/10.1145/2702123.2702609 (cit. on p. 160).

Swipe right: The emergence of dating-app facilitated sexual assault. a descriptive retrospective audit of forensic examination caseload in an australian metropolitan service. (2020). *Forensic Science, Medicine and Pathology*, *16*, 71–77. https://doi.org/10.1007/s12024-019-00201-7 (cit. on p. 194).

Tabassum, M., Kosinski, T., & Lipford, H. R. (2019). "I don't own the data": End User Perceptions of Smart Home Device Data Practices and Risks. Retrieved January 24, 2021, from https://www.usenix.org/conference/soups2019/presentation/tabassum (cit. on pp. 253–255, 279).

Taskforce, A. (2019, May). *Report of estonia's ai taskforce* (tech. rep.). Republic of Estonia Government Office, Republic of Estonia Ministry of Economic Affairs, and Communications. Estonia. https://ec.europa.eu/knowledge4policy/ai-watch/estonia-ai-strategy-report (cit. on pp. 38, 131).

Taylor, A. S., Harper, R., Swan, L., Izadi, S., Sellen, A., & Perry, M. (2007). Homes that make us smart. *Personal and Ubiquitous Computing*, *11*, 383–393 (cit. on p. 354).

Teo, W. L., Lee, M., & Lim, W.-S. (2017). The relational activation of resilience model: How leadership activates resilience in an organizational crisis. *Journal of Contingencies and Crisis Management*, *25*(3), 136–147 (cit. on p. 353).

Teyssier, M., Bailly, G., Lecolinet, É., & Pelachaud, C. (2017). Survey and perspectives of social touch in hci. *Proceedings of the 29th Conference on l'Interaction Homme-Machine*, 93–104 (cit. on p. 17).

the Council, E. P. a. (2018, May). General Data Protection Regulation (GDPR). https://gdpr-info.eu/ (cit. on p. 280).

Thomas, J. B., Clark, S. M., & Gioia, D. A. (1993). Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. *Academy of Management journal*, *36*(2), 239–270 (cit. on p. 15).

Thompson, R. (2022). Bumble makes cyberflashing detection tool available as open-source code. *Mashable*. https://mashable.com/article/bumble-cyberflashing-private-detector-open-source (cit. on pp. 195, 198, 240).

Toff, B., & Nielsen, R. K. (2018). "i just google it": Folk theories of distributed discovery. *Journal of communication*, *68*(3), 636–657 (cit. on pp. 41, 197).

Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022). Capable but amoral? comparing ai and human expert collaboration in ethical decision making. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517732 (cit. on p. 160).

Tolmie, P., Crabtree, A., Rodden, T., Colley, J., & Luger, E. (2016). "this has to be the cats" personal data legibility in networked sensing systems. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 491–502 (cit. on p. 251).

Tolmie, P., Crabtree, A., Rodden, T., Colley, J. A., & Luger, E. A. (2016). "This has to be the cats" - Personal Data Legibility in Networked Sensing Systems. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, 490–501. https://doi.org/10.1145/2818048.2819992 (cit. on pp. 60, 256–258).

Toma, C. L. (2022). Online dating and psychological wellbeing: A social compensation perspective. *Current Opinion in Psychology*, *46*, 101331 (cit. on p. 233).

Tong, S. T., Hancock, J. T., & Slatcher, R. B. (2016). The influence of technology on romantic relationships: Understanding online dating. *Social Computing and Social Media: 8th International Conference, SCSM 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17–22, 2016. Proceedings 8*, 162–173 (cit. on pp. 44, 191).

Trewin, S., Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N., & Manser, E. (2019). Considerations for ai fairness for people with disabilities. *AI Matters*, *5*(3), 40–63 (cit. on p. 50).

Tsoukas, H. (2017). Don't simplify, complexify: From disjunctive to conjunctive theorizing in organization and management studies. *Journal of management studies*, *54*(2), 132–153 (cit. on p. 355).

Tufekci, Z. (2015). Algorithmic harms beyond facebook and google: Emergent challenges of computational agency. *Colo. Tech. LJ*, *13*, 203 (cit. on pp. 53, 196).

Turing, A. (2009). *Computing machinery and intelligence*. Springer. (Cit. on pp. 32, 176, 330).

