

UMEÅ UNIVERSITY

Digital Agents as Social Actors: From Theory to Agent-Supported Lifestyle Changes

Kaan Kilic

Doctoral Thesis, December 2026
Department of Computing Science
Umeå University
Sweden

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

kaank@cs.umu.se

Copyright © 2026 by authors

ISBN 978-91-6850-228-1
ISSN 0348-0542
UMINF 26.18

Printed by Scandinavian Print Group, Hägersten, 2026

Abstract

Most digital interventions for behavior change to improve health employ methods to support individuals such as providing research-based rationales, setting goals, sending reminders about goals, and tracking progress. However people differ, both between individuals and within the same individual over time, in how much support or challenge they want or need, and existing frameworks offer limited ways to model this variation or to act on it. Person-tailored support with artificial intelligence (AI) also brings questions on privacy, security, motivation, and how the AI-based support is perceived.

This thesis investigates how a digital companion can both support and challenge a person in lifestyle change while staying aligned with that person's motives, readiness, and shifting preferences. Activity Theory and Argumentation Theory form the theoretical framework to develop the companion's underlying models: agent roles, dialogue protocol, and two novel argumentation schemes for support and challenge. Participatory design is the overarching design methodology engaging an interdisciplinary team of health and domain experts, end users, and the research team iteratively design, prototype and refine subsequent versions of agent-based coaching systems. Across nine empirical user studies, data were collected through questionnaires, participant ratings, semi-structured interviews, use of prototypes, and observations during multi-week onboarding periods. Questionnaire data were analyzed quantitatively, and the interview and observation data qualitatively, through thematic analysis.

This thesis makes four contributions. First, an activity-theoretical reframing distinguishing exogenous (healthcare driven) from endogenous (person driven) behavior change, moving beyond persuasion and adherence to an intervention. Second, a formal model of supportive and challenging dialogue built on two novel argumentation schemes Argument from Position to Support and Argument from Position to Create Tension, each specified with critical questions and a dialogue protocol, which structures the grounds upon which an agent can act in a supportive or challenging manner. Third, a person-tailored, multi-domain coaching system built through participatory design, together with embedded clustered behavior change techniques. The approach is extended to a swarm of collaborating agents and single agents, whose role the person can choose or write, that supports recovery activities to manage every-

day stress. Finally, empirically grounded design implications for onboarding and role personalization. Together, these provide a theoretical foundation, empirical insights, and a functional system of AI agents that tailor support and challenge to the individual; further identifying multi-agent coordination as a viable path toward personalized coaching for improved health through lifestyle change.

To summarize, the thesis contributes a deepened understanding of how an AI-based agent and a team of agents can be adapted to people's individual and varying needs.

Sammanfattning

De flesta digitala interventioner för beteendeförändring i syfte att förbättra hälsa använder metoder för att stödja personer i att ändra beteenden och vanor, såsom ge motiveringar grundade i forskning, sätta mål, påminna om uppsatta mål och följa upp resultatet av dessa. En stor utmaning är dock att människor skiljer sig åt – både sinsemellan och över tid – när det gäller hur mycket stöd eller utmaning de önskar eller behöver från ett digitalt stöd. Personanpassade stöd behöver också utformas utifrån hur teknologi baserad på artificiell intelligens (AI) bör utformas utifrån integritet, säkerhet, hur motivation påverkas och hur AI-baserat stöd upplevs av personen.

Denna avhandling undersöker hur en digital följeslagare baserad på AI-teknologi, en AI-agent, kan både stödja och utmana en person vid livsstilsförändring, samtidigt som den förblir i samklang med personens motiv, beredskap och skiftande preferenser.

Verksamhetsteori och argumentationsteori formar det teoretiska ramverket för utveckling av den digitala följeslagarens modeller: agentens roller, dialogprotokoll och två nya argumentationsscheman för stöd respektive utmaning. Deltagandedesignmetodik tillämpas, där hälso- och domänexperter, slutanvändare och forskare iterativt tar fram och förfinar versioner av agentbaserade coaching-system. Inom ramen för nio empiriska användarstudier samlades data in genom enkäter, deltagarskattningar, semistrukturerade intervjuer, användning av prototyper samt observationer under flera veckor långa introduktionsperioder. Enkätdata analyserades kvantitativt, medan data från intervjuer och observationer analyserades kvalitativt genom tematisk analys.

Avhandlingen presenterar följande fyra huvudsakliga bidrag. För det första bidrar arbetet med och utgår ifrån ett verksamhetsteoretiskt perspektiv som skiljer på exogena (hälso- och sjukvårdsdrivna) och endogena (persondrivna) beteendeförändringar. Det andra bidraget är en formell modell för stödjande och utmanande dialog som består av två nya argumentationsscheman och ett dialogprotokoll som strukturerar på vilka grunder en agent kan agera stödjande respektive utmanande. Det tredje bidraget innefattar implementationer av resultaten i ett agent-baserat coachingsystem för person-anpassat stöd till förändring i flera levnadsvaneområden. Arbetet tar också ett steg mot att utvidga detta till en svärm av samverkande agenter med olika roller. Det visades hur stöd och utmaning kan samordnas inom ett team av agenter, vilket bidrar till att digitala

coachningsagenter kan ta individens motiv, beredskap och skiftande preferenser i beaktande och anpassa dialogen till detta. Slutligen presenteras kvalitativa resultat kring hur deltagare upplever och hanterar stöd och utmaning i vardagen, vilket ger empiriskt grundade implikationer för design av onboarding och personanpassning av agentens roller. Tillsammans ger dessa bidrag en teoretisk grund, empiriska insikter och implementerade system av AI-agenter som anpassar stöd och utmaning till individen, och identifierar vidare samordning mellan flera agenter som en gångbar väg mot personanpassad coachning för förbättrad hälsa genom levnadsvaneförändring.

Sammanfattningsvis bidrar avhandlingen med en fördjupad förståelse för hur en AI-baserad agent eller ett team av agenter kan anpassas till människors individuella och skiftande behov.

Preface

Paper I **Kaan Kilic**, Saskia Weck, Timotheus Kampik and Helena Lindgren. Argument-based human-AI collaboration for supporting behavior change to improve health. *Frontiers in Artificial Intelligence 6:1069455. doi: 10.3389/frai.2023.1069455*, 2023.

Paper II Helena Lindgren and **Kaan Kilic**. Participatory design and evaluation of digital coaching for improving health—the star multicomponent lifestyle intervention. *Frontiers in Digital Health 7:1600535. doi: 10.3389/fdgth.2025.1600535*, 2025.

Paper III Victor Kaptelinin, **Kaan Kilic** and Helena Lindgren. Behavior Change Beyond Intervention: An Activity-Theoretical Perspective on Human-Centered Design of Personal Health Technology. *Frontiers in Digital Health 8:1754399. doi: 10.3389/fdgth.2026.1754399*, 2026.

Paper IV **Kaan Kilic** and Helena Lindgren. To Challenge or Support: Managing Behavior in Dialogues for Promoting Health Using Argument Schemes with a Digital Companion for Human-AI Teaming. Manuscript submitted, 2026.

Paper V **Kaan Kilic**, Vera Kaelin, Julian Alfredo Mendez, and Helena Lindgren. Human-Swarm Teaming for Personalized Behavior Change Support: An In-The-Wild Feasibility Study. Manuscript submitted, 2026.

Paper VI **Kaan Kilic**, Vera Kaelin, and Helena Lindgren. Exploring Recovery Activities in Collaboration with a Digital Companion: A Case Study of Aime. Manuscript submitted, 2026.

In addition to the papers included in this thesis, the following publications were published within the doctorate but not contained in this thesis.

Paper VII Baskar, J., Kaelin, V.C., **Kilic, K.,** Lindgren, H., Argumentation-Based Person Tailored Storytelling for Older Adults: A Reflective Hybrid AI Architecture. *ArXiV Preprint.* *doi: 10.48550/ARXIV.2605.10531*, 2026.

Paper VIII Kaelin, V.C., **Kilic, K.,** Neely, A.S., Gavelin, H.M., Wennberg, P., & Lindgren, H. Characteristics of digital twins to visualize, monitor and predict outcomes on participation in meaningful activities. *Frontiers in Psychiatry 16:1694966. doi: 10.3389/fpsyt.2025.1694966*, 2026.

Paper IX Baskar, J., **Kilic, K.,** Kaelin, V.C., Lindgren, H. Towards collaborative planning for health promotion through person-tailored storytelling and argumentation. *In 1st Workshop on Human-AI Collaborative Systems co-located with 28th European Conference on Artificial Intelligence (ECAI 2025), Vol. 4072. CEUR-WS, Bologna, Italy, 70-83*, 2025.

Paper X Persiani, M., Guerrero, E., Brännström, A., **Kilic, K.,** Kampik, T. Fantastic Argumentation Tools and Where to Find Them. *Fifth International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2024), Vol. 3757. CEUR-WS, Hagen, Germany, 56-68*, 2024.

Paper XI **Kilic, K.,** Lindgren, H. Constructing and Argumentative Agent in the Healthcare Domain: A Participatory Design Approach. *Inter.HAI Workshop at HAI*, 2023.

The PhD studies were funded by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS), and the research project "Digital Companions as Social Actors: Employing Socially Intelligent Systems for Managing Stress and Improving Emotional Wellbeing" was funded by the Marcus and Marianne Wallenberg Foundation. The work related to StarCoach was partly funded by Forte, the Swedish Council for Health, Working Life and Welfare (STAR-C, Dnr. 2018-01461).

Acknowledgements

Where to even begin? I'd first like to start by thanking my supervisor, **Helena**. There is no amount of gratitude I could show nor say that could ever repay the mentoring, but more importantly, the genuine care you have shown throughout my PhD. I am forever thankful to you for everything you have done for me. I'd also like to thank my co-supervisor **Victor** who has introduced me to different ways of thinking throughout my academic journey.

Vera, thank you for being an example of how an excellent academic should be like and a great support during my PhD. **Tim**, thank you for all the advice especially at the beginning of my PhD. **Julian**, thank you for being a "brother-in-arms" in hunting for bugs as well as being a place to turn to for advice. The bug hunting party wouldn't be complete without you, **Lennart**, who I'd like to also thank for the tents which have traveled many miles and housed many adventures for me and my family. A thanks to you, our unofficial "consultant" but even more a lifelong friend, **Eren**. Yakında seninde savunmanı görme dileğiyle. Finally, thanks to you, **Jaya**; it was great collaborating with you!

As for my reference person **Polina**, thank you for all your advice and support! Thank you, **Thomas**, for being my teaching mentor at the beginning of my journey, which has helped immensely in my teaching. Thank you, **Kai-Florian**, for your time and support as the director of studies and thank you **Frank** for your conversations and support, including outside of the PhD. Thank you, **Erik**, for your hospitality and for being the best grilled salmon maker in Northern Europe.

Special thanks to my legendary office mate, **Igor**. Here's to having many more fika together! Thank you to all the PhDs and postdocs I've met along the journey especially those in our department, past and present. Thank you to the NTK crew namely from our department **Arvid**, **Oliver**, **Jed**, **Tanaz**, and others, namely, **Erik** and **Moa**.

Thanks to the best IT team/car experts/lock pickers in the world **Tomas**, **Andreas**, **Mattias**, and **Bertil**. Thank you, **Jonny**, for having the whole department running smoothly. Thank you, HR and finance team, for all your support and understanding, especially **Ennie** and **Heini** for all your patience. And thank you **Ellinor** and **Helena** for helping with the courses.

Thank you to the WASP-HS PhD students for all the great times we had together! Special thanks to the broskis **Joakim** and **Felix**, see you on Dust

II, the CT side, of course. Special thanks to **Ericka** and **Eva**, who have been so understanding and helpful throughout my PhD journey. I truly appreciate you both so much. Lastly, a big thank you to WASP-HS and those who help it run; the work you do has helped hundreds already.

Thank you to **Hatice** and her lab who welcomed me during my visit to Cambridge. I learned a lot and will never forget the experience. A very special thanks to **Arzu** who made it all possible and was always there when I needed it. Thank you, **Pedro**, for all your help when we first arrived in Umeå. The kindness you showed still brings tears to my eyes.

Thank you to **Juan-Carlos**, **Lois**, and **Roy**, whose input have helped strengthen this thesis. And thank you to all the **participants** who took part in our studies.

Finally, a big heartfelt thank you to my **family** who have always believed in my abilities. Hope I make you proud. A special thanks to my hayat arkadaşım, **Aray**. You stuck with me through thick and thin and have been so patient. I owe all of this to you, your love, and your never-ending support. Seviyop. Haydi artık sonunda rahatça oturup kahve içelim, offff.

Abbreviations

AI	Artificial Intelligence
AIF	Argument Interchange Format
APCT	Argument from Position to Create Tension
APS	Argument from Position to Support
BCSS	Behavior Change Support Systems
BCT	Behavior Change Technique
CASA	Computers As(Are) Social Actors
CQ	Critical Question
DoLE	Division of Labor Efficiency
ER	Efficiency Ratio
GenAI	Generative AI
HAI	Human-AI Interaction
HAT	Human-AI Teaming
ICF	International Classification of Functioning, Disability, and Health
JITAI	Just-In-Time Adaptive Intervention
LLM	Large-Language Model
MHBC	Multiple Health Behavior Change
MI	Motivational Interviewing
MoA	Mechanisms of Action
OWL	Web Ontology Language
PD	Participatory Design
PSD	Persuasive System Design
RAG	Retrieval-Augmented Generation
RDF	Resource Description Framework
RQ	Research Question
SDT	Self-Determination Theory
TAS	Teamwork Assessment Scale
TTM	Transtheoretical Model
VIP	Västerbotten Intervention Program
ZPD	Zone of Proximal Development

List of Figures

2.1 Activity System Model 8
2.2 The hierarchies of activity 9
2.3 Toulmin argument model 11

3.1 Participatory design process of *StarCoach* 32

4.1 Argument construction pipeline of Generation I 37
4.2 *StarCoach* architecture 39
4.3 *StarCoach* Star Profile and Activity pages 40
4.4 *StarCoach* coach 41
4.5 Design process: from theory to action 43
4.6 The *Blumeai* swarm architecture 45
4.7 *Blumeai* app pages 46
4.8 Aime's pages. 49
4.9 The agent chat, the Activity suggester, and the customization of the agent's role and avatar in Aime. Adapted from [134]. . . 50

5.1 Behavior Change, network of three interacting activity systems 58
5.2 *StarCoach* as a hybrid behavior change system 59
5.3 Argument-based dialogue activity of Paper IV 62
5.4 Thesis Flow 66

6.1 Direction, Grounds, and Level 71
6.2 Basic Needs of SDT 76

List of Tables

2.1 Dialogue types of Walton and Krabbe 12
2.2 Argument from Expert Opinion Scheme 14

3.1 Methodologies, studies, and methods used to address each research question. 30

4.1 The four system generations. 35
4.2 Agent roles and their supportive/challenging weights 41
4.3 Overview of the embedded BCTs in *Blumeai* 47

5.1 Mapping of AT concepts and dialogue structures 60

Contents

List of Figures xiii
List of Tables xv

1 Introduction **1**
1.1 Research Questions 3
1.2 Contributions 4
1.3 Delimitations 5
1.4 Thesis Outline 5

2 Background and Theoretical Frameworks **7**
2.1 Modeling the situation with Activity Theory 7
2.2 Communication via Argumentation 10
2.2.1 A brief history of argumentation applied in AI 10
2.2.2 Argumentation Schemes 13
2.2.3 Argumentation in persuasion and health 14
2.3 Integration through dialogue 15
2.3.1 The Behavior - Activity Link 16
2.3.2 The Argumentation Connection 16
2.4 From Interaction to Teaming 18
2.4.1 Computers as Social Actors 18
2.4.2 From tool to teammate 19
2.5 Support vs. Challenge 20
2.6 Behavior Change to Improve Health 21
2.6.1 A brief history of behavior change 21
2.6.2 SDT, TTM, and Activity Theory 24
2.6.3 Behavior change support and techniques 25
2.6.4 Stress and recovery 26
2.7 Summary 26

3 Methodology **29**
3.1 A Participatory Design Approach 31
3.2 Study Phases 32
3.3 Data collection and analysis: Constants and Change 33

4 From Theory to System **35**
4.1 Generation I: Exploring Agent Roles, Micro-Dialogues with Arguments and Knowledge Base (StarCoach prototype) 36
4.2 Generation II: StarCoach 37
4.2.1 Four modules 38
4.2.2 The supportive and challenging coach 38
4.2.3 The transparency expectancy 44
4.3 Generation III: The Blumeai Swarm 44
4.4 Generation IV: Aime 48
4.5 Human-AI Team 51

5 Research Contributions **53**
5.1 Summary of the contributions 53
5.2 Paper I 55
5.3 Paper II 56
5.4 Paper III 57
5.5 Paper IV 59
5.6 Paper V 63
5.7 Paper VI 64

6 Discussion **67**
6.1 Aligning with a dynamic target 67
6.2 Beyond a single scale of support and challenge 70
6.3 What changed when moving to genAI? 73
6.4 Team vs. Leader 74
6.5 Habits and relations 76
6.6 Reflections on Methodology 78
6.7 Addressing the RQs 80

7 Limitations and Ethical Concerns **83**
7.1 Limitations 83
7.2 Ethical Considerations 84

Paper I **113**

Paper II **136**

Paper III **159**

Paper IV **178**

Paper V **248**

Paper VI **274**

Chapter 1

Introduction

If you could create a digital companion to support you, what would it be like? How would it act? How would it behave? What role would it have? How would it speak? When considering a human companion, we tend to have clearer expectations and hunches about the answers to these questions. When considering Artificial Intelligence (AI) companions, however, the answers become less obvious. What *should* an AI companion be like, how should it behave, what role should it play? These seemingly simple questions with, perhaps, simple answers become more difficult when the companion is artificial rather than human. Coupling that with the fact that everyone may have different expectations of an AI companion [45, 64, 110, 257], makes defining the "perfect companion" a difficult problem to address.