Turner, J. R., Allen, J., Hawamdeh, S., & Mastanamma, G. (2023). The multifaceted sensemaking theory: A systematic literature review and content analysis on sensemaking. *Systems*, *11*(3), 145 (cit. on pp. 31–33, 35, 348).

Usmani, U. A., Happonen, A., & Watada, J. (2023). Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. *2023 5th international congress on human-computer interaction, optimization and robotic applications (HORA)*, 1–7 (cit. on pp. 50, 53, 361).

Uthirapathy, S., & Sandanam, D. (2024). Aspect-based sentimental analysis on social media data using deep learning methods. *Proceedings of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security, IACIDS 2023, 23-25 November 2023, Lavasa, Pune, India* (cit. on p. 72).

Valdez, A. C., Heine, M., Franke, T., Jochems, N., Jetter, H.-C., & Schrills, T. (2024). The european commitment to human-centered technology: The integral role of hci in the eu ai act's success. *arXiv preprint arXiv:2402.14728* (cit. on p. 53).

Valentine, J. L., Miles, L. W., Hamblin, K. M., & Gibbons, A. W. (n.d.). Dating app facilitated sexual assault: A retrospective review of sexual assault medical forensic examination charts [PMID: 36310506]. *Journal of Interpersonal Violence*, *0*, 08862605221130390. https://doi.org/10.1177/08862605221130390 (cit. on pp. 190, 193).

van Maanen, P.-P., Wisse, F., van Diggelen, J., & Beun, R.-J. (2011). Effects of reliance support on team performance by advising and adaptive autonomy. *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 02*, 280–287. https://doi.org/10.1109/WI-IAT.2011.117 (cit. on p. 160).

Van Kleek, M., Liccardi, I., Binns, R., Zhao, J., Weitzner, D. J., & Shadbolt, N. (2017). Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5208–5220. https://doi.org/10.1145/3025453.3025556 (cit. on pp. 61, 259).

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings of the national conference on artificial intelligence*, 900–907 (cit. on p. 86).

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *Br J Soc Psychol*, *44*(Pt 3), 443–461 (cit. on pp. 38, 131).

Van Wynsberghe, A. (2021). Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, *1*(3), 213–218 (cit. on p. 14).

Varanasi, R. A., & Goyal, N. (2023). "it is currently hodgepodge": Examining ai/ml practitioners' challenges during co-production of responsible ai values. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580903 (cit. on p. 134).

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174014 (cit. on p. 134).

Velkova, J., & Kaun, A. (2021). Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society*, *24*(4), 523–540 (cit. on p. 232).

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2). https://doi.org/10.1145/3476068 (cit. on pp. 131, 144, 156, 157).

Villani, C., Bonnet, Y., Rondepierre, B., et al. (2018). *For a meaningful artificial intelligence: Towards a french and european strategy*. Conseil national du numérique. (Cit. on pp. 38, 131).

Vitale, F., Chen, J., Odom, W., & McGrenere, J. (2020). Data dashboard: Exploring centralization and customization in personal data curation. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 311–326 (cit. on pp. 259, 281).

Voit, A., Niess, J., Eckerth, C., Ernst, M., Weingärtner, H., & Woźniak, P. W. (2020). 'it's not a romantic relationship': Stories of adoption and abandonment of smart speakers at home. *Proceedings of the 19th international conference on mobile and ubiquitous multimedia*, 71–82 (cit. on pp. 250, 253, 288).

Voskobojnikov, A., Wiese, O., Mehrabi Koushki, M., Roth, V., & Beznosov, K. (2021). The u in crypto stands for usable: An empirical study of user experience with mobile cryptocurrency wallets. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14 (cit. on pp. 201–203).

Walker, A. M., & DeVito, M. A. (2020). "'more gay'fits in better": Intracommunity power dynamics and harms in online lgbtq+ spaces. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15 (cit. on pp. 234, 235, 341).

Wang, X., & Yin, M. (2023). Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581366 (cit. on pp. 39, 126).

Wang, X., & Du, X. (2018). Why does advice discounting occur? the combined roles of confidence and trust. *Front Psychol*, *9*, 2381 (cit. on pp. 38, 131).