This problem is known as AI alignment. Though there is not a universally accepted definition, for our human-centered purposes, AI alignment can be defined as an iterative process of designing, developing, deploying, and governing AI systems so that their behavior, objectives, and effects are responsive to human intentions, values, interests, autonomy, and agency, while accounting for diversity, uncertainty, limitations, and potential disagreements or harm [65, 87, 98, 118, 174, 244].

Alignment is crucial as we encounter digital assistants, conversational chatbots, and the increasingly popular large-language models (LLMs) more and more in services we use daily [155, 156, 270]. Despite their intelligent facade, these systems are very much limited and prone to mistakes [6, 132, 138]. They miss the very things that we, as humans, use to operate. They are currently ungrounded in the real-world, lack lived experience, and rely on probabilistic guessing which leads to hallucinations and mere mimicry [18, 208]. Regardless, people view these systems as social actors [30] and may form bonds with them [111]. It is also apparent that AI is increasingly developing at an exceptionally swift rate, improving on the limitations listed above.

These facts uncover possibilities to utilize these AI systems for beneficial purposes [30]. Strong social support and bonds are major elements in helping

people achieve their goals and achieve healthier lifestyles [93, 175]. As AI technology develops, digital assistants can transform into digital companions that can enact roles with greater positive influence in daily life [168, 273]. However, it is important to align the companion to the individual's needs and motives and build it together with the people the companion aims to affect [23, 68] due to the sensitivity of such influential systems where transparency, control, and mutual influence are all things that need to be taken into consideration [159].

To frame this thesis and give context, this thesis situates digital companions as social actors designed to support behavior change for better health and lifestyle. Though many desire to have a healthier lifestyle [97], the hindrances to changing for most are habit systems, conflicting motivations, and contextual factors [37, 73]. To address this, the usual approach of digital interventions taken by researchers is to have a digital intermediary to bridge the domain expert's advice to the changee. This extrinsically regulated motive is not sustainable to the behavior change that one often strives to achieve as opposed to intrinsic [218], and to achieve this, the three basic needs of autonomy, competence, and relatedness must be met [218].

People are dynamic [22, 62, 216]. They do not stay the same, even from hour to hour [217]. Then, how can an AI companion stay aligned to and support a person whose motives, needs, preferences, and mood changes constantly, also in situations where disagreeing or challenging the person might be the best way to support them? Research has shown that a person needs both support and challenge, delivered at different moments depending on the person's state [36, 218, 223].

To do this calls for a newer generation of non-static digital assistants: the dynamic socially intelligent digital companion. These are a kind of agent handling complex human activity in consideration with not only the "doing" of the activity but also the relational, cultural, and collaborative aspects. They are adaptive to the individual and collaborate with the person in their path to their goal.

However, for the digital companion to be meaningfully able to do this, the challenge of designing the companion as well as the interaction between the two must be tackled. The agency and autonomy of both parties must be clearly defined but also be able to cooperate with one another as team players [120, 196]. This teaming between the human and AI is known as human-AI teaming (HAT) which O'Neill et al. [196] define as the *"interdependence in activity and outcomes involving one or more humans and one or more autonomous agents, wherein each human and autonomous agent is recognized as a unique team member occupying a distinct role on the team, and in which the members strive to achieve a common goal as a collective."*

In this thesis, we start with a rule-based, pre-LLM companion that eventually evolves into a small team of specialized companions teaming up with the person, situating the work within the fields of HAI and, as the agents take on more autonomous and coordinated roles, HAT. The final generation returns to a single generative companion whose role the person can choose to create

themselves, and follows it into a new domain, recovery from everyday stress, for up to 12 weeks.

1.1 Research Questions

This thesis is guided by three overarching research questions, each addressed by the papers included in this dissertation:

RQ1 How can theory and empirical findings inform an agent's strategies for improving AI alignment in lifestyle change; supporting and challenging the person through roles, behaviors, and argument-based dialogue, and how can these strategies, together with theory on human activity and argumentation, inform the agent's knowledge and reasoning models?

RQ1 identifies the origin of the content of alignment. Some of the sub-questions this RQ tries to address are: what does an agent need to know to be able to reason about to position itself in relation to the person?; what roles can it employ?; what grounds can it use?; how does the person's motives, long-term goals, short-term goals relate to the actions being taken?; how can theories on human activity and argumentation supply the structure? RQ1, in a nutshell, is the question that bridges theory to the agent's internal model.

RQ2 How can supporting and challenging behavior be formalized and implemented in a digital companion so that its dialogue and behavior can be aligned with an individual's motives, readiness, and shifting preferences?

RQ2 addresses how the agent's internal model established in RQ1 can become behavior that the companion conducts. It asks what is needed to make the notions of support and challenge into machine-useable constructs: explicit schemes, questions that can be raised against the schemes, protocol to establish the admissable moves and when they can be used, and an implementation in which the person's motives, values, preferences, and readiness mold what the agent does and says.

RQ3 How do people experience collaborating with such a companion (and a team consisting of a team of agents with different roles), in daily life, in terms of personal relevance, autonomy, and the negotiation of conflicting goals, and what design implications follow?

RQ3 is about the implementation being used real life. It addresses whether the position and behavior of the agent is experienced as relevant personally, how situations where the agent and the person's positions mismatch, what the presence of the agent changes or does to the regard of the implementation, and what should be updated in the design as a result.

1.2 Contributions

The following is a summary of key contributions of this thesis:

1. **Reframing of personal health technology from "delivering an intervention" to sustaining own evolving activity.** With Activity Theory, this thesis posits that behavior change is something a person does over time. It is an endogenous and developing process, not something a system or healthcare administers, or should administer to them (Paper III). This conceptual foundation necessitates everything else; if the individual is changing, an agent cannot be aligned to "once upon a time" profile of them. It reframes the design target from complying with a plan to keeping pace with the evolving person while showing that it is possible to embed both motives in a design, both the exogenous and endogenous motives, which is in itself a novel, theory-based contribution.

2. **Identifying support versus challenge as a core axis of a helpful and fruitful human-AI companionship.** The findings across the thesis is that to genuinely help someone sometimes means pushing back but there is no single setting that always works. Some people want a forceful, "pushy" companion, while others prefer affirmation and empathy. The same individual's preference can shift over time and systems must be able to accommodate such a change (Papers I, II, VI).

3. **Integrated argumentation theory with activity theory into a formal, inspectable dialogue framework.** This thesis maps activity theory constructs (motives, goals, contradictions) onto dialogue constructs and pairs them with argument schemes and a dialogue protocol (Paper IV). This makes the system accountable because the agent's reasoning is based on schemes and is traceable. Things like a person's disengagement can be localized to specific mechanisms rather than speculation.

4. **Design, development and evaluation of a multi-domain personal health system through participatory design.** Beyond frameworks, the works in this thesis led to the development of a system co-designed with a twelve person multidisciplinary expert team, embedding a structure of Behavior Change Techniques (BCTs) across multiple lifestyle domains. Onboarding and feasibility studies were conducted to evaluate various aspects of the developed system (Paper II).

5. **Extended single-agent approaches to a multi-agent teaming architecture and evaluated human-AI teaming.** Since single-agent architectures must unify the responses despite the answer being a multi-faceted one, this work explored recasting BCTs as differentiated roles within a multi-agent swarm. The work also proposed metrics to assess teamwork through a feasibility study to contribute towards a validation of human-AI teaming (Paper V).

1.3 Delimitations

This thesis is written at a time where AI capabilities are changing at a rapid pace and alternative definitions of central concepts are emerging. Therefore, some delimitations are needed and listed in this section.

- The use of the term "AI alignment" evokes machine learning methods in many scholars' minds. We opt for the human-centered or human-AI interaction approach and definition that focuses more on the AI system's goals and behaviors being responsive to a person's motives, goals, and preferences as they evolve over time.
- It is important to highlight that better health and lifestyle, including recovering from everyday stress, is the use case, the underlying problem of how AI systems represent and respond to a dynamically changing person is not specific only to this case, and this thesis treats it as an application domain for a more general claim about human-AI interaction (HAI) and alignment.
- Research presented in this thesis is interdisciplinary, utilizing theories from computer science, psychology, social sciences, healthcare, and philosophy.
- The two argumentation schemes Argument from Position to Support (APS) and Argument from Position to Create Tension (APCT) introduced in this work should be viewed as meta-level argumentation schemes. In other words, arguing about how to argue.
- The aim of this thesis is not to ultimately show evidence that a digital companion changes health behavior. It contributes to the field of computer science through novel theories, formal models, methods, design implications, and techniques that can eventually become building blocks of systems supporting lifestyle change. The user studies examine how people experience and use these building blocks rather than their effect on behavior or health.

1.4 Thesis Outline

This thesis is organized into 7 chapters, finishing off with 6 research articles. Chapter 2 delves more into the theoretical framework and the background which build the spine of this research. Chapter 3 continues on to reflect on the methodology across the studies as whole. Chapter 4 introduces the apps that were built to house the digital companion for the participants. Chapter 5 gives an overview of each paper and the research contributions and Chapter 6 discusses what the thesis does and does not. We conclude with Chapter 7, reflecting on the limitations, and ethical concerns.

Chapter 2

Background and Theoretical Frameworks

The definition of AI alignment that is given in Chapter 1 implies that it represents a multidisciplinary challenge. This thesis builds on and draws from theories in several research domains.

To design the human-AI interaction and human-AI collaboration, Activity Theory is used [129, 150, 259]. For communication purposes and structuring arguments and dialogues, Argumentation Theory is employed [16, 261, 262]. Theories and methods that have been utilized for behavior change have also contributed to the research of this thesis are, namely Self-Determination Theory [218], Behavior Change Support and Techniques (with Mechanisms of Action) [40, 199, 212], and Models of Behavior Change [80, 210].

In this chapter, the background and theories that have built the spine of this thesis are outlined. The combination of the theories for their use in this thesis is elaborated on in Section 2.3.

2.1 Modeling the situation with Activity Theory

Aligning person-tailored support to improve health to a person's motive, goals, situation requires a model of what the person is trying to achieve. Activity Theory (AT) is a theory capturing human activity and various aspects relating to this. Established by Leontiev and Vygotsky [150–152, 258, 259] and further developed by Engeström [70–72], AT takes the motive-driven purposeful activity as the unit of analysis, rather than only isolated, goal-directed individual actions carried out by a person. These purposeful activities are directed at an object, driven by an underlying motive, and mediated by tools, social rules, division of labor, and community of which the person is a part of (Figure 2.1).

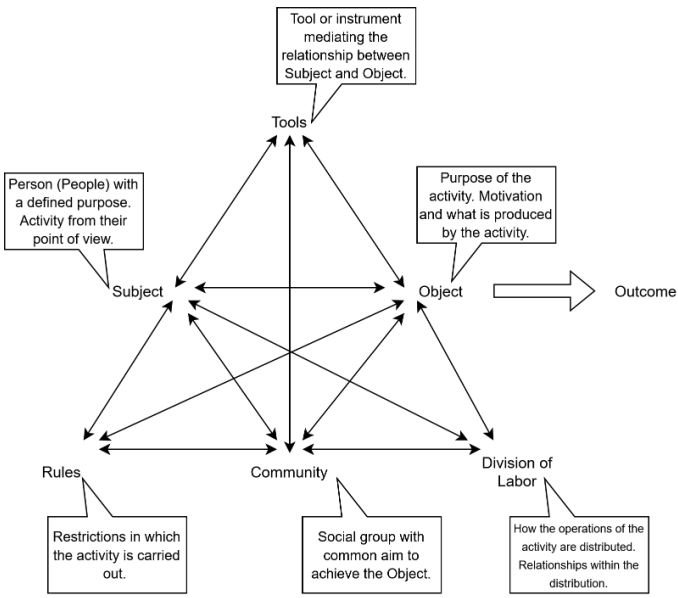


Figure 2.1: Activity System Model [72] with six nodes with differing functions all contributing to an activity and activity systems.

AT allows us to map the person to society at large and takes into consideration the *network* of activity systems or multiple activities with overlapping objects. This is important as behavior cannot be understood in isolation from the cultural and social environment, according to Leontiev and Vygotsky [150, 259]. AT is used as a means of connecting the individual with society and argues that both shape each other through shared objectives. It furthers its argument by stating that the subject interacts with the world through mediating tools, which should not require conscious thought, be it physical like a computer or symbolic like language. AT also surfaces mismatches or contradictions within the system, for instance, when a tool does not work as intended it need to be focused and altered, such as language, thus also altering the ongoing activity. A contemporary example can be an office worker (Subject) who has been to a health checkup and found out they are at a raised risk of cardiovascular disease, so wants to be more physically active (Object). They use a conversational agent on their mobile phone to plan walks and get reminded for the walks (Tools). But long work days and an uneven amount of household duties (Division of Labor), long mandatory workplace meetings (Rules), and friends whose time together mostly involves sedentary activities (Community) all prevent the Subject from carrying out their objective. An agent which sees only the Subject could misinterpret the missed walks as low motivation and send more reminders. Mapping the whole system can emerge the problem as a contradiction in the system rather than an individual issue, for example be-

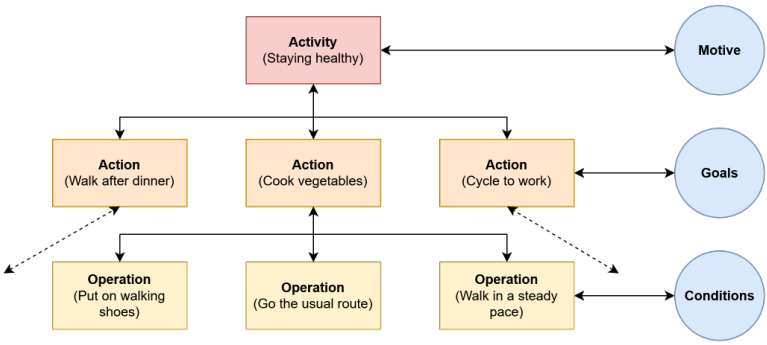


Figure 2.2: The hierarchies of activity split into Activity, Action, and Operation following the example of Staying healthy as the activity/motive. An activity is made up of actions which are made up of operations [129].

tween the time in which the agent suggests a walk and the person's working time. Once the person is mapped to the larger social system, we can pinpoint where the system breaks down. This point of breakdown is also called "contradictions" in AT terms.

An activity consists of three dynamic levels of action: *activity* oriented towards an object, goal-oriented *actions*, which can be more than one level in turn distinguished from the *operational* level, which is sub-conscious, tasks which have been internalized and become automated such as the use of language or tools (Figure 2.2) [150]. Leontiev argues that one concrete way to distinguish activities from one another is their object, or their true motive [150]. There is always an individual's need behind an activity, such as those elicited by Self-Determination Theory (autonomy, relatedness, competence). This is the first level in the hierarchy that is overarching the rest. An example of an activity at this level can be *staying healthy* where the individual's motive is to have a long and active life free of cardiovascular disease. The second level is *Action* which are the conscious steps individual(s) take to achieve the broader activity's motive. These actions are clear in their beginning, end, and goals. Each action requires focus and attention to carry out. An example of an action can be taking a 20 minute walk after dinner. The final level is *Operations* which are driven by Conditions. Operations are routine actions that are carried out without conscious thought, due to their familiarity and are not thought about unless something breaks. Operations are identified by the available tools or setup, in other words, conditions. An example of an operation can be can be putting on one's shoes and going for a walk in a familiar route without consciously thinking about the route, with the conditions of, for instance, the route starting at the door of the workplace.

The hierarchy is fluid in the sense that it shows behavior as something that constantly moves up and down the hierarchy between actions and operations

based on developments or disruptions. An activity moving down the hierarchy is known as automatization [128]. Automatization is the process of when an action becomes an operation, such as a person new to weight lifting having to think about each part of an exercise until the movement becomes automatic. The inverse is de-automatization where an operation returns to becoming an action. An example can be the person joining a new gym with new, unknown equipment, or recovering from an injury, and having to think about the movement anew.

2.2 Communication via Argumentation

Modeling the person and the situation is only half the job. The agent must also be able to act upon the model created. It must be able to decide *what* to communicate and also know *why*, in a way that is not hidden and ambiguous [237]. Argumentation theory offers an avenue to transparency in that regard. It empowers the agent to posit its stance as an argument with claims that are grounded, and open to be questioned or challenged as opposed to only being a binary supportive or challenging entity.

2.2.1 A brief history of argumentation applied in AI

Argumentation theory, as we know it today, can be traced back a millenia [55] but to Toulmin, it was not enough. Formal logic was too abstract and formal and it is difficult to carry these formalisms or syllogisms to real-life arguments. Therefore, in his book *The Uses of Argument*, Toulmin proposed a model for analyzing and building logical arguments using real-world data (grounds) [251]. Through the grounds (evidence), a claim could be presented that is backed by warrants (logical bridge between grounds and claim) and backing (support for the warrant). Additionally, qualifiers (limits to the claim) and rebuttals (exceptions) can also be used to limit or strengthen the argument. In Figure 2.3, an example of AI being used as a tool for supporting consistent study habits is given through Toulmin's argument model with the atypical addition of "Response".

Perelman and Olbrechts-Tyteca [204] continued the line of re-scoping formal logic with practical real-world approaches. In their book, Perelman and Olbrechts-Tyteca [204] argue that human beings communicate differently than systems with strict rules, however, the argument and how it is received is dependent on the one receiving it. The premise of the argument, or the starting point, should be started with what the receiver is already accepting. The "real" are hard facts and truths that is accepted without the need for further proof whereas the "preferable" represents values and preferences guiding choices but are not hard truths. The ways to deliver these arguments are through association (bring together concepts to expand beliefs and values) or dissociation (break down concepts to resolve contradictions).