Ward, J. (2017). What are you doing on tinder? impression management on a matchmaking mobile app. *Information, Communication & Society*, *20*(11), 1644–1659 (cit. on p. 232).

Warren, S. D., & Brandeis, L. D. (1890). The Right to Privacy. *Harvard Law Review*, *4*(5), 193–220. https://doi.org/10.2307/1321160 (cit. on p. 256).

Waterman, R. H. (1990). *Adhocracy: The power to change* (Vol. 3). Knoxville, Tenn.: Whittle Direct Books. (Cit. on p. 15).

Weick, K. E. (1988). Enacted sensemaking in crisis situations [1]. *Journal of management studies*, *25*(4), 305–317 (cit. on p. 337).

Weick, K. E. (2009). *Making sense of the organization, volume 2: The impermanent organization*. John Wiley & Sons. (Cit. on p. 337).

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization science*, *16*(4), 409–421 (cit. on pp. 15, 34, 328, 336, 338, 340, 353).

Weick, K. E., & Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage publications Thousand Oaks, CA. (Cit. on pp. 14, 51, 326, 335, 337, 339, 340, 351, 360).

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7–9 (cit. on p. 58).

Weizenbaum, J. (1983). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *26*(1), 23–28 (cit. on p. 12).

Wheale, N. (1991). Recognising a'human-thing': Cyborgs, robots and replicants in philip k. dick's" do androids dream of electric sheep?" and ridley scott's" blade runner". *Critical Survey*, 297–304 (cit. on pp. 173, 183).

Whitten, A., & Tygar, J. D. (1999). Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. *Proceedings of the 8th USENIX Security Symposium*, 169–184 (cit. on pp. 61, 259).

Whitty, M. T. (2015). Anatomy of the online dating romance scam. *Security Journal*, *28*, 443–455. https://doi.org/10.1057/sj.2012.57 (cit. on pp. 194, 237).

Wickramasinghe, C. S., Marino, D. L., Grandio, J., & Manic, M. (2020). Trustworthy ai development guidelines for human system interaction. *2020 13th International Conference on Human System Interaction (HSI)*, 130–136 (cit. on p. 14).

Widder, D. G., Zhen, D., Dabbish, L., & Herbsleb, J. (2023). It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them? *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 467–479. https://doi.org/10.1145/3593013.3594012 (cit. on p. 135).

Wilhelm, M., Otten, T., Schwaetzer, E., & Schumacher, K. (2022). Keep on smiling: An investigation of the influence of the use of emoticons by chatbots on user satisfaction. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–6 (cit. on p. 313).

Wilson, C. (2013). *User interface inspection methods: A user-centered design method*. Newnes. (Cit. on p. 72).

Winter, S. G. (2003). Mistaken perceptions: Cases and consequences. *British Journal of Management*, *14*(1), 39–44 (cit. on p. 339).

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174230 (cit. on p. 135).

Worthy, P., Matthews, B., & Viller, S. (2016a). Trust Me: Doubts and Concerns Living with the Internet of Things. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16*, 427–434. https://doi.org/10.1145/2901790.2901890 (cit. on p. 257).

Worthy, P., Matthews, B., & Viller, S. (2016b). Trust me: Doubts and concerns living with the internet of things. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 427–434 (cit. on p. 250).

Wu, Y., & Kelly, R. M. (2020). Online dating meets artificial intelligence: How the perception of algorithmically generated profile text impacts attractiveness and trust. *Proceedings of the 32nd Australian Conference on Human-Computer Interaction*, 444–453 (cit. on pp. 40, 190).

Wulf, V., Müller, C., Pipek, V., Randall, D., Rohde, M., & Stevens, G. (2015). Practice-based computing: Empirically grounded conceptualizations derived from design case studies. In *Designing socially embedded technologies in the real-world* (pp. 111–150). Springer London. https://doi.org/10.1007/978-1-4471-6720-4_7 (cit. on pp. 64, 65).

Wulf, V., Pipek, V., Randall, D., Rohde, M., Schmidt, K., & Stevens, G. (Eds.). (2018). *Socio-informatics: A practice-based perspective on the design and use of IT artifacts* (First edition) [OCLC: on1000434665]. Oxford University Press. (Cit. on p. 65).

Wulf, V., Rohde, M., Pipek, V., & Stevens, G. (2011). Engaging with practices: Design case studies as a research framework in cscw. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 505–512. https://doi.org/10.1145/1958824.1958902 (cit. on p. 65).

Xiao, J., Stasko, J., & Catrambone, R. (2007). The role of choice and customization on users' interaction with embodied conversational agents: Effects on perception and performance. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1293–1302. https://doi.org/10.1145/1240624.1240820 (cit. on p. 160).

Xu, W., Gao, Z., & Dainoff, M. (2023). An hcai methodological framework: Putting it into action to enable human-centered ai. *arXiv preprint arXiv:2311.16027* (cit. on p. 14).

Xu, Y., & Warschauer, M. (2020). What are you talking to?: Understanding children's perceptions of conversational agents. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13 (cit. on pp. 47, 48, 91, 92, 115).

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/https://doi.org/10.1006/obhd.2000.2909 (cit. on pp. 37, 130).

Yao, Y., Basdeo, J. R., Kaushik, S., & Wang, Y. (2019). Defending My Castle: A Co-Design Study of Privacy Mechanisms for Smart Homes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12. https://doi.org/10.1145/3290605.3300428 (cit. on pp. 61, 258).

Yao, Y., Huang, Y., & Wang, Y. (2019). Unpacking people's understandings of bluetooth beacon systems-a location-based iot technology (cit. on pp. 90, 94, 115).

Yildirim, N., Pushkarna, M., Goyal, N., Wattenberg, M., & Viégas, F. (2023). Investigating how practitioners use human-ai guidelines: A case study on the people + ai guidebook. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580900 (cit. on p. 134).

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300509 (cit. on pp. 39, 126).

Ylirisku, S., Halttunen, V., Nuojua, J., & Juustila, A. (2009). Framing design in the third paradigm. *Proceedings of the SIGCHI conference on human factors in computing systems*, 1131–1140 (cit. on pp. 67, 351).

Yu, K.-H., & Kohane, I. S. (2019). Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety*, *28*(3), 238–241. https://doi.org/10.1136/bmjqs-2018-008551 (cit. on pp. 38, 131).

Yuan, C. W., Bi, N., Lin, Y.-F., & Tseng, Y.-H. (2023). Contextualizing user perceptions about biases for human-centered explainable artificial intelligence. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15 (cit. on p. 38).

Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., & Bozzon, A. (2023). Disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581161 (cit. on pp. 134, 155).

Yurrita, M., Murray-Rust, D., Balayn, A., & Bozzon, A. (2022). Towards a multi-stakeholder value-based assessment framework for algorithmic systems. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 535–563. https://doi.org/10.1145/3531146.3533118 (cit. on pp. 133, 159).

Zamenopoulos, T., & Alexiou, K. (2018). *Co-design as collaborative research*. Bristol University/AHRC Connected Communities Programme. (Cit. on p. 297).

Zargham, N., Reicherts, L., Bonfert, M., Voelkel, S. T., Schoening, J., Malaka, R., & Rogers, Y. (2022). Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. *Proceedings of the 4th conference on conversational user interfaces*, 1–14 (cit. on pp. 288, 317).

Zeng, E., Mare, S., & Roesner, F. (2017). End user security and privacy concerns with smart homes. *thirteenth symposium on usable privacy and security (SOUPS 2017)*, 65–80 (cit. on pp. 279, 280).

Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874* (cit. on pp. 30, 84, 85).

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852 (cit. on pp. 39, 126).

Zhou, Z., Li, Z., Zhang, Y., & Sun, L. (2022). Transparent-ai blueprint: Developing a conceptual tool to support the design of transparent ai agents. *International Journal of Human–Computer Interaction*, *38*(18-20), 1846–1873 (cit. on p. 13).

Zimmerman, M. A. (1995). Psychological empowerment: Issues and illustrations. *American journal of community psychology*, *23*, 581–599 (cit. on pp. 51, 326, 342, 361).

Žižek, S. (2021). Blade runner 2049: A view of post-human capitalism. *Lacanian Perspectives on Blade Runner 2049*, 41–51 (cit. on p. 174).

Zytko, D., & Furlo, N. (2023). Online dating as context to design sexual consent technology with women and lgbtq+ stakeholders. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580911 (cit. on p. 195).