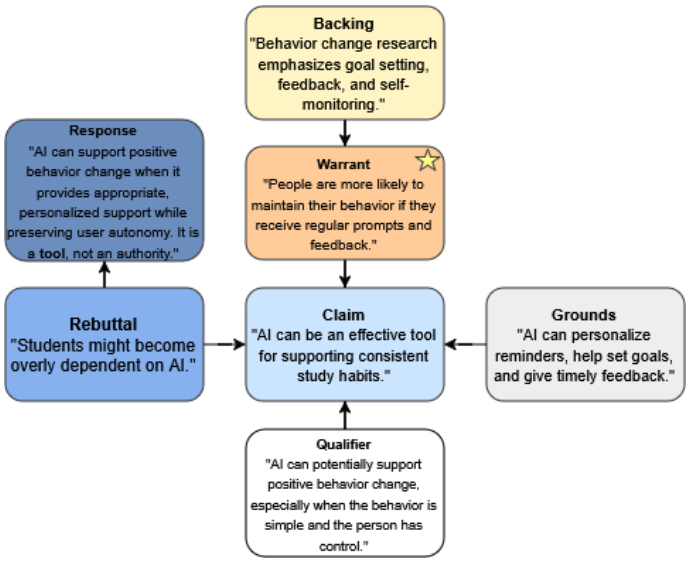


Figure 2.3: Toulmin argument model [251]. The example given to explain the model follow the claim "AI can be an effective tool for supporting consistent habits". The Warrant has a star next to it to emphasize its role as the logical bridge between the grounds and the claim.

In 1970, Hamblin [101]'s constructive work *Fallacies* was published. This work also criticized the deficiency of approaches till then stating that pure formal approaches cannot truly explain why real-world arguments fail. These failures were attributed to context and intent which formal mathematical structures could not capture. Though *Fallacies* as the name suggests, mostly focused on fallacies, a key idea from Hamblin [101]'s work was how real-world dialogue contexts may alter how legitimate an argument is even though the truth values might say otherwise.

Then, after about two and a half decades, Walton and Krabbe [262] mapped out their dialogue typology to shift arguments from isolated structures to conversational context. They outlined six dialogue types: persuasion, inquiry, negotiation, information-seeking, deliberation, and eristic. An explanation of each type is taken from [135] and given in Table 2.1. Distinguishing the dialogue types allowed for looking into the dynamics of dialogue shifts and shifting the "burden of proof" which is the idea determining *which* participant in the conversation is required to back up their claims at a given moment, and the level of evidence they need to provide to go forward in the conversation.

During the same time, argumentation entered the field of AI from another direction. Defeasible reasoning, where a conclusion is drawn from incomplete information and withdrawn with the arrival of new information [207], had been a struggle for AI researchers. Dung [66]'s groundbreaking work showed that

Type	(Dialogue) Purpose	Example
Information Seeking	Find new information	Asking questions
Inquiry	Collaboratively create new knowledge	The two parties cannot alone create new knowledge, collaboration is needed
Deliberation	Decide about what to do	Form a plan and decide on the order of execution
Persuasion	Convincing of other party(ies) to adopt a belief one party(ies) has	One side holds a belief and wants to convince the other side(s)
Negotiation	Eliciting gains and compromises during decision making	Calculate gains and losses
Supportive	Provide the other agent support/ Reinforce other agent	Expressing emphatic behavior, affirmation, creating rapport
Challenging	Create tension (cognitive dissonance) in other agent	Pose statement that is in conflict with previously posed statement for the purpose to create tension/cognitive dissonance

Table 2.1: Dialogue types defined by Walton and Krabbe [262] with the addition of Supportive dialogues introduced by Baskar et al. [14] and Challenging introduced in [136], highlighted in italic. Table taken from [135].

defeasible reasoning can be represented in a graph format of arguments with attack relations where arguments that "survive" can be calculated from the graph itself. This abstract method was then built upon with frameworks such as ASPIC+ [183], which described the arguments that were built and how they attacked one another, and also by value-based frameworks [16], where one argument defeats another argument based on the values held by the audience. Through these lines, argumentation became a way to compute conclusions and also a way to structure how parties exchanged reasons [17, 234].

Another use of argumentation was multi-agent systems. When agents with varying information and varying goals need to cooperate and collaborate, they need to have a way to exchange not only proposals but also the reasons behind the proposal. Agents with negotiation capabilities can do so through argumentation to explain why a proposal is unacceptable and alter the other party's position [203]. Through building on formal dialogue games of Hamblin [101] and Mackenzie [167], researchers in argumentation turned Walton and Krabbe [262]'s dialogue types into protocols, which are rules about, for example, which dialogue moves are allowed and when the dialogue ends [173, 209]. The Argument Interchange Format (AIF) provided arguments with a common representation, allowing them to be stored, connected, and exchanged between systems [48]. During this period of argumentation, the parties were mainly machines, and the parties in the dialogue normally used argumentation to win

an argument, persuade, or maximize gain.

With time, argumentation transformed from dialogue between machines into dialogue between machines and humans. A reason for this is that the argument structures that were used were close to how humans reasoned and provided reasons in daily life [63, 184], which, in turn, allows the human to follow the reasoning and question it. Medicine was a domain that this was tested on early. Argumentation-based decision support showed clinicians arguments for and against the available options which gave the clinician the power to choose between the options [84]. This decision support was then extended to managing chronic conditions at home [143]. As explainable AI started to gain traction, argumentation has been used as a way to explain the outputs of AI-based systems. For instance, argumentation was used to organize explanations as a set of arguments which the person could dispute and inspect [56]. With further advances in systems, eventually systems were able to reach a level where they could imitate natural language, even debating with humans on topics [238], to argument structured chatbots [43], to LLMs whose responses are organized as human contestable arguments [86].

The next step of taking explainable system to collaboration is where this thesis places itself. In the case of an AI systems and human collaborating to achieve a common goal, argumentation plays a larger role than explainability and transparency. It enables both parties to access the same argumentative moves such as putting forward a claim, question each other's reasons, supply information that alters outcomes, and explain their reasons to the other party. This "hybrid intelligence" way of approaching the issue [2], combines the strengths of humans and AI systems. In the health domain, the ACKTUS platform was built specifically with this idea in mind, using argument structures based on the AIF to person-tailor dialogues between a person and digital agent [161]. ACKTUS also housed a *supportive* dialogue type in addition to the types of Walton and Krabbe [262]'s to encapsulate the agent's encouragement and affirmation towards a person [14]. Paper I [136], further contributed with a *challenging* type (Table 2.1). In early work of argumentation, arguing was a sort of competition between the parties involved to achieve their own goals. In human-AI collaboration, the agent supports and argues on behalf of the person's goal, with the aim of each dialogue to inch towards that goal, irregardless of winning or losing the argument.

2.2.2 Argumentation Schemes

Walton [261] and Walton, Reed, and Macagno [263]'s work set the foundation for *Argumentation Schemes*. Argumentation schemes are templates representing structures of daily human reasoning. They are comprised of premises, conclusions, and critical questions to find flaws and gaps in the reasoning. An example of a widely known argumentation scheme, *Argument from Expert Opinion* is given in Table 2.2. Which scheme to use and when also shapes how the message is received by the person as one scheme might be more persuasive

in certain cases over other schemes [249]. Using argumentation schemes allows AI systems to understand human reasoning, view the disputes, and see the conflicts (or supports) of delivered arguments. They also enable presumptive reasoning which facilitates decision making under uncertainty and new incoming information.

Major Premise:	Source E is an expert in subject domain S containing proposition A.
Minor Premise:	E asserts that proposition A is true (false).
Conclusion:	A is true (false).
CQ1: Expertise Question:	How credible is E as an expert source?
CQ2: Field Question:	Is E an expert in the field that A is in?
CQ3: Opinion Question:	What did E assert that implies A?
CQ4: Trustworthiness Question:	Is E personally reliable as a source?
CQ5: Consistency Question:	Is A consistent with what other experts assert?
CQ6: Backup Evidence Question:	Is E's assertion based on evidence?

Table 2.2: Argument from Expert Opinion Scheme from [263] paired with its set of critical questions CQ. Table taken from [135].

The critical questions are the vital part of the schemes because they are the specifiers of a scheme's defeat conditions and also the shifters of the burden of proof used for evaluating the argument. Therefore, schemes are templates for *how* to argue but also how to *challenge*. An example of usage of argumentation schemes in coaching dialogues can be *practical reasoning* [8]: an agent has goal G, action A is a means to goal G, therefore A ought to be performed. The critical questions of this scheme map onto coaching dialogues well, i.e., is A feasible, are there any side effects, is G really the person's goal, is G in conflict with other existing goals? Atkinson and Bench-Capon [8]'s value-based approach have been applied in personalized systems, as defeats in the argument graph are relative to the value ordering, meaning that parties can accept the same premises and still disagree, without either party being wrong, due to ranking of values differently.

In this thesis, schemes are used as executable templates, making the agents' utterances traceable to a scheme and its premises, and the set of possible agent moves is foreseeable in advance. Schemes are also used as "meta-level schemes", or arguing about how to argue, with a person.

2.2.3 Argumentation in persuasion and health

Argumentation has been used to support healthier behavior for over 20 years. Grasso, Cawsey, and Jones [92] introduced Daphne, a system for advising for healthy nutrition, borrowed from Perelman and Olbrechts-Tyteca [204]'s New

Rhetoric to start from what the person already accepted and to resolve conflicts between the system's advice and person's beliefs through dialectical argumentation. Portia was used to adapt persuasive arguments about healthy eating to the person's values and emotional state [172]. In more recent years, *computational persuasion* through automated argument-based dialogue with behavior change as its application domain was researched by [112, 113]. In the framework of [113], the system contains the person's beliefs and concerns, selects arguments from an argument graph, and as the dialogue goes on, updates the model. Hunter's later work proposed that addressing the person's own concerns were better at persuading than systems that did not [99], and built chatbots that chose their replies from crowd-sourced argument graphs [47].

The difference between these works and the work presented in the thesis, albeit close, is the question of who holds the goal. In Hunter's computational persuasion, it is the system who holds the goal of persuasion set by the designer. In such systems, the outcome is supposed to be a person who accepts a belief or does an action, and the model of the person serves to achieve that goal. However, in this thesis, the object's owner is the person themselves (Section 2.1). And so, the agent's question is not only about which argument to present but whether it should argue at all and in which direction (e.g., supportive vs. challenging). Meta-level argumentation schemes serve this purpose, and this is the reason why they are "meta-level" as they reason about which position the agent takes before any object-level argument is made. Computational persuasion approaches update the person's belief model, and though the approach in this thesis does the same, it also updates the agent's model of how the person wants to be treated.

2.3 Integration through dialogue

How is human and agent activity framed? How is authority allocated, and by who? How are the reasons between the parties exchanged? These questions start to weigh more when an AI agent supports a person in personal and value-based goals over time. A setting where these questions are imperative is health behavior change. An agent's influence can undermine a person's autonomy, the very thing it was meant to support. The lines between persuasion and manipulation can become distorted and ethically questionable [125], and over-control and over-sycophancy can become modes of failure. Such concerns make behavior change a real test for the questions above. Behavior change theory alone cannot address them as it focuses on what an intervention does *to* a person but not what a person and an agent do *together*. The following sections aim to address this gap through Activity Theory and Argumentation Theory.

2.3.1 The Behavior - Activity Link

The frameworks reviewed later in Section 2.6.1 all share units of analysis, with having an individual with internal states and constraints with their behavior to be changed. The frameworks integrating various frameworks (Section 2.6.1) further include social factors as determinants acting on an individual. However, Activity Theory offers a different unit of analysis, the activity system consisting of subject, object, rules, community, division of labor, and mediating tools [70, 150, 259].

Integrating behavior change approaches with activity theory offers three avenues which would not be possible with one or the other exclusively. Firstly, the behavior and activity link. A BCT such as self-monitoring is, in the activity theory sense, an action or, possibly, even an operation, which is a subsidiary of the object-oriented activity giving it sense and meaning. Two subjects logging the same behavior might be conducting a completely different activity. BCTs alone cannot classify this but activity theory structures and connects the distinctions to the sociality of the activity as well as to the internal state of the person in the transitions between the hierarchy of activity (Figure 2.2).

Next avenue is the distinction of how an agent is viewed in the process. Traditional persuasive technologies viewed the system as a tool acting on a person or a social entity copying a human persuader [78]. Activity theory enables another lens by treating the agent as a mediating artifact reorganizing the activity *and* as a subject in its division of labor.

Finally, activity theory reframes barriers from the individual into the systemic by letting some contradictions be surfaced and worked on rather than simply removing them. This "contradictions, not deficits" is a distinction between how behavior change approaches and activity theory differs. Behavior change models try to map the source of the change in deficits such as capability, opportunity, or motivation whereas activity theory maps the source of the change in contradictions within and between activity systems as opportunities to change and develop. Simply put, activity theory augments behavior change with the unit of analysis and is a framework where the "dynamic duo", human and agent, can be described as jointly conducting an activity and not an intervention acting on a person.

2.3.2 The Argumentation Connection

Human reasoning defeasible [245]. A recommendation such as reducing screen time, however, is based on presumptive inference from the general to an individual circumstance, and it holds until something in those circumstances defeat it. Essentially, this is a reasoning structure which argumentation theory can formalize. The practical reasoning scheme [8] mentioned in Section 2.2.2 is a form that most behavior change systems utilize, even if the system does not necessarily present it in that way. Once we become aware of this, we also realize the ways these schemes can fail. In the instance of the practical reasoning

scheme, the critical questions provide this mode of failure through answering whether the goal is the person's own. That is, if it is the person's own goal, whether other actions are better alternatives, whether the recommendation is feasible, and whether the outcome benefits outweigh the consequences. A system that cannot answer these questions are not transparent and has no means to represent of it being wrong hence, no way of recovering from it being wrong. This issue is a crossroads for argumentation theory and activity theory.

The critical questions are at the middle of those crossroads because this is where activity theoretic structures become actionable through dialogue as each of the questions refer to a relation that the activity model is describing, e.g., fitting between goal and motive, and alternative mediating means. Contradictions in the activity system, which are visible analytically, gain a foothold into conversation through "challenges" the agent can raise. Dialogue protocols give the interaction a commitment structure, so a person who is opting for a goal one day but changing it in another is understood as updating their commitments within an ongoing dialogue. Therefore, the agent can track and query the update and adapt resulting in an agent whose behavior is linked to the dialogue itself and can reason and evaluate its own arguments through the usage of argumentation schemes' critical questions.

Furthermore, argumentation schemes are semi-formal and are better at capturing real-life as opposed to formalisms due to the real-world being incomplete and constantly changing. Semi-formal approaches' conclusions are "presumptively" true based on what we know *now*, but can be rejected with new evidence. Semi-formal approaches also rely on natural language, which allow the context, tone, and intent of the parties to be variables in the equation. Moreover, they mirror heuristics and patterns that humans naturally use in the real-world such as arguing from analogy.

A theory discussed later in Section 2.6.2, SDT, recommends influence via meaningful rationales, genuine choice, non-controlling language, etc. These are, at the very core, specifics of dialogue where reasons are put forth, challenged, defended. Essentially, they are strategies for argumentative dialogue á la Walton and Krabbe [262], Modgil and Prakken [183], and Prakken [209]. Argumentation theory is the other side of the coin to SDT's interpersonal style or Persuasive System Design (PSD)'s dialogue support. We use argumentation theory to take the autonomy support posited by SDT and other behavior change theories mentioned in this work, and formalize it with dialogue moves, burdens of proof, and critical questions. Activity theory is used to situate the dialogue within an activity system and accounts for the mediation. Allocation of autonomy is used to determine who is the decision making authority over what aspects of the activity, and when it shifts. In this research, BCTs and their MoAs dictate what is being delivered through dialogue and why it (should) work.

2.4 From Interaction to Teaming

2.4.1 Computers as Social Actors

Nass, Steuer, and Tauber [189], in their work, used results from social psychology and replaced one of the persons with a computer. They found that the findings still held, with people continuing to be polite to computers and assigned gender stereotypes to computer voices. The participants also were judgmental of a computer while trusting another computer more. This phenomenon lead to the idea of Computers Are Social Actors (CASA), also widely known as the "media equation", where people treat media and computers as people and places [213]. Even people who believe that computers are not social beings have the same response. Nass and Moon [188] calls this mindlessness, where some social cues like interactivity, language, or some role, can trigger sociality which people then continue sub-consciously.

People prefer computers whose "personality" is similar to their own [190]. This finding has bearing on how digital companions should be constructed. It indicates that a companion's role and tone are a part of how well a companion could fit a person, for example. Further, computers that were flattering were liked more and were rated as better, even if the flattery was randomly said [79], meaning that sycophancy that we see in LLMs today is not something new and is built into how people respond to computers in the first place. Moreover, people who were told they were on the same team as a computer, and whose rewards depended on it, were more likely to follow the computers proposals [187]. Therefore, a sense of being in a team can be recreated through the setup, prior to teamwork taking place.

Work following CASA focused on building agents in place of computers. Relational agents using social dialogue and empathy to build a relationship to the person were developed to keep engagement higher in health interventions [21]. With the added capabilities to conversational agents, CASA was due for an update as people now also carried more communication with machines and how they related to machines needed to be examined [88, 164]. So, social responses to companions of today which have longer conversations and a memory of the person, are likely to be stronger than the computers of the original CASA studies [236].

The title of this thesis is, in a way, a nod to this history. Fogg [78] advanced CASA with his description of "computer as a social actor". In Fogg's work, the social actor is a persuader. This thesis goes beyond and positions the technology from a persuader to a teammate sharing the person's object. As [189] showed, people are willing to accept computers as social actors, and [187] showed that a feeling of teamwork can come about with technology.