Zytko, D., Furlo, N., Carlin, B., & Archer, M. (2021). Computer-mediated consent to sex: The context of tinder. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–26 (cit. on pp. 195, 236).

Zytko, D., Grandhi, S. A., & Jones, Q. (2015). Frustrations with pursuing casual encounters through online dating. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 1935–1940. https://doi.org/10.1145/2702613.2732905 (cit. on p. 195).

Zytko, D., Grandhi, S. A., & Jones, Q. (2014). Impression management struggles in online dating. *Proceedings of the 18th international conference on supporting group work*, 53–62. https://doi.org/10.1145/2660398.2660410 (cit. on pp. 195, 231).

Zytko, D., Regalado, V., Furlo, N., Grandhi, S. A., & Jones, Q. (2020). Supporting women in online dating with a messaging interface that improves their face-to-face meeting decisions. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 1–30. https://doi.org/10.1145/3415208 (cit. on p. 195).

Webpages

9241-11:2018(en), I. (2018). *Ergonomics of human-system interaction – part 11: Usability: Definitions and concepts*. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en (cit. on p. 116).

Agarwal, S. (2023). *Death to captchas*. https://www.technologyreview.com/2023/10/24/1081139/captchas-ai-websites-computing/ (cit. on pp. 42, 171).

AIAAIC. (2023). *What's driving ai and algorithmic incidents and controversies*. https://www.aiaaic.org/home (cit. on pp. 54, 197).

Amazon. (2024). *Ask alexa to tell you why she just did something*. https://www.cnet.com/home/smart-home/amazon-alexa-adds-new-commands-to-tamp-down-privacy-concerns/ (cit. on p. 289).

Association, B. S. (2016). *One in three believe that the rise of artificial intelligence is a threat to humanity*. https://www.britishscienceassociation.org/news/rise-of-artificial-intelligence-is-a-threat-to-humanity (cit. on p. 85).

Bartlett, M. (2020). *How tinder's algorithm is micromanaging your dating life*. https://thespinoff.co.nz/tech/18-07-2020/how-tinders-algorithm-is-micromanaging-your-dating-life (cit. on p. 199).

Brown, A. (2018). *'least desirable'? how racial discrimination plays out in online dating*. (Cit. on p. 233).

Davies, J. (2020). *Arm 2020 global ai survey*. https://newsroom.arm.com/blog/read-arm-2020-global-ai-survey (cit. on pp. 30, 83, 85, 86, 115).

Dent, S. (2019). *Tinder ditches its hidden desirability scores*. https://www.engadget.com/2019-03-18-tinder-dumps-desirability-scores.html?guccounter=1 (cit. on p. 199).

Documentation, B. S. (2022). *Beautiful soup documentation*. https://beautiful-soup-4.readthedocs.io/en/latest/ (cit. on p. 201).

Eindhoven University, o. T. (2024). *Future everyday*. https://www.tue.nl/en/research/research-groups/future-everyday (cit. on p. 17).

Elliott, T. (2017). *What is artificial intelligence called?!* https://timoelliott.com/blog/2017/06/what-is-artificial-intelligence-called.html/comment-page-1 (cit. on pp. 10, 12).

GoodAI. (2019). *Understanding the public perception of ai*. https://medium.com/goodai-news/understanding-the-public-perception-of-ai-a14b0e6b6154 (cit. on pp. 82, 83).

Google. (2024). *Tough on bots easy on humans*. https://www.google.com/recaptcha/intro/invisible.html?ref=producthunt (cit. on pp. 42, 170, 171).

Greenberg, A. (2014). *Google can now tell you're not a robot with just one click*. https://www.wired.com/2014/12/google-one-click-recaptcha/ (cit. on p. 179).

HCI Lab, S. U. (2021). *Hci sorbonne*. https://hci.isir.upmc.fr/research/ (cit. on p. 17).

Human-Computer Interaction Institute, C. M. U. (2021). *Research*. https://hcii.cmu.edu/research (cit. on p. 17).

Identities, S. (2024). *Research*. https://www.speculativeidentities.com/research/voight-kampff (cit. on p. 173).

Junkie, T. (2020). *How to calculate and increase your tinder elo score*. https://social.techjunkie.com/calculate-increase-tinder-elo-score/ (cit. on pp. 44, 199).

Kantor, J. (2022). *Times article changes a starbucks policy, fast*. https://archive.nytimes.com/www.nytimes.com/times-insider/2014/08/22/times-article-changes-a-policy-fast/ (cit. on pp. 53, 196).

Leal, S. (2016). *Oh, okay: Tinder is secretly ranking your desirability*. https://www.marieclaire.com/sex-love/news/a17973/tinder-secret-dating-score/ (cit. on p. 199).

MidJourney. (2022). *Midjourney – an ai for generating art*. Retrieved November 29, 2024, from https://www.midjourney.com/ (cit. on p. 11).

Mikonoss, P. (2019). *The secret rules of tinder – how to improve your score & get more matches*. https://thefrisky.com/the-secret-rules-of-tinder-how-to-improve-your-score-get-more-matches/ (cit. on p. 199).

Noessel, C. (2019). *8 reasons the voight-kampff machine is shit (and a redesign to fix it)*. https://scifiinterfaces.com/2019/12/10/8-reasons-the-voight-kampff-machine-is-shit-and-a-redesign-to-fix-it/ (cit. on p. 174).

Notopoulos, K. (2016). *The dating app that knows you secretly aren't into guys from other races*. https://www.buzzfeednews.com/article/katienotopoulos/coffee-meets-bagel-racial-preferences (cit. on p. 190).

OpenAI. (2022). *Introducing chatgpt*. Retrieved November 29, 2024, from https://openai.com/index/chatgpt/ (cit. on p. 11).

Pegasystems. (2019). *What consumers really think about ai: A global study*. https://www.pega.com/ai-survey (cit. on pp. 30, 83, 85, 359).

Richard Nieva, B. F. R. (2019). *Amazon alexa adds new commands to tamp down privacy concerns*. https://www.cnet.com/home/smart-home/amazon-alexa-adds-new-commands-to-tamp-down-privacy-concerns/ (cit. on p. 289).

Santiago, E. (2023). *How ai perceptions have changed in the last decade*. https://blog.hubspot.com/marketing/artificial-intelligence-is-here (cit. on pp. 30, 36, 85).

Schwab, K. (2019). *Google's new recaptcha has a dark side*. https://www.fastcompany.com/90369697/googles-new-recaptcha-has-a-dark-side (cit. on p. 179).

Smith, K. L. (2020). *What is the tinder elo score? the 'desirability rating' is no longer used by the dating app*. https://www.popbuzz.com/life/news/tinder-elo-score/ (cit. on pp. 44, 199).

Staff, R. (2021). *Match tops sales estimates as tinder, hinge keep sparks flying*. https://www.reuters.com/article/us-match-group-results-idUSKBN2A22V1 (cit. on p. 201).

the Future of Life Institute (FLI). (2024). *The eu artificial intelligence act*. https://artificialintelligenceact.eu/ (cit. on pp. 11, 346, 361).

Tinder. (2023). *Swipe right*. https://tinder.com/ (cit. on p. 191).

Tinderpress. (2022a). *Powering tinder® — the method behind our matching*. https://www.tinderpressroom.com/powering-tinder-r-the-method-behind-our-matching/ (cit. on pp. 45, 199, 238, 239).

Tinderpress. (2022b). *Privacy faqs*. https://www.help.tinder.com/hc/en-us/articles/5349453268877-Privacy-FAQs#h_01G63PGMBC1RQ5ESM3RC72ZF0M (cit. on pp. 239, 339).

VIDAselect. (2022). *Best tinder bios for guys: 6 examples that will make her swipe right*. https://www.vidaselect.com/best-tinder-bios/ (cit. on p. 199).

Waters, S. (2021). *I'm not a robot! so why won't captchas believe me?* https://www.wired.com/story/im-not-a-robot-why-captchas-hard-to-solve/ (cit. on pp. 42, 43, 179, 180).

zoovu. (2018). *The 6 stages in the evolution of ai and customer experience*. https://medium.com/@zoovu/the-6-stages-in-the-evolution-of-ai-and-customer-experience-15e6467e2bfc (cit. on p. 82).