2.4.2 From tool to teammate

Moving from interaction to teaming within a human-agent relationship is not a question of sophistication [228]. Human-AI teaming is already equipped with a sizable empirical backing [196], however, it is widely focused on tasks where the object of the joint work is given in advance and success of a task is decided within a single session. A coaching relationship is different in which the object is owned by the person and is dynamic throughout. Teaming, therefore, is a question of where the agent is positioned within an activity [129, 147]. The tool plays a mediating role between the subject and the object by extending the capabilities of a person without itself being visible or noticeable, or having any stake in the object [26, 259]. The agent could be designed to act in this way as a tool. If the agent is designed to act as a teammate, it becomes embedded in the team/community node in the activity system model. Its tasks dwell in the division of labor node instead of the tool node, and shares the work directed towards the joint object [72]. By doing so, it becomes a stakeholder in the pursuance of that object [32, 51, 95]. No matter how fluent, eloquent, "intelligent" it may seem [188, 213], an agent that only reacts to requests is still just a tool, since everything that it does is dictated by the person in advance [50]. An agent that commits to the object over the immediate request is something a tool cannot do [44, 117], and this commitment of the agent is where teamwork begins [119, 140].

Roles that agent can play are not necessarily a gimmick for the agent simply switching roles to appease the person and can be viewed as a means to an end; the end being the shared object of activity. The addition of the argumentation layer makes them commensurable since no matter the role being played, the critical questions and the dialogical states still are the same. An agent challenging the person can be a sign that the agent is working together with the person instead of trying to appease them [205, 230]. Moreover, it shows that the agent is a participant as a stakeholder in the outcome and is ready to disagree with the person for them to achieve the person's goal [116, 221].

So, why is this relation considered a team and not a sycophantic assistant with a built in override? Well, this is because with this shift of the agent from a tool to a stakeholder, the object is shared and authority over the object is not so symmetric [77]. Where that authority sits is not fixed for the duration of the relationship and it moves as the person's ability, willingness, and preferences shift [31, 109, 224]. The person is the motive holder while the agent holds a derivative of the motive as a commitment and is now accountable to the person for the given commitment. In activity theory mapping, the agent itself has become a subject, in a sense, with its own activity system derived from the person's [71]. Here, again, argumentation supplies the necessary tools through accountability. Every challenge raised by the agent is transparent and traceable to some unsatisfied condition [180, 237] and can be defeated by the person adding new information to what the agent lacked [207, 209]. A person with a goal and a system organized around that goal, with each party able to act

and be questioned, and both responsible to the activity's development and commitment, are all fulfilling the conditions for teamwork that matters for the interaction [32, 140, 221].

2.5 Support vs. Challenge

Support that only encourages is not enough nor optimal for growth [36, 61, 259]. People sometimes need to be challenged, pushed, or told what they would not find easy, and what is considered helpful shifts between persons and also within the individual over time [141, 153]. Ryan and Deci [218] calls this "optimal challenges" and they argue that it facilitates intrinsic motivation, doing an activity for its own inherent satisfaction. Furthermore, it has been shown that positive, supportive feedback compared to negative feedback improved intrinsic motivation [58, 60].

In connection with Self-Determination Theory (SDT), Miller and Rollnick [182] posit that expressing empathy (support) comes hand in hand with developing discrepancies (gentle challenge) to motivate the person in question (Motivational Interviewing (MI) [24]). The distinction in Miller and Rollnick [182]'s approach is from where the challenge comes from, in this case, from the person recognizing alternatives rather than confrontation. In persuasive dialogue systems, MI was found to be more persuasive than argumentation [192]. Another difference between SDT and MI is that SDT is a theory on motivation whereas MI is a counseling method. SDT postulates that change may begin with external reasons but states that these reasons hold when they are internalized and experienced as the person's own. MI draws the reasons out of the person, so these two have been intertwined in research with SDT being used to understand the "how" and "why" of MI [254].

Challenges and barriers also supply a bar in which overcoming it through support can be advantageous for achieving the person's goal [195]. It is advantageous because it allows the *identification* of the barrier that can be lowered by coaching or support in general [165, 195]. Barriers also surface contradictions within the person's own activities according to Activity Theory [72]. These contradictions can be used as drivers to improve upon and develop the person's practices to achieve their objectives [243].

The main idea of challenge being a driver for positive behavior change is incomplete without the support that aids in overcoming the challenge. Both are needed for optimal growth of the person, with challenge being in the "just right" zone or, in the words of Vygotsky [259], in the Zone of Proximal Development (ZPD) which is the area that is challenging to the person but can be overcome with the proper help of a more knowledgeable peer.

In this thesis, we apply challenge or "tension" to the person, operationalized through our argumentation schemes, in agreement with the person. We also highlight that the person's preference of support over challenge is not a stable straight line. Participants who first chose maximum challenge initially for their

companion revised their selection during use, and the same person could opt for a more supportive tone in one activity while opting for a challenging approach for another. Alignment of support and challenge, therefore, is alignment to an always dynamic target. The coaching style to a person cannot be tuned only to the person's "type", so to say, but must adapt to the person's needs as it shifts. As a comparison to earlier work, Smith et al. [239] and Dennis, Masthoff, and Mellish [60] has suggested that adapting the emotional support a system contributes to a person's personality, whereas this thesis suggests that though it is a possibility, it is more of a starting point.

2.6 Behavior Change to Improve Health

Behavior change is the application domain of this research. This thesis concerns itself with how humans and artificial agents, together, comprise an activity, how authority and autonomy over decisions is shared and distributed, and how reasons can be bartered in such a relationship. Health behavior change is a research domain where autonomy allocation is vital, lines between persuasion and manipulation matter ethically, and where an extensive amount of work and theory already exist in which a new framing can be positioned against.

This section has two objectives: i) it aims to detail the state of the art in the social science and psychological theories on behavior change that this thesis draws from (Self-Determination Theory, Behavior Change Techniques, several models, and complications). ii) it argues that despite the important and extensive work already existing on behavior change interventions to improve health, there is still a gap. The gap is that behavior change describes more what interventions *does* to a person rather than describing what a system and person *do together*. To address this gap, we borrow from other theories, namely, Activity Theory, Argumentation Theory, and human-AI interaction/teaming, which are elaborated more on earlier in this chapter.

2.6.1 A brief history of behavior change

In the 1950s, free mobile units were offered to the people of the United States for tuberculosis screening. Very few actually used them [5, 107]. We now understand some of the reasons for this was the approach taken by the experts of the time. These approaches have been modeled as "Information Deficit Model" [235] and "Knowledge-Attitude-Behavior Model" [20], which both assume that if the public is given more facts, it will change their minds. Essentially, the idea was that the public behaves in unhealthy ways because they do not know better. Countless examples exist ([131, 272] to name some), even today, to prove that this assumption was flawed, and furthermore countless campaigns also showed that this was flawed. This is because knowledge itself is not a good predictor of behavior, and neither are intentions; the "intention-behavior gap", i.e., people form intentions but do not act upon them, is a widely accepted

phenomenon in behavior change literature [75, 231, 232]. In fact, medium to large change in intention, in reality, translate to smaller to medium changes in behavior [265].

So, we turned to theory. Social cognitive models were used to find the determinants mediating between information and action. The Health Belief Model [215] with perceived susceptibility, severity, and barriers; the Theory of Planned Behavior [1] with attitudes, norms, and behavioral control; and self-efficacy and expectancies with Social Cognitive Theory [11, 12]. Other developments at the time contributed with goal setting, self-regulation, monitoring, discrepancies, and feedback with ideas from control theory [42] and goal-setting theory [163].

Parallel to the development of the models, behavior change was, and still is, something that was done between people. Medical professionals and experts were helping people change through conversation before the conception of behavior change apps. Motivational Interviewing (MI) was a result of studies with problem drinkers [181], but soon was used in collaborative counseling to elicit the person's own reasons for change, by also respecting their autonomy [182, 214]. Advances in health coaching followed a similar trajectory through placing the expert as a partner in the person achieving their goals as opposed to the traditional ways of giving instructions [268]. A more recent example, the Västerbotten Intervention Program (VIP) is built around this health dialogue between a nurse and the person [194]. Behavior change has always had a relational and a dialogical side to it, which has been inherited by digital companions. In MI vocabulary, developing discrepancy and eliciting the person's own motivation, can be translated into challenge and support (Section 2.5) grounded in the person's own motives.

More recently, however, another wave of theories have muddied the waters. These new theories suggest that human behavior is not always deliberative. Dual process theory separates automatic determinants (unconscious) from the reflective (conscious) [108, 124, 246], and behavior often an automatic response tied to context rather than deliberate decision [89, 148, 269]. These findings are important from a design standpoint. It shows that interventions addressing "the reasons" may not be enough to change behaviors that are currently not reasoned.

After decades of research on behavior change there are still inconsistencies between the theories, terminologies, and conceptual overlap [57, 193]. In health psychology, two responses are dominant to the issue, both relevant to this thesis. One, is integration; combining theories under a broad framework large enough to consume the determinants of the theories (examples are IMBP [76], Health Action Process Approach [226], Integrated Behavior Change Model [100], COM-B and Theoretical Domains Framework [39, 176, 177]); and the second: standardization of vocabulary of the components of intervention and their mechanisms (examples are BCT taxonomies, Behavior Change methods taxonomy, Theory Techniques Tool, Decentralized Construct Taxonomy, Behavior Change Intervention Ontology [122, 142, 169, 179, 206]).

At the same time, human-computer interaction (HCI) and AI joined in

through technology designed to support change behavior. Fogg [81] introduced captology, or the computers as persuasive technologies. Fogg [78] describes computers' roles in persuasion which can be as a tool that makes a behavior easier, a medium in which people simulate the behavior, or a social actor that persuades through social relationships. The Fogg Behavior Model suggests that a behavior occurs when motivation, ability, and a prompt intersect at a given moment [80], akin to COM-B but made for the designers who take the decision for when and how a system should act. Oinas-Kukkonen [198, 199] took this a step further by specifying ethical and design requirements for persuasive technology aimed to support behavior change to improve health, defining them as Behavior Change Support Systems (BCSS). This shift changed the structure of persuasive systems from distinct persuasive features to socio-technical systems. Personalization followed thereafter, with the idea that individuals all respond differently to the same approach and systems can learn and adapt which strategy is suitable for which individual [126, 202]. HCI has also surfaced that behavior theories are vaguely included in the design of technology, and that these designs can likewise feed back into the behavior theories [54, 103].

A natural next step is moving from the question of *what* to deliver to *when* to deliver. The introduction of smartphones and wearables allows to gauge the person's state as well their context multiple times per day, importantly, in their daily life as opposed to the traditional clinic [233]. This new opportunity also allows for delivering support at those times [105]. Just-in-time adaptive interventions (JITAIs) aim to deliver the right amount of support at the right time based on the changes in the person's state [186]. A JITAI can be defined by its parts. *Decision points* are the moments at which the system decides to act, *tailoring variables* are the information about the person (e.g., location, stress levels, activities), *intervention options* are what can be delivered, which includes non-communication, and *decision rules* which is the connection of tailoring variables to an option. Delivering support "just-in-time" is vital because delivering unsolicited support, possibly at a time when the person is busy or indifferent to it, can come off as annoying or be ignored. JITAIs aim to time the delivery of support at moments of opportunity where the person seems to be receptive to it [146, 186].

The idea of JITAIs resonates well with the approach we take in this thesis as the agent also adapts over time to a person who is dynamic. The difference between the approaches, however, originates from who the decision maker is. In a JITAI, the decision rules are composed by the designer or learned by the system. The person is more a source to feed the tailoring variables. Furthermore, the person has the choice to ignore a prompt sent by the system but they cannot question the reason for why the prompt was sent nor alter the rule that sent it. JITAIs also do adapt what is delivered and when it is delivered, but unlike the approach taken in this thesis, they do not typically adapt the agent's position and stance toward the person (i.e., cannot decide whether to support or challenge). Compared to JITAIs, in our work, the rule for how the agent should act and behave is something the person and agent can argue about. In

activity theoretical terms, a JITAI adapts the tool to the state and situation of the subject whereas the companion also negotiates its place in the division of labor.

In this thesis, the health psychology branch lives within the content, BCTs and their MoAs, being delivered by the system. The persuasive branch places the system as an actor with a particular behavior and stance toward the person. However, the origin of these branches takes systems as things *acting on* the person whereas the aim should be to act *with* the person. Ogden [197] warns against systematization, stating that this approach can risk losing variability in individuals and their context, and variability, can be used as a target for alignment, as this thesis does.

2.6.2 SDT, TTM, and Activity Theory

From the massive expansion of behavior change theories described in the previous section, two of them majorly carries into this thesis. The first one is Self-Determination Theory (SDT) [220] which provides the criteria for the outcome and asks the question of *what kind* of motivation rather than *how much* [59, 218, 220] of motivation is needed for behavior change. SDT assumes three human needs underlying motivation: *autonomy* (experiencing one's actions as self-authorized), *competence*, and *relatedness*. These requirements are held to be essential as opposed to simply desired for motivation. Motivation, runs as a continuum between *extrinsic* motivation and *intrinsic* motivation, running through *external regulation* (reward-punishment) through *introjection* (guilt) to *integrated regulation*. The main point here is that autonomous regulation is a predictor of maintenance whereas controlled regulation (external and introjected) predicts short-term compliance with relapses occurring in between [191, 219, 247]. This point is an important distinction as many of the interventions fail at the maintenance stage.

The Transtheoretical Model (TTM) [210] provides the temporal aspect. According to TTM, change is a movement through precontemplation (no intention to change within 6 months), contemplation (intention but ambivalence), preparation (plan to take action within a month), action (actively modifying behavior for less than 6 months), and maintenance (sustained behavior for longer than 6 months), with relapses being treated as normal instead of failure. The main point for the TTM is the idea of stage matching which is customizing treatment or interventions according to the person's specific readiness to change and stage of change. The TTM has been used in research as a way to provide a way for systems to talk to persons about where they are in a change process and a way to vary strategies over time [46, 49, 227, 264].

Activity Theory also plays an important role in the development of autonomy, competence, and relatedness. It supports the satisfaction of these three needs through tool-mediated activities. Tool-mediated activities provide the subject with the means to direct their own actions, make choices, and also pursue their goals using the tool. Tools that are shared and community practices

enable collaboration and join personal efforts toward a collective object, which builds relatedness. Furthermore, tool-mediated activities provide opportunities for feedback and skill building through tools to support competence building.

2.6.3 Behavior change support and techniques

According to Oinas-Kukkonen [198], *a behavior change support system (BCSS) is a socio-technical information system with psychological and behavioral outcomes, designed to form, alter, or reinforce attitudes, behaviors, or an act of complying, without the use of coercion or deception.* Oinas-Kukkonen, by his defintion of BCSS, "bans" coercion, which is a distinction from Michie's approach where coercion is a valid intervention function [177]. Furthermore, Oinas-Kukkonen and Harjumaa [200] supplies a design model, the Persuasive System Design (PSD) model, which organizes 28 principles into: primary task support (reduction, tunneling, tailoring, personalization, self-monitoring), dialogue support (praise, rewards, reminders, suggestion, similarity, liking, social role), system credibility support, and social support. Furthermore, the principle of personalization has become a line of research in itself within behavior change support (systems) [171], and central to the current discourse on AI alignment. For this thesis, the "dialogue support" category is key for what it does *not* contain, The principles outlined by Oinas-Kukkonen and Harjumaa [200] are surface level conversational moves; the PSD aims to model the persuasive stance of a system toward a person, but not dialogue, in a technical sense. "Without the use of coercion or deception" found in the definition of BCSS above, surfaces the ethical question in persuasive technologies [19], which includes pressure from a system and when it turns into coercion [125], an important thing to note when considering digital companions designed to challenge.

Building these systems, however, requires addressing other issues, namely, identifying the ingredients of an intervention, which is what Behavior Change Techniques (BCT) [179] aimed to identify. BCTs are the replicable and observable components of an intervention, originally identified in interventions delivered by health professionals. The BCT Taxonomy v1 is the current standard, with 93 clustered observable and irreducible techniques such as goal setting, feedback, social support, habit formation, credible source, etc., devised with an international expert team [179], that can be combined together for a more encompassing approach [178]. As a coupling to the BCTs, Mechanisms of Action (MoAs) [40, 53] (e.g., goals, social influence, skills, intentions, etc.) explain the *why* these techniques work. MoAs are the process through which a BCT affects behavior, e.g., beliefs about capabilities, self-image, emotion, etc. The link between MoAs–BCTs were evidenced by Carey et al. [40] and an expert consensus on the plausible links were given by Connell et al. [53].

Moreover, this section would not be complete without touching on Multiple Health Behavior Change (MHBC) simply due to health risk behaviors clustering; they do not occur singly [212]. MHBC studies the modification of

two or more related behaviors simultaneously or sequentially [212]. Sequential approaches take longer but reduce the burden on the person, the opposite of simultaneous approaches which can sometimes overwhelm. Though the effect size of MHBC approaches is marginally smaller compared to single behavior focused approaches, the aggregate risk reduction is larger [115, 267]. Furthermore, the idea of coaction, effective action in one behavior can indicate effective action on another behavior [121], can be an effective tool in informing an agent about readiness to change in a person through inference based on a change in another behavior domain.

2.6.4 Stress and recovery

Stress is another dimension in this thesis. One in six people in Sweden report feeling stressed, more so for younger people [82], and stress induced illnesses are the leading cause of sick leave [83]. Recovery research describes how people replenish the resources that stress siphons. An activity is recovering only if it brings about at least one of four recovery experiences: psychological detachment, or mentally switching off; relaxation, a state of low-arousal; mastery, the experience of challenge or learning; and control, deciding for oneself what to do and how [241, 242]. Recovery activities can further be characterized along several dimensions, from physical to social to spiritual [3], and short activities of a few minutes can already improve wellbeing [15]. Conversational agents such as Woebot and Wysa deliver short, therapy-based exercises through chat [114, 211], and recent work calls for keeping a human in the loop for such agents, especially in crises [229].

There are two points that make recovery relevant to this thesis. The first is that the domain has its own version of the problem a companion is meant to address. In what [240] calls the recovery paradox, the people who need recovery the most are the least able to find and carry out recovery activities, because stress drains the resource needed to do so. Support in finding, choosing, and keeping activities is therefore a part of the problem. Second, recovery is defined by the experience an activity produces and not its type. A walk can be relaxing for one person and a source of mastery for another, and it is not recovering at all if it produces none of these experiences. This is another form of the behavior-activity link (Section 2.3.1) where two people logging the same activity may be carrying out different activities, and only the motive and the experience tell them apart. Assessing the experience close to the activity itself, through ecological momentary assessment (EMA) [233], is one way to capture it, and is the approach taken in Generation IV (Chapter 4).

2.7 Summary

This chapter brought together the theories and frameworks that the rest of the thesis uses to build its work upon. Section 2.5 conveys that support alone is not

enough for growth and that challenge delivered, in agreement with the person, is a part of what makes change possible, and the balance between support and challenge is a dynamic target of the person.

Behavior change (Section 2.6) gives us the empirical domain and vocabulary. Behavior change theory itself does not supply a unit of analysis in which a person and an agent do something together but what an intervention does to a person. To address this gap, we employ Activity Theory (Section 2.1) by treating behavior as an object-oriented activity mediated by tools, rules, community, and division of labor, relocating barriers from deficits to contradictions in the activity system. Argumentation Theory (Section 2.2) equips the agent with the ability to act on that model by allowing it to posit its stance as a scheme backed argument with premises that can be questioned and challenged.

Section 2.3 ties behavior change, activity theory and argumentation theory together. Critical questions from argumentation theory takes relations that activity theory surfaces, and gives them a way into conversation that the agent can use. Agreements and disagreements in dialogue showcase a person's shifting goals as updates, and, with that, the agent's behavior becomes traceable to the specific dialogue.

Finally, Section 2.4 re-maps the agent from tool to a teammate within the activity system. The agent moves from the tool node to the division of labor node, sharing the object with the person and becomes accountable for it, making the agent a stakeholder in the person's object, i.e., a team member.

Chapter 3

Methodology

This chapter outlines the methodological approaches used to answering the research questions (Table 3.1). The work in this thesis follows a constructive, "research through design" approach applying participatory design engaging domain experts in particular. Frameworks, formalizations, and artifacts were built, and then studied as participants used them in real life. Paper I introduces argumentation schemes and a role taxonomy and tests them as requirements. Paper IV formalizes those schemes into a dialogue protocol and implements it for daily use. Paper V extends the architecture into generative multi-agent system and asks whether it holds up in the wild. Paper VI returns to a single generative agent whose role the person can choose or create, and follows how participants perceive and team up with it over up to 12 weeks. In each of these cases, the prototype is the vehicle through which the aspects of the research challenges explored. The work spans four generations of systems, described in Chapter 4: Generation I, a rule-based argument-based architecture (Paper I), Generation II StarCoach, a rule-based app with a configurable coach (Paper II and IV), Generation III Blumeai, embedding a swarm of LLM-agents (Paper V), and Generation IV Aime, an LLM-based recovery app with customizable companions (Paper VI).

Table 3.1: Methodologies, studies, and methods used to address each research question.

Study	**Methodology**	**Methods** (data collection and analysis)
RQ1	Theory and empirical insight informing the agent's strategies and model	
S1, S2, S6	Participatory design; knowledge engineering	Three iterations with healthcare partners and a twelve-person multidisciplinary expert team. **S1**: purposes, use cases, and high-level architecture defined. **S2**: conceptual models, mock-ups, and early prototypes developed and revised, with co-design sessions with domain experts and target users. **S6**: the experts used prototype versions for days or weeks between regular meetings and commented in a shared document; the prototype was updated between meetings. Domain knowledge was elicited and modeled in ACKTUS and reviewed by the experts and 65 specialized nurses; 33 BCTs were identified and clustered by purpose.
S3, S4	Survey study	**S3**: online questionnaire with 40 adults in Scandinavia, aged 29–60, on perceptions of behavior change in five domains, of digital companions as social actors, and on preferred roles. **S4**: questionnaire with 82 adults relating readiness to change to preferred roles and behaviors, with nine dialogue scenarios rated for the agent's behavior and role. Quantitative analysis of preference patterns; thematic analysis of open answers.
S5	Formative expert evaluation	Nine experts in epidemiology, public health, nutrition, nursing, social work, and ethnology, over three sessions: individual use of a prototype with observation and interviews; a workshop in pairs running five dialogues and describing how they should ideally unfold; a group evaluation of the revised dialogues.
–	Conceptual analysis; theory integration	Support-challenge framework and the APS and APCT schemes built from S3–S5; activity-theoretical framework of three interacting activity systems and five design concepts, exemplified by StarCoach; activity-theoretical concepts mapped onto dialogue constructs (motives as grounds, barriers as counter-arguments).
RQ2	Formalizing and implementing support and challenge	
–	Formal modeling	Critical questions for APS and APCT; dialogue protocol and seven rules; support-challenge level computed from role weights and the person's preferences, with a default weighting toward support set from S5.

Table 3.1 (continued)

Table 3.1 (continued)

Study	Methodology	**Methods** (data collection and analysis)
–	Constructive design and prototyping	Four system generations: i) a prototype with argument-based dialogue scenarios; ii) StarCoach, with a rule-based coach whose statements are grounded in expert-approved content and person's preferences; iii) a swarm of LLM-based agents with a triage agent modeled as a finite-state machine with preference-weighted transitions, role agents carrying a cluster of BCTs, RAG based on national guidelines, and four log-based coordination metrics; iv) Aime, with a genAI agent in predefined or person-created roles, an expert-authored fallback activity bank, and a safety filter. i, ii, and iv were all co-designed with experts.
S7, S8, S9	Analysis of system behavior in use	**S7**: the formal model used to trace breakdowns (which could lead to a participant's disengagement) to a specific dialogue rule. **S8**: coordination metrics computed from logs to test whether the agents divided the work as designed. **S9**: agent conversations coded turn by turn.
RQ3	Experiences of collaborating with the agent(s) in daily life	
S7	Qualitative field study (onboarding) of StarCoach	Four weeks with five participants and three VIP nurses; semi-structured interviews and observation over three sessions; usage logs. Analyzed through: i) inductive thematic analysis of the app with descriptive usage statistics; and ii) focusing analysis of the agent's appearance, sociality, behavior, and participants' expectation through the activity system.
S8	Mixed-methods in-the-wild feasibility study of Blumeai	Four weeks with 18 participants, six returning and twelve new; semi-structured interviews and observation at baseline and follow-up; scripted questions posed to the swarm; TAS (descriptives); coding of dialogue topics and of reactions to the agents' proposals.
S9	Multiple-case study of Aime	Four participants over two weeks, two of them for 12 weeks; logs and in-app ratings (when and how they were input); Godspeed, Perceived Cooperativity, and Teaming Perception scales; usability and acceptance questionnaires; 10 semi-structured interviews; within-case and cross-case analysis.

Where the studies are reported. S1–S2: [159, 162], summarized in Paper II; S3–S5: Paper I; S6: Paper II; S7: Papers II and IV; S8: Paper V; S9: Paper VI.

3.1 A Participatory Design Approach

Participatory Design (PD) [27, 94, 225] gives future users and domain experts a direct say in design decisions throughout the process. It was chosen because StarCoach was in connection with VIP which situated the clinicians as co-owners of the design [158]. Domain experts hold the power to veto over what the agent could say and what content could be embedded in StarCoach. PD is the methodological umbrella of this thesis. StarCoach is the application that

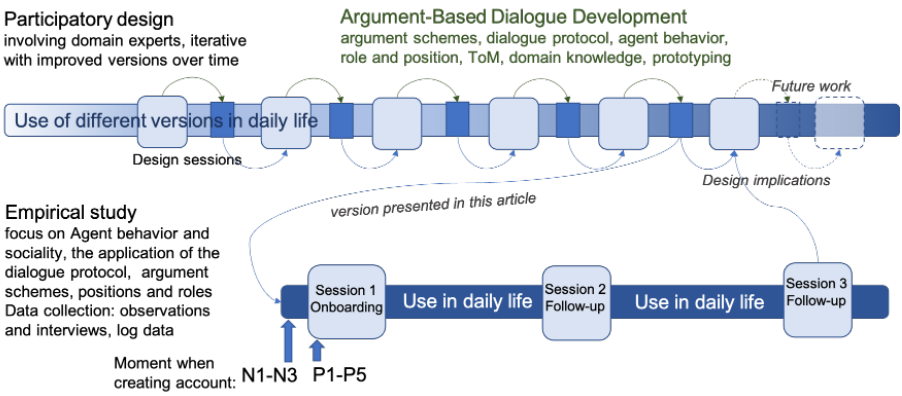


Figure 3.1: Participatory design process of StarCoach and the 4-week onboarding study. Paper II [158] documents the process. The onboarding is analyzed at the app level in Paper II and the agent level in Paper IV. Taken from [135] under the Creative Commons Attribution (CC BY).

Papers II, III and IV all depend on in some form, and Paper I feeds into. It is also a product of the PD process run over multiple iterations with a twelve person multidisciplinary expert team ranging from medicine, nursing, psychology, ethnology, social work, epidemiology, nutrition, and health economics. The team also includes people in the target group of the Västerbotten Intervention Program (VIP) [158, 194]. Paper II can be seen as a documentation of the process. Though Paper IV does not employ PD directly, it still consumes the output. In the case of Paper I, its formative sessions with domain experts feed into what the PD process in Paper II would refine. Paper VI continued with the PD approach where Aime, the app of Paper VI, was co-designed by experts in cognitive psychology, stress rehabilitation, occupational therapy, human-computer interaction, and AI [134].

3.2 Study Phases

Paper I is the initial phase where two online questionnaire studies were conducted via Prolific (122 participants across both studies in Scandinavia, aged 29-60), and a multi-session study with nine domain experts using an early prototype [136]. The purpose of this phase was to explore particular aspects from potential users' point of view before studying the system in daily use. The explorations were about what roles, what balance of support and challenge did people expect from a digital companion, and how did this differ between people who had read scenarios versus the experts who had used the prototype.

This establishing work fed into the PD process documented in Paper II which turned the elicited design proposals, which were negotiated and evaluated

iteratively, into StarCoach. Once a first version of StarCoach was completed, we ran a four week onboarding deployment with five participants and three VIP nurses. Paper II analyzes this study at level of the design process and how the modules were used, what factors led to adherence, how the multidisciplinary expert team's decisions actually played out in use. Paper IV looks at this study from the level of the agent dialogue behavior; how the formal argumentation schemes performed, and whether the CQs of these schemes held up with what participants did [135, 158].

Paper V is the third phase after the architecture was complementing the rule-based approach with an LLM-based swarm. The four week feasibility approach is similar in structure of Paper II and IV's studies with an onboarding period, semi-structured interviews, and observation (see Figure 3.1). The difference was that we expanded methodologically as the questions needed to be answered are essentially different from when considering a single dyadic agent [137]; measuring the coordination of several agents as a team, and the person's experience connected to that called for an instrument the earlier generations did not need (Teamwork Assessment Scale and coordination metrics in Paper V).

Paper VI is the fourth phase. It uses a multiple-case design with replication logic [271]: four adults followed for two weeks, two of them for 12 weeks, with each person's use of Aime and their experiences as the unit of analysis [134]. Each case combined app logs, in-app ratings, agent perception and HAT scales, usability and acceptance questionnaires, interviews at each timepoint, and the agent conversations coded for each turn. Cases were analyzed one by one and then compared, and the eight propositions the paper reports were formed from patterns in the cases and checked against each case for literal or theoretical replication.

In summary of the four phases, the arc was exploratory research, co-design, to feasibility and finally, a longer, case-based study of a single generative companion in everyday use. In conjunction with these papers, theoretical analyses and development was also done, most notably in Paper III.

3.3 Data collection and analysis: Constants and Change

Across all the studies that require the use of a system, this thesis follows a qualitative and exploratory approach consisting of semi-structured interviews, observation, and thematic analysis [33, 34] to get to the core of what people experience and feel when using these systems [135–137, 158].

The sample sizes are tailored to the qualitative study designs, and to carefully and qualitatively explore a range of different viewpoints is justified as deploying such technologies are set in high-risk environments (healthcare in this case) and much work needs to be put into exploration and formation [96, 104, 222]. A sample size is adequate when the study's aim is well defined and

the sample is specific to that aim, with theories to draw from and the resulting dialogues and analysis are of a sufficient quality [96, 104, 222].

Paper I's quantitative work measured stated preference before use based on the 122 participants who had read but not experienced the scenarios. Papers II and IV gives descriptives that are reported alongside the qualitative findings. Paper V introduces metrics for the swarm and person that the other studies did not require and it should be noted that this one is different in nature to the other studies since this is a point where the system's behavior is also a primary source of evidence compared to only the person. Paper V also utilizes the Teamwork Assessment Scale (TAS) [133] as we could not find a human-AI teaming instrument that fit at the time. This was also a contribution in itself as it was also used to test this instrument's fit for this context.

How we evaluate the agent, the interaction, and teamwork changed in Paper VI compared to Paper V. Firstly, new HAT instruments, the Perceived Cooperativity Scale and the Teaming Perception Scale [9, 10], became available, so the agent was rated using them, alongside the Godspeed likability and perceived intelligence scales for assessing how the participants perceived and experienced the agent [13]. In-app ratings as well as logs were explained for when and how they were input (in the moment vs. in batches later) and compared with the interview accounts. Additionally, the System Usability Scale (SUS) [35] was used to measure ease of use of Aime, nine items adapted from UTAUT2 [256] to understand persons' acceptance, adoption, and rejection of Aime, and finally a questionnaire on trust and integration written for Paper VI were administered.

Interviews were coded against a set of topics and then compared with the same person's logs, ratings, and conversations, with discrepancies treated as findings. The conversation transcripts and log timestamps were the only source in Paper VI that was not self-reported, and they often explained what the ratings and interviews did not.

Chapter 4

From Theory to System

The theory introduced in Chapter 2 lay the ground and framing and feed into the decisions about design in this thesis alongside the outcomes of the participatory design process. These decisions are apparent in the code, data model, and in the interaction design. This chapter summarizes four generations of agent-based behavior change systems developed as part of the research presented in this thesis. This chapter also highlights the changes from generation to generation. Further details can be found in the articles.

Paper I [136] supplies an argument construction architecture using the ACKTUS knowledge base, an ontology based platform for modeling and managing knowledge bases [161]. It also introduces the notion of roles and explores participants' and domain experts' view on how a digital coach with different roles could contribute to lifestyle changes. Papers II [158] and IV [135] give the design and the agent-level implementation of *StarCoach* respectively. Paper V updates the architecture with a swarm of generative agents in the *Blumeai* application. Paper VI [134] returns to a single GenAI agent in the *Aime* application designed for supporting a person to find and evaluate recovery activities

	Generation I	**Generation II**	**Generation III**	**Generation IV**
System	Argument construction architecture	StarCoach	Blumeai	Aime
Paper	Paper I	Paper II, Paper IV	Paper V	Paper VI
Implementation	Rule-based, OWL/RDF, AIF	Rule-based, OWL/RDF, AIF	LLM agents, RAG for guidelines	Single LLM agent, expert activity bank
Positioning	Computed from User model	Person, explicit	Person's weights and classifier	Person, by choosing or creating the role
Utterances	Templates, slot-filling	Templates, slot-filling	Generated	Generated
Inspectable	Full-argument chain	Full argument chain	Routing decisions	Active role only

Table 4.1: The four system generations.

and monitoring everyday stress. An overview of the four generations are given in Table 4.1. The next generation of the work is outlined in Chapter 7.

4.1 Generation I: Exploring Agent Roles, Micro-Dialogues with Arguments and Knowledge Base (StarCoach prototype)

Generation I is the outcome of three studies in Paper I [136], which explored what people expect from a digital coach and how its role and behavior could be tailored to them. In two online questionnaires, 122 participants in Scandinavia chose which role a digital coach should take (assistant, coach, expert, or companion). In the second questionnaire, 82 of the 122 participants also chose how challenging, empathic, and friendly it should be and rated scenarios between an agent and two personas. In the third study, nine domain experts used an early prototype across three sessions and edited the dialogues based on their expertise. The expert sessions and outcomes from the studies with 122 participants were used to develop the architecture shown in Figure 4.1. The results of the study for Generation I is summarized in Section 5.2.

The first generation is an architecture which uses a pipeline starting with the agent. A dialogue commences by the person or the agent. The Construct Arguments module retrieves the goals, values, activities, and prior arguments associated with the person, altogether with the contextual data in the knowledge base. The result is then passed into a Formulate Arguments module which writes into the repository so that it can be used in further dialogues. An Evaluate Arguments module then assess the stored arguments, and finally, a Compute Position module combines the coach's role and the person's current situation to decide whether what reaches the interface is a supporting argument or a counter argument. When there are no more arguments left to be constructed, the dialogue is closed through a component to give the dialogue an ending [136].

There are some specific decisions taken when building this pipeline that are based on theory (Sections 2, 2.1, 2.2). The first is the data. The knowledge base is built on the ACKTUS platform [161], whose ontology extends the International Classification of Functioning, Disability, and Health (ICF) with sub-concepts under Personal Factors and under Activity and Participation [201]. Information nodes in ACKTUS carry values such as how important, and how fun an activity is, and how prepared the person is to change. The ACKTUS ontology implements the activity ontology, which is based on Activity Theory, the nodes store motives and values of the behavior that makes the behavior meaningful to the person [129, 150, 160].

Furthermore, as a second decision, the arguments in ACKTUS are stored as objects instead of strings. The ACKTUS ontology embeds a modified version of the Argument Interchange Format (AIF) in which an argument is a combi-

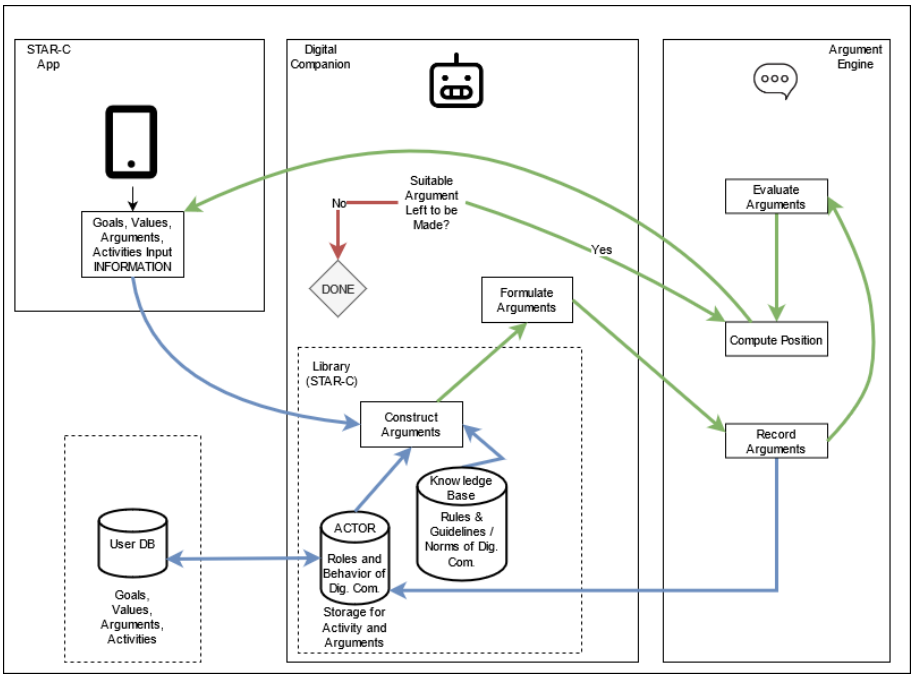


Figure 4.1: Argument construction pipeline of Generation I. Arguments are constructed from the person's model and the domain knowledge base. Taken from [136] under the Creative Commons Attribution (CC BY).

nation of premise nodes connected to a conclusion node in a graph database [48]. Each combination of the premises and conclusions is associated with an argumentation scheme which is also modeled in the platform.

The final decision was the conclusion type determining the kind of dialogue that would take place. A conclusion can be a decision, an activity of what to do next, or advice. These then correspond to dialogue types of inquiry, deliberation, and persuasive or supportive dialogue respectively [262].

This type of dialogue generation, however, lacks the way of deciding how much to challenge the person. The Compute Position module combines the role with the situation but the grounds on *which* support over challenge is not formalized. The formalization is delivered in Paper IV [135].

4.2 Generation II: StarCoach

Following the prototype in Generation I, we developed StarCoach, a mobile application that supports behavior change across five lifestyle domains for the prevention of cardiovascular disease. One domain is balancing life relating to monitoring stress, others are physical activity, nutrition, alcohol use and

tobacco use. The app is designed and developed through a participatory design process engaging twelve domain experts and participants in the Västerbotten Intervention Program (VIP) [158, 194]. Chapter 3 discusses more in detail about the participatory design method.

4.2.1 Four modules

The app is organized into four modules emerging during the participatory design process: the Star Profile, the Activity module; the Digital Coach module and the personalization module [158].

The Star Profile shows the person's status across the five domains in a single form, in the shape of a star, with each domain being connected to person-tailored evidence-based advice taken from the Swedish national guidelines. The Activity module allows for goal setting through value directions, or high level behavioral goals defined by the person themselves connecting to the five lifestyle domains, and through "baby step" activities towards the goal defined by frequency of the activity, the duration, importance, fun, and with whom the activity is conducted with. Activities are tracked on a checklist that is color coded and can be reviewed across daily, weekly, and monthly time frames. The Digital Coach module provides reminders, feedback, and social and emotional support through an embodied avatar embedded in the app. Finally, the Personalization module tailors the information and coaching style to the person, based on their profile, preferences, and readiness to change [210].

Within the four modules, 33 BCTs are embedded, ranging from goal setting, feedback and self-monitoring, social support and habit formation, which are then organized into target MoAs such as motivation, self-efficacy, and goal commitment [53, 179]. The architecture of the StarCoach app is given in Figure 4.2. The Star Profile and Activity modules as they appear in the app is shown in Figure 4.3.

Participants treated the Star profile as a portrayal of who they were trying to become, and making it turn greener became the overarching goal that organized their other activities [158]. Presenting the five lifestyle domains in a single structure turns the clinical assessment angle into an object of the person's own activity, indicating the distinction drawn in Paper III [127] of endogenous and exogenous behavior change. The same goes for value directions and baby step activities and goals. Letting the person set their ambition level themselves rather than being imposed from the guidelines positions the system's support from externally guided support to self-directed change.

4.2.2 The supportive and challenging coach

In StarCoach, there are five roles the person is able to choose from for their agent: assistant, coach, expert, companion, and alter ego (challenging). The agent's approach is based on two questions: how empathic it should be, and

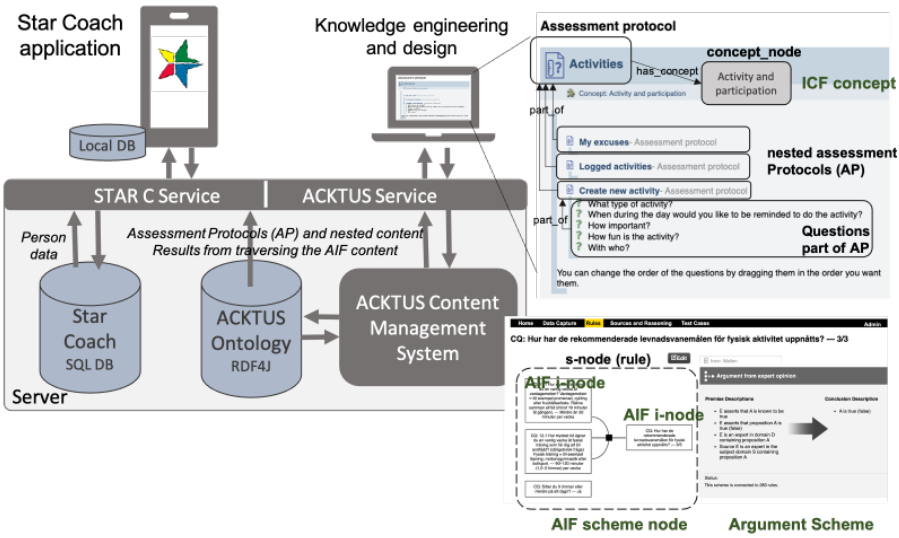


Figure 4.2: The StarCoach architecture, including ACKTUS as the content management system. Taken from [158] licensed under the Creative Common Attribution (CC BY).

how challenging (Figure 4.4). The answers to the questions are then used as weights in the dialogue.

The connection between role and weight sets a supportive weight ω_s and a challenge weight ω_c. The resulting challenge level determines whether the agent instantiates the meta-level argument schemes *Argument from Position to Support* or *Argument from Position to Create Tension* when answering the first critical question of each scheme [135]. The alter ego role, for example, at maximum challenge chosen in the personalization module is assigned the weight $\omega_s = 0.2$ and $\omega_c = 0.8$, which directs the agent toward being more challenging. The weights and the agent roles can be seen in Table 4.2.

The person opens a dialogue when entering the home screen and when assessing goal achievement. The agent opens dialogues in the morning, at lunch time, and at dinner time, based on the person's given information about these times. The argument schemes were implemented through ACKTUS, with instances of arguments related to their schemes and agent roles and behaviors connected to those schemes [135].

The thing that makes challenge useful and not annoying to the person is the origin of the content being delivered. When the agent "challenges" a person on an activity, say a planned walk, the grounds the agent uses is the motive the person themselves provided, in this case them wanting more energy during the day, with the addition of effects of everyday movement taken from the national guidelines [135]. Essentially, the agent holds the person accountable to what

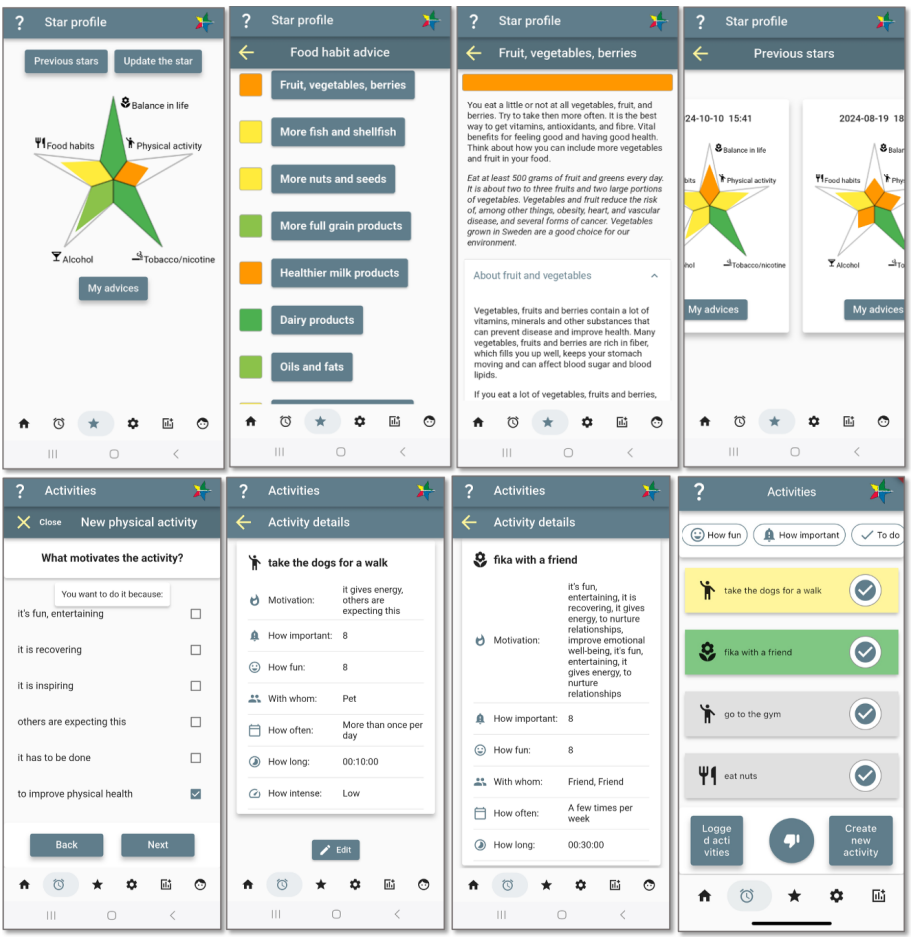


Figure 4.3: The StarCoach app with the Star Profile at the top, followed by the Activity module with the value directions and baby-step activities and checklist. Taken from [158] licensed under the Creative Common Attribution (CC BY).

Role	ω_s	ω_c	Scheme
Supporting Companion	0.8	0.2	APS
Challenging Alter Ego	0.2	0.8	APCT
Expert	0.6	0.6	APS / APCT
Coach	0.4	0.7	APS /APCT
Neutral Assistant	-	-	None

Table 4.2: Agent roles and their supportive/challenging weights with the scheme for each combination directing the agent. Taken from [135] licensed under Creative Common Attribute (CC BY) with the addition of the "Scheme" column.

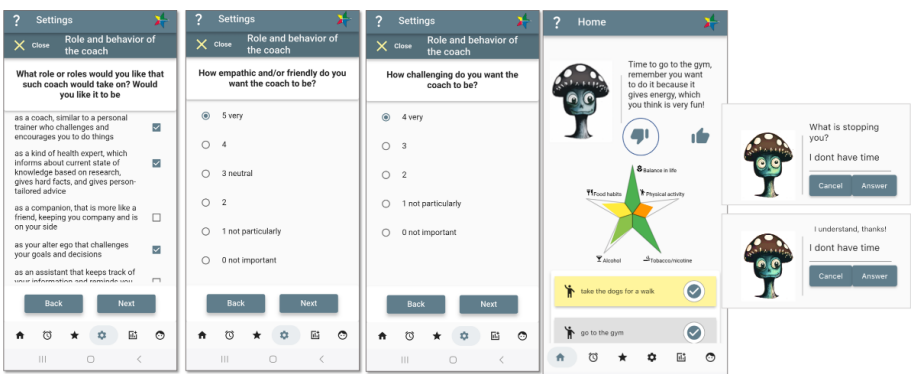


Figure 4.4: Left (3): Personalization module of the coach's role (1), empathy and challenge (2,3). Right (4): A micro-dialogue with the agent where the agent is suggesting an activity grounded in the person's own motive and follows the refusal with a question about what is stopping them. Taken from [158] licensed under Creative Commons Attribute (CC BY).

they stated as a collaborator without being seen as outside pressure.

The embodiment decisions taken also try to follow this thread for personalization. The agent can take the form of an animal or a mushroom, respond non-verbally to how the person answers, and can be put to sleep at the person's behest [158]. The sleeping function added to the agent came out of design sessions and is a good example of appropriation in that the person can decide when the agent is present in their activity.

An overview of the process of the agent's foundations to utterance can be seen in Figure 4.5.

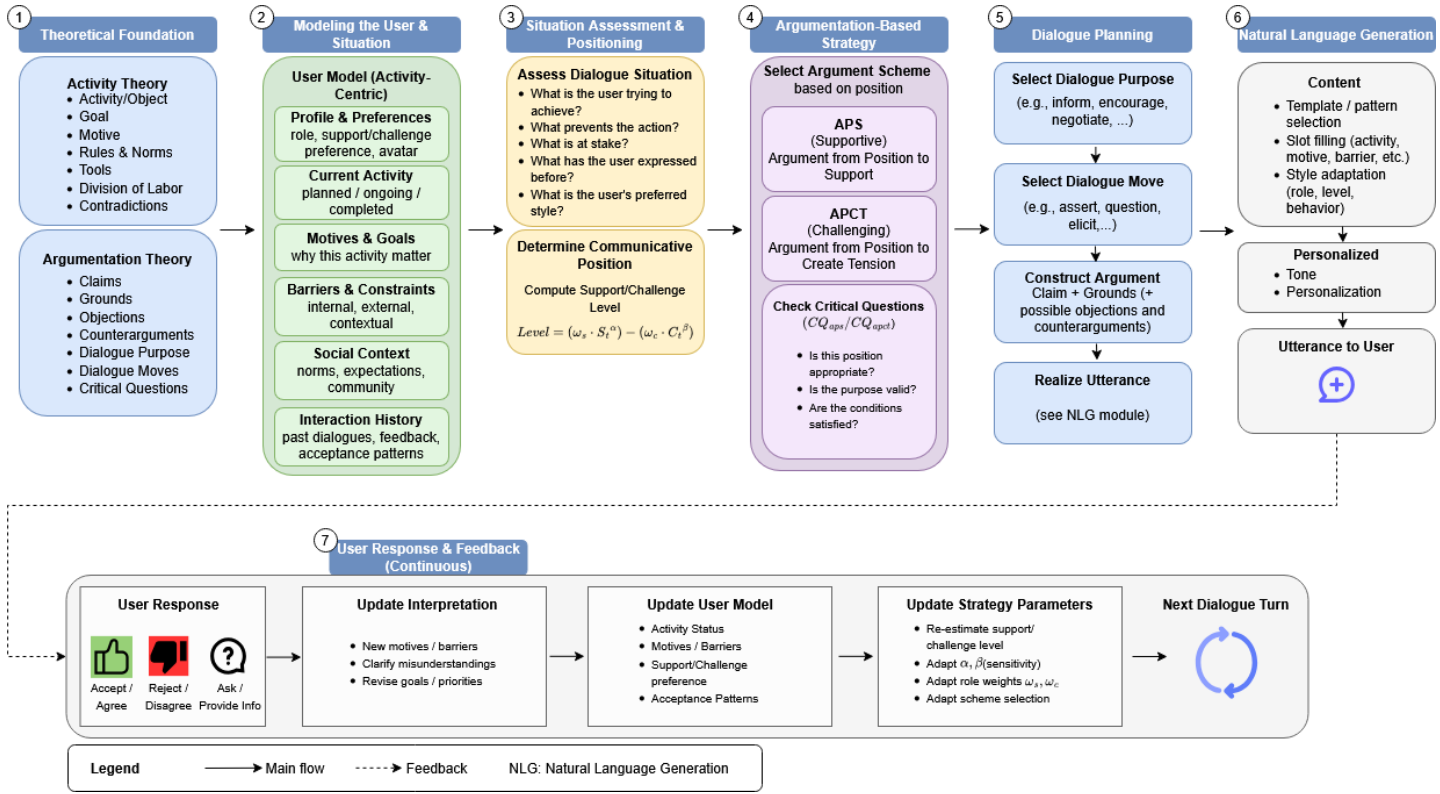


Figure 4.5: Design Process from theoretical foundations to conversational action.

4.2.3 The transparency expectancy

The agent in StarCoach is rule-based, using OWL/RDF ontologies and AIF-based dialogue structures. The utterances are produced via templates populated from the user model and from expert-authored content. This approach was taken as it is a way to traceably and transparently make decisions in line with ethical AI guidelines, and because the responsible domain experts wanted to be able to evaluate and control the statements delivered by the agent. The domain experts viewed that a rule-based system carried less risk of producing dubious advice related to health as opposed to genAI [158].

That is to say, in Generation I and Generation II, every sentence the agent conveys can be traced back through the scheme it uses, the critical questions it answers, the premises drawn from the user's model, and the guideline content that informed the advice. On one hand this limits the agent's language by the content written by the domain experts, and the support/challenge parameters remained static until the person updated it themselves, but on the other hand, allowed for a safer approach during the studies conducted with StarCoach.

4.3 Generation III: The Blumeai Swarm

Compared with StarCoach, *Blumeai* supports four of the five lifestyle domains: physical activity, recovery, nutrition, and alcohol, for cardiovascular disease prevention, and complements the single configurable coach with a swarm architecture accessible through a chat [137]. The change in architecture came out of empirical findings that different people preferred different coaching roles and communication styles, and a single agent tasked with employing all the roles risks inconsistency and identity confusion. Rather than having one agent shifting its role and style, several agents hold one role each, and the system routes between them.

Though, technically, the amount of agent that can be in a swarm architecture is unlimited, we opted to use four specialized agents based on the roles preferred by the participants in earlier studies: the empathic companion, the persuasive challenging coach, the medical domain expert, and the neutral assistant. A triage agent is also present, tasked with routing between the agents, and a Retrieval Augmented Generation (RAG) agent [154] that injects national guideline content to ground the specialized agents in expert written content and infuse their utterances with this content. All the agents have access to general health knowledge with the expert agent covering the medical specifics of the four lifestyle domains. Furthermore, if a specialized agent thinks another specialized agent is more suited to reply to a query of the person, it can then transfer to the other suitable agent. The swarm architecture of Blumeai can be seen in Figure 4.6.

The system is formalized as a finite state machine whose states correspond to the triage agent and the four specialized agents, and whose input alphabet is a set of semantic signals: emotional, motivational, medical, factual, and re-

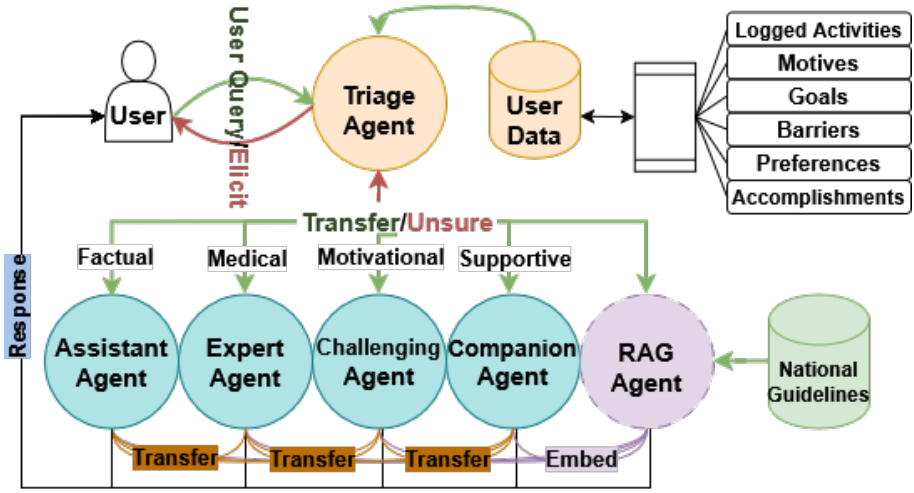


Figure 4.6: The Blumeai swarm architecture. The Triage agent routes queries to one of four specialized agents, with the RAG agent supplying guideline content. The person's activity context taken from their user model is available to all agents. Taken from [137] licensed under the Creative Commons Attribution (CC BY).

route, inferred from the person's query by the language model. The transition function takes the current state, the signal, and the person's preference weights over roles. If there are several agents as candidates for a signal, the one with the highest preference weight is selected. In the case where one agent is the only candidate, it is selected regardless of preference. When no agent matches, the person is transferred back to the Triage agent. Figure 4.7 shows the Blumeai app. An interesting paper as some background from Masthoff [170] on group recommendation shows a similar approach in the reverse way where one system serves multiple people whose preferences also shift.

Two functionalities have been carried over from Generation I and II into Generation III. First, the person's role preferences still continued to be input by the person, similar to the empathy and challenge scales from StarCoach; the functionality changed from weights over schemes to weights over agent, however. Moreover, BCTs and MoAs were mapped onto role specific agents in Blumeai, keeping the 33 techniques [53, 179] in StarCoach intact. The techniques and their mechanisms of action can be seen in Table 4.3.

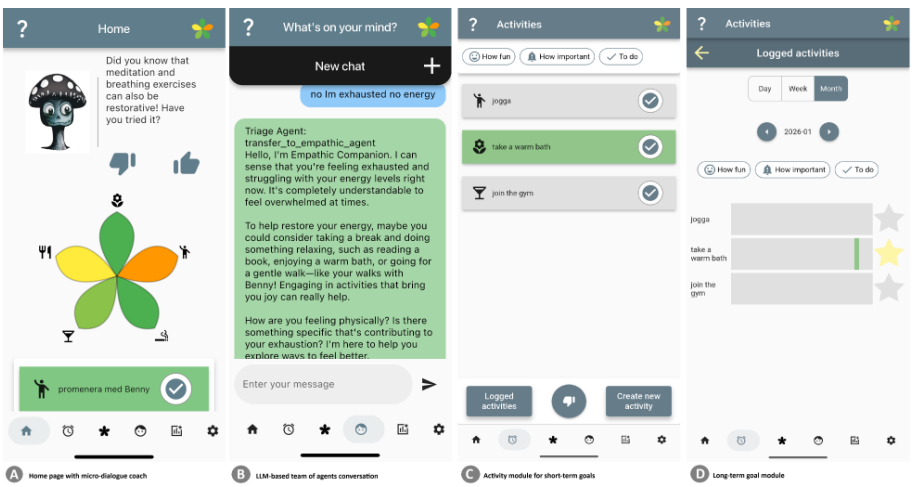


Figure 4.7: Blumeai app pages. (B) shows a dialogue with the transfer notice as responsibility is shifted to another agent. Taken from [137] under Creative Commons Attribution (CC BY).

Table 4.3: Overview of the BCTs embedded in the different modules of Blumeai. Taken from [137] licensed under the Creative Common Attribute (CC BY).

Behavior Change Technique (BCT) (Michie et. al. 2015)	**Flower profile (long-term goals)**	**Activity module (short-term goals)**	**Micro-dialogue agent**	**LLM-based agent team**
1.1. Goal setting (behavior)		✓		✓
1.2. Problem solving	✓	✓		✓
1.3. Goal setting (outcome)	✓	✓		✓
1.4. Action planning		✓	✓	✓
1.5. Review behavior goal(s)		✓		✓
1.9. Commitment		✓	✓	✓
2.2. Feedback on behavior			✓	✓
2.3. Self-monitoring of behavior	✓	✓		
2.4. Self-monitoring of outcomes of behavior	✓	✓		
2.7. Feedback on outcome(s) of behavior			✓	✓
3.3. Social support (emotional)			✓	✓
4.1. Instruction on how to perform behavior			✓	✓
5.1. Information about health consequences	✓		✓	✓
5.4. Monitoring of emotional consequences		✓	✓	✓
5.5. Anticipated regret		✓	✓	✓
6.3. Information about others' approval		✓		✓
7.1. Prompts/cues			✓	✓
8.2. Behavior substitution		✓		
8.3. Habit formation		✓		
8.4. Habit reversal		✓		
8.7. Graded tasks		✓		
9.1. Credible source	✓		✓	✓
9.2. Pros and cons			✓	✓
9.3. Comparative imagining of future outcomes	✓			✓
13.2, Framing/reframing		✓	✓	✓
13.3. Incompatible beliefs			✓	✓
15.1. Verbal persuasion about capability			✓	✓
15.3. Focus on past success		✓	✓	✓

Though one thing is worth mentioning for Generation III. The swarm's implementation as a complement to the rule-based agent had to address the inspectability that was lessened even though the single agent remained in the app as a micro-dialogue coach. Every transfer in the swarm as well as an utterance within the same agent labeled a notice to the person before or during an utterance of which agent was taking part in the dialogue. As LLMs are black boxes, the person could not see exactly why a particular sentence was produced even if everything the agents uttered were grounded in the person's activities, preferences, and the national guidelines, but they could see which role was now speaking and when responsibility was shifted. Though not as traceable as an argument chain, it is considered a partial solution for the same requirement.

4.4 Generation IV: Aime

Aime (Figure 4.8) is a mobile app designed to support a person to find and evaluate recovery activities, which can help managing stress and every-day life balance if integrating them in daily life. As such it provides a more focused support towards one of the five lifestyle domains of StarCoach and Blumeai.

Aime continues the line of work of StarCoach and Blumeai by taking up the recovery branch or flower petal that the two previous generations had covered alongside physical activity, nutrition, and alcohol. Aime also embeds a GenAI agent which the person can choose from a predefined set of agents [136] or customize [134]. Recovery activities help a person replenish the mental and physical resources that stress consumes [241, 242]. Finding and keeping up such activities is not always easy when stressed, since stress can drain some of the resources a person needs to look for them [240], which is also a reason why Aime places emphasis on helping the person find, try out, and evaluate activities that work for them.

The development of Aime was also conducted through a participatory design process engaging experts in cognitive psychology, stress rehabilitation, occupational therapy, and AI. Particular outcomes from this was a stronger focus on the qualities of activity engagement and on capturing the person's experiences of each activity and mood, through ecological momentary assessment (EMA).

Some design decision from Chapter 2 were implemented in Aime. The first design decision was the idea of meaningful activity being the focus of the whole app, following Activity Theory [150]. Activities are created by the person, not prescribed by an external expert, which supports the mastery and control experiences in itself. Each activity is described by a set of characteristics developed as a part of the participatory design process based on dimensions of activity proposed by Alameer, Uitdewilligen, and Hülsheger [3] and Kaelin et al. [123]. The activity and its motive carry over from the StarCoach activity module. An activity bank containing activities suggested by participants in earlier studies [7, 136] gives ideas for new activities and serves as the agent's fallback when the person has logged too little for the agent to work with. In

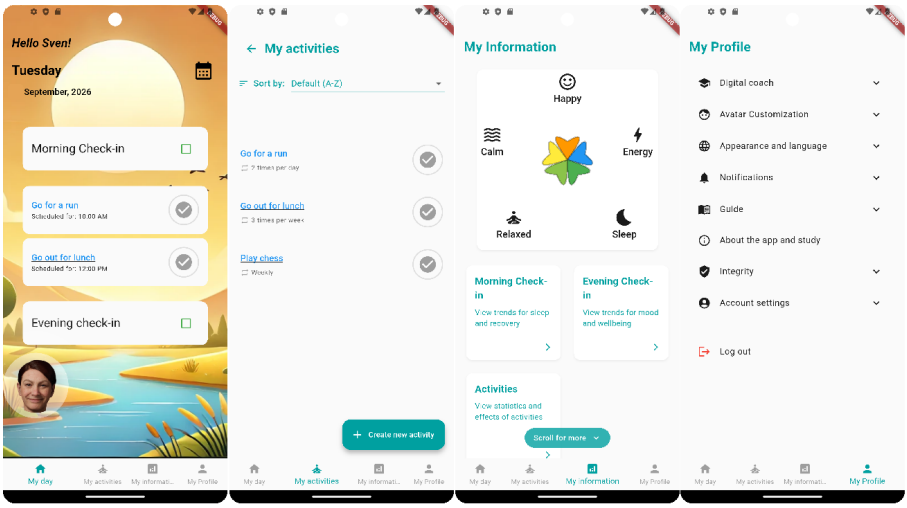


Figure 4.8: Aime's pages.

terms of Paper III, the exogenous space is still there in the forms of expert-authored content embedded in the app, as well as the tasks to assess activities imposed on the persons, but it is a resource that the person can use without having to only comply with it [127].

Another design decision taken in the participatory design process concerned how to capture information about experiences of a recovery activity and emotional wellbeing as part of the interaction. Logging an activity opens an event-based ecological momentary assessment [233] with five activity-effect items (relaxing, stress-releasing, fun, energizing, and sense of presence; 1–7), four mood items (depressed-happy, anxious-calm, exhausted-energetic, stressed-relaxed; -3 to +3), and whether the person came from work of free-time. Morning and evening check-ins asked about sleep quality, restedness, and mood, which the domain experts specified as valuable to capture recovery patterns over time. The *My information* page visualized emotional wellbeing as a flower whose five petals grow and shrink with the logs and check-ins, carrying over the flower goal image of StarCoach and Blumeai into Aime (Figure 4.8).

Moreover, a design decision on the agent was also applied. The agent uses the LLM GPT-5 mini and its context included the current season where the person lives, the past week's activity logs, each activity's type, company, and motive, and the morning and evening check-ins. The person selects among four roles taken from Paper I [136]: *Companion* (empathic and supportive), the *Alter ego* (direct, reminding the person of their intentions and challenging them to act), the *Expert* (knowledge-driven), and the *Assistant* (fact-based summaries of the person's activities). The Assistant is the default role. A difference from previous applications is the possibility to *customize* a role. The

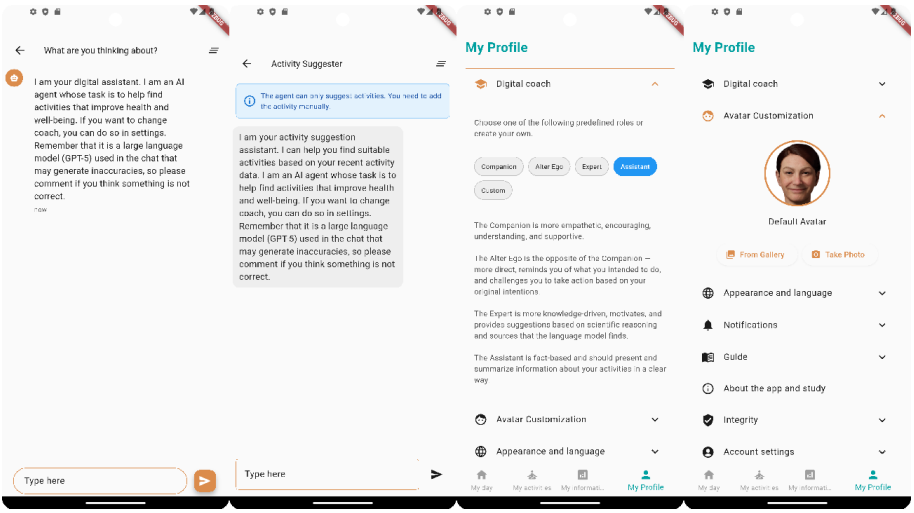


Figure 4.9: The agent chat, the Activity suggester, and the customization of the agent's role and avatar in Aime. Adapted from [134].

person can name the agent and author its instructions. The avatar can be replaced with any image, regardless of role. A separate Activity suggester agent is available in the activity module which reasons over the person's earlier logs responding to questions that a person may have about recovery activities. Lastly, a safety filter ends a conversation that is taking a direction towards mental illness or the consequences thereof, then telling that the topic is beyond the capability of the agent asking the person to start a new dialogue (Figure 4.9).

Compared with the earlier generations, there are a few things that changed in Aime such as the return to a single agent. The difference between Aime's single agent and the previous generations is the fact that Aime's agent's position and behavior is not tied to weights. The role's description is given in natural language, through text, written either by the design team with the experts or by the person themselves. The meta-level choice between supporting and challenging (Section 4.2.2) is still made by the person when they pick the role, but it is carried inside the role's instructions instead of being computed. The Custom role takes the appropriation, one of the design concepts of Paper III, further than previous generations did. In StarCoach, appropriation worked within boundaries that were set by the design team such as choosing the avatar, pick a predefined role, and decide when the agent was present by putting it to sleep. Furthermore, in StarCoach, the agent opens dialogues with the person three times a day. In Aime, every dialogue with the agent is initiated by the person. Additionally, the thing that is inspectable by the person in Aime is the role that is currently active and the data in the agent's context. The argument

chain of Generation I and II as well as the routing decisions of Generation III do not appear in Generation IV. Paper VI describes that this lack of a traceable chain also risks a provenance problem that the earlier generations did not have to face, the agent could sometimes make errors, such as misreading the person's own ratings, discussed further in Chapter 6.

4.5 Human-AI Team

"Alone we can do so little; together we can do so much." This is a quote from Keller [130] about teaming. Multiple studies have shown that this holds true even for teams within AI [29, 139, 166, 274]. The four generations of the apps developed and studied in this thesis stand in a different position to teaming. Activity Theory supplies the required vocabulary and method to explain the differences. A team is described not only by the members forming the team but also by its rules, community, and its division of labor [72]. These activity system nodes are used in [135] to pinpoint a participant's disengagement to a breakdown in renegotiation.

When looking at these nodes from Generation I and II's perspective, these two generations are dyads, in that the role of the agent is clear since the person selects it and the avatar is embodied in that role visibly. Basically, in Generation I and II, the person provides the data building the user model while the agent supplies the guideline knowledge, both dependent on each other to shape the dialogue on activities relevant to the person. Rules are explicit, given the dialogue protocol and rules. But the division of labor node is much more concise because there is only two actors: the person and the agent that is assigned a particular role by the person.

Generation III expands the division of labor node and gives it a major role in its architecture. Responsibility for a dialogue turn is allocated to the agent whose role fits the signal, hand offs shift the responsibility, and the reallocation of responsibility is apparent. The (re)allocation also places the person into the "team" since the person's input enters into the allocation. The preference weights are arguments to the transition function which also decide which agent takes the turn if there are more than a single candidate. The person, therefore, is a member shaping how the team distributes the work as a participant in the team's division of labor. Adding on to that, the visibility of the responsibility shift to the person also makes them aware. Shared awareness of who is doing what in a team is one of the conditions given in various teaming literature [119, 140, 196] and Blumeai's architecture enables this to occur.

The person's membership in the team is assumed in the evaluation of Paper V. The Teamwork Assessment Scale (TAS) [133] used in [137], an instrument developed for assessing human-human teaming, was used to evaluate what the person thought about the swarm as a team. This instrument was opted for as to our knowledge, no validated human-AI teaming instrument was available at the time of the study. Many participants in the study evaluated themselves as

part of the team, in fact leading it. By the time of Paper VI, HAT instruments had arrived [9, 10]. Chapter 6 expands more on this while also weighing the pros and cons of using a human team instrument for human-AI teams.

We implemented into Blumeai methods to capture how much the person engaged with the team, how work is divided, how evenly the division is spread, and whether the allocation served the person at all, information obtained from the system logs. The four metrics applied in Paper V were designed specifically to be able to study this and are presented as measurement instruments in the methodology Chapter 3. Their applicability and what they resulted in is detailed more in Chapter 5.

Generation IV returned to the dyad but the division of labor was different from Generation II. The person created the activities, logged them, rated them, and opened every conversation. The agent was interpreting the logs, suggesting activities, and responding in the position it was given. Every exchange was initiated by the person, which placed the agent closer to the tool role in the activity system model of Activity Theory (Chapter 2). The teaming ratings in Paper VI agree with this statement where taking the initiative was the lowest rated cooperativity item for two of the four cases at two weeks, and for three by their last time point. Three of the four cases also rated the agent's contribution to the team above their own. The fourth reversed this pattern due to having to repair a conversation where the agent was not able to follow the thread, hence, according to this case, she did the teaming work for the agent [134].

Chapter 5

Research Contributions

In this chapter, first the short summaries of each paper included in this thesis and which research questions they address is given. The summaries are followed by the detailed contributions each of the papers. Figure 5.4 shows an overview of how the included papers build on one another, and how the theoretical and methodological foundations feed into the design and implementation.

5.1 Summary of the contributions

This section presents the papers' contributions in the context of the thesis:

Paper I

- An empirical account of people's expectations of a digital companion: the roles they envision (assistant, coach, expert, companion, alter ego) and their preferences for supportive versus challenging, and empathic versus factual, behavior.
- Two novel argumentation schemes (Argument from Position to Support and Argument from Position to Create Tension) for tailoring supportive and challenging dialogue, together with a scheme-based mechanism for switching the agent's role.
- A two-dimensional framework mapping agent behavior along emotional support and emotional challenge, linked to roles and to the person's stage of change based on the Transtheoretical Model of Behavior Change.

Paper II

- *StarCoach*, a person-tailored, multi-domain health-promotion application supporting both short and long-term goals, produced through participatory design with a multidisciplinary expert team.
- A framework that clusters behavior-change techniques (BCTs) by their purpose, organizing, engaging, learning about effects, and re-

flecting on contradictions, to study multicomponent lifestyle interventions.
- Qualitative onboarding findings on usage, adherence, and perceived effects including the observation that participants anchored the new routine in already habituated activities, a possible adherence strategy to test in future studies.

Paper III

- An activity-theoretical, design-oriented framework that models behavior change as an interplay of three activity systems; intervention (healthcare), development (the person), and design/empowerment (the technology) balancing external and personal agency.
- The distinction between exogenous (externally driven) and endogenous (person driven) behavior change as a design lens that moves beyond usability, engagement, motivation, and persuasion.
- Five concrete design concepts derived from the framework: interaction spaces, contradiction-driven development, holistic transformation, appropriation, and transitions in agency (from assisted to autonomous change).
- An application of the framework to real systems, including StarCoach, showing how it may guide decisions on system structure, interaction style, personalization, and ethical responsibility.

Paper IV

- A theoretical framework integrating Activity Theory and Argumentation Theory to structure purposeful human–agent dialogue for behavior change.
- Argument from Position to Support (APS) and Argument from Position to Create Tension (APCT) schemes, extended with critical questions and made operational for implementation and evaluation.
- A dialogue protocol for argument-based human–agent dialogue, and its implementation in a companion used in daily life to promote healthier habits.
- Qualitative findings on how participants experienced collaborating with the agent, its behavior, sociality, roles, and appearance, and preliminary evidence on the applicability of the schemes and protocol.

Paper V

- A swarm-based multi-agent architecture for person-tailored, multi-domain behavior change support, in which behavior-change techniques (BCTs) and their mechanisms of action are mapped onto distinct, role-specific agent behaviors.
- Four coordination metrics computed directly from system logs: communication frequency, division-of-labor efficiency, division-of-labor balance, and efficiency ratio, which capture how work was distributed among the agents and whether the distribution matched the purpose.
- Findings from a four week in-the-wild feasibility study with 18 par-

ticipants, showcasing how they experienced the swarm as a team and how they reacted to its proposals and coaching, and assessing applicability of the design and metrics as a basis for a larger controlled study.

Paper VI

- *Aime*, a mobile app designed to support a person finding and evaluating new recovery activities, through in-app ratings of activities and mood for personal reflection, developed through participatory design engaging domain experts.
- A set of measures of activity engagement and emotional wellbeing accessible to the person to elaborate upon themselves and in dialogue with a customizable agent in the context of their chosen recovery activities.
- The role that fit was different for each person, and people found it by switching
- Continued use meant periodic, conditional, or temporary use, not daily use.
- Design implications such as a "pause" mode that checks back after a certain period of time; context tags such as deadlines or sick leave; a one-tap button to add agent suggestions as activities; counter-questions from the agent before suggestions; offering periodic debriefs with a human.

5.2 Paper I

Kilic, K., Weck, S., Kampik, T. and Lindgren, H. (2023) Argument-based human–AI collaboration for supporting behavior change to improve health. *Front. Artif. Intell.* 6:1069455. doi: 10.3389/frai.2023.1069455

Paper I aimed to gather knowledge about how people viewed argument-based support through digital companions promoting healthy lifestyles. The goal was to investigate what kind of AI coaching roles participants preferred, how much challenge vs. empathy the participants tolerated, how preferences vary by readiness to change, and how participants perceive argumentative dialogue with AI. The results indicated that the participants differ vastly in whether they prefer a supportive companion vs. a neutral assistant vs. a factual expert vs. a "tough coach". Essentially, Paper I showed that the *style* of persuasion mattered as much as the content being delivered.

In addition, we presented a framework for digital companions using computational argumentation to support healthy behavior change. The paper aimed to study how digital companions can interact with users in more adaptive, persuasive, empathic, and personalized ways to improve behavior such as physical health and stress management. In doing so, we obtained information to build an argument-based framework for tailoring a digital agent's roles and behaviors. The framework connected coaching roles, stages of behavior change, and

dialogue strategies into one while mapping the agents' behavior along two dimensions: emotional support/empathy and emotional challenge/assertiveness.

Finally, we introduced a novel application of argumentation schemes for tailoring the digital agent's role and behavior and for switching between or merging roles.

These outcomes laid the building block on which we built upon to advance the research frontier in the following papers[1].

5.3 Paper II

Lindgren, H. and Kilic, K. (2025) Participatory design and evaluation of digital coaching for improving health—the star multicomponent lifestyle intervention. *Front. Digit. Health* 7:1600535. doi: 10.3389/fdgth.2025.1600535

Paper II documents the system used in Papers I, IV, and V, the process that produced it, and where the companion's knowledge comes from.

Paper II gives an account of the participatory design process that was carried out over three phases with a multidisciplinary team of twelve domain experts from medicine, nursing, psychology, ethnology, social work, epidemiology, nutrition, health economics, and computer scientists, together with participants in the target group for the Västerbotten Intervention Program (VIP). Working prototypes were used by the domain experts in scheduled meetings and further development was made through their comments and suggestions. The comments and suggestions rested on seven recurring themes: the knowledge model and its visualization, adaptation, goal setting, assessment and communication of accomplishment, roles and behaviors of the digital coach, ethics and values, and onboarding.

The first contribution of Paper II is *StarCoach* itself (Figures 4.3 and 4.4). StarCoach is a person-tailored intervention covering five lifestyle domains. It is built from three modules. The star profile gives a holistic, normative picture of the person's health status, with each edge representing a value direction and its fill and color set by criteria defined by the healthcare experts. The activity module lets the person define their own "baby-step" goals, attach motives to them, and log them daily on a checklist. The digital coach delivers reminders, feedback, and short argumentative dialogues, and the person shapes its role and its behavior. The knowledge behind the three modules is engineered in ACKTUS as an ontology with argument structures following the Argument Interchange Format [48], and are authored and reviewed by the experts and 65 specialized nurses. The coach is rule-based to make sure every utterance was traceable to expert approved content.

The next contribution is a framework for studying multicomponent interventions carrying multiple BCTs (Table 4.3). Thirty three BCTs were identified

[1]Since this paper was the first one to be authored, the other papers build upon the idea that is presented in Paper I .

across the three modules and then clustered by what they are for. The clustering supplied a unit of analysis more informative and finer than individual BCTs for the application.

The final contribution was the qualitative account of the onboarding phase which took place for four weeks and with eight participants (five residents and three VIP nurses). Results found that people anchored the new technology in their daily habitual activities. Further, it was the habituated activities that became the tool through which StarCoach became useful for the person. The star was seen as a goal image, and some participants updated their desires for their coach role during the study which showed that people's preferences in how a coach communicates and behaves change over time.

5.4 Paper III

Kaptelinin, V., Kilic, K. and Lindgren, H. (2026) Behavior change beyond intervention: an activity-theoretical perspective on human-centered design of personal health technology. *Front. Digit. Health* 8:1754399. doi: 10.3389/fdgth.2026.1754399

Paper III argues that the focus of current digital health systems are too heavily leaning towards the *intervention*, or externally driven behavior change administered by healthcare professionals, and insufficiently on the person's own agency. As a contribution to advance the field, we proposed an alternative framework that is grounded in Activity Theory, adjusting the scope of behavior change to a more collaborative and evolving process. The process includes healthcare professionals, technology designers, and most importantly, the individual themselves.

Framework

The framework introduced a relatively new conceptual model of behavior change as an interaction, or inter-play, between three different activity systems (Figure 5.1).

The *intervention* system represents healthcare professionals and clinical interventions that attempt to create conditions for the individual's behavior change. The *development* system represents the person's own ongoing transformation of daily life practices, motivations, and routines. The final system, *design/empowerment*, represents designers and technologies that mediate and support both intervention and self-development. Through this framework, we argue that sustainable behavior change emerges not from the intervention alone, but from the dynamic interplay between these systems. Furthermore, we introduce a conceptual distinction between **exogenous** behavior change which is driven externally by clinicians, interventions, or systems, and **endogenous** behavior change, driven internally by the individual's own goals and contradictions.

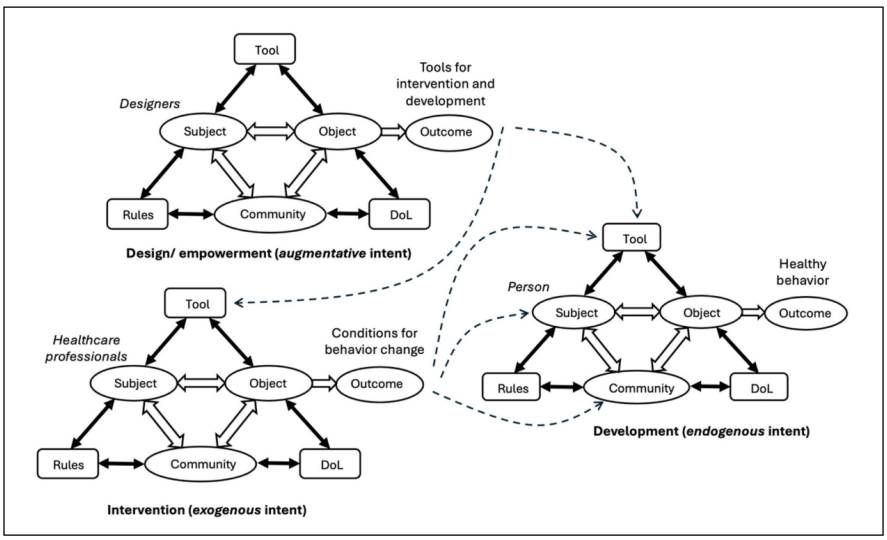


Figure 5.1: The proposed framework: behavior change as a network of three interacting activity systems (DoL, division of labor). Intervention system (exogenous) represents healthcare professionals creating conditions for change. Development (endogenous), represents the person's own transformation activities.

Design Concepts and AI Companions

Paper III suggests five design concepts as a guideline for future health technologies:

1. *Exogenous vs. Endogenous interaction spaces*: Systems should separate clinician-controlled spaces and user-controlled spaces.
2. *Contradictions as drivers of development*: Technologies should help users identify and work through conflicts in life (e.g., family obligations) instead of demanding compliance.
3. *Holistic and sustainable transformation*: Behavior change should be embedded in a person's broader life, not as isolated habits.
4. *Appropriation*: Users should be able to adapt technologies for their own purposes and meanings.
5. *Transitions in agency*: Systems should support movement between externally guided behavior and self-directed behavior over time.

Moreover, the paper outlines differences between virtual therapists and digital companions. Virtual therapists are AI systems that act as extensions of clinical authority whereas digital companions support self-reflection, autonomy, emotional processing, and the individual's ongoing development. Future AI systems (as well as current ones) should not only prescribe some action to individuals but should adapt its role dynamically, support reflection, negoti-

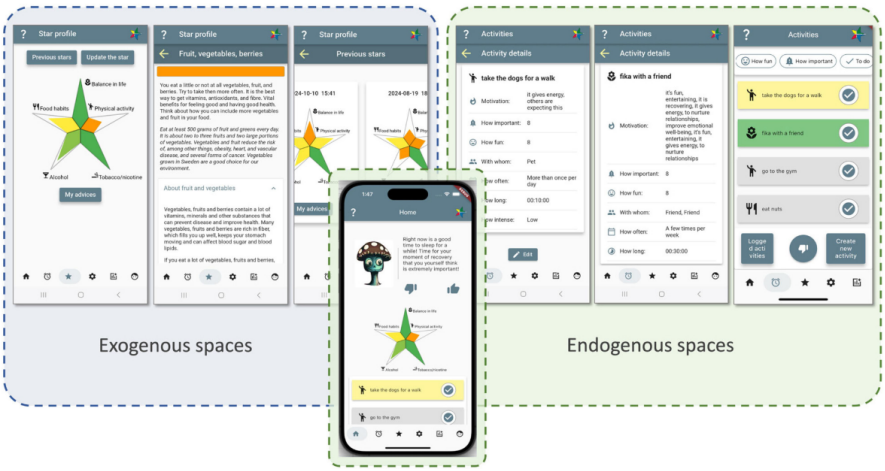


Figure 5.2: *StarCoach* as a hybrid behavior change system: exogenous and endogenous spaces in a single application. The exogenous space embeds evidence-based advice written by domain experts. The endogenous space supports user defined short-term goals and self-reflection.

ate contradictions, evolve with the user, and most importantly, preserve user agency.

The contributions outlined in Paper III are then exemplified and portrayed in a mobile application *StarCoach*. *StarCoach* combines the healthcare oriented behavior change components with the user owned reflective spaces, creating an exogenous and endogenous hybrid. An adaptive digital coach that can shift between the roles of expert, alter ego, and empathic companion is embedded (Figure 5.2).

5.5 Paper IV

Kilic, K. and Lindgren, H. (2026) To Challenge or Support: Managing Behavior in Dialogues for Promoting Health Using Argument Schemes with a Digital Companion for Human-AI Teaming. ***User Modeling and User-Adapted Interaction***

In Paper IV Activity Theory and Argumentation Theory are combined into one framework and is formalized, implemented, and explored.

The framework introduced in this paper calls the exchange between the person and the agent a *dialogue activity* in the activity theoretical sense. It has a subject, an object, mediating instruments, a community, rules, and a division of labor. The object of the activity is the topic. It's outcome is an

Activity Theory Concept	Dialogue Component
Subject	Actor (Human/Agent)
Object	Activity Topic
Motive	Argument ground
Contradiction leading to barrier	Counter-argument
Goal-directed action	Dialogue purpose
Operation	Dialogue move
Rules	Dialogue protocol / rules

Table 5.1: Mapping between activity-theoretical concepts and dialogue structures proposed in Paper IV.

agreement about how to handle that topic, which can be an agreement to not do the activity at all. The agent's instruments are its model of the person, the domain knowledge, the argument schemes, the dialogue types, and the set of dialogue moves it can take. The division of labor is made explicit through having the person set the goal, define activities with their motives and barriers, logging those activities, and assessing their own wellbeing. The agent, on the other hand, reminds, inquires, deliberates, persuades, supports, and challenges.

Through modeling in this way, the concepts introduced by Activity Theory and Argumentation Theory can line up one to one (Table 5.1), and the activity theoretical hierarchy of activity, action, and operation then is mapped onto motive, dialogue purpose, and dialogue move. Moreover, the person's motives become grounds for the agent's arguments and the barriers arising from contradictions in daily life become the counter-arguments the agent has to respond to. The agent can use this to distinguish a long-term motive from an immediate goal that is in conflict with it, and argue about the conflict.

Furthermore, the two argumentation schemes from Paper I are extended with critical questions (CQs) which decide if the scheme applies in a given moment. *Argument from Position to Support* (APS)'s CQs ask whether the agent is in a position to support, whether the person wants support from it, whether the agents wants to give it, and whether it know how the person wants to be supported. *Argument from Position to Create Tension* (APCT)'s CQs asks similar questions but pertaining to challenge. Paper IV argues that these schemes operate at the meta-level as *meta-level argumentation schemes* at the agent level, meaning they fix the position from which the following arguments are delivered. They are "socio-centric" as opposed to "ego-centric" when compared to traditional argumentation schemes. An argumentation scheme, Argument from Expert Opinion, as an example, has one answer regardless of who is asking the question whereas APS and APCT ask what two parties want from each other and requires more than one to have an answer.

It is around these schemes that Paper IV formalizes the building blocks of the dialogue activity with: the actors, an activity as a tuple of topic, value directions, motives, barriers, importance, and fun, a dialogue protocol of sit-

uations, topics, purposes, moves, and reactions, dialogue statements recording who said what in which context with which position and role, and seven rules governing topic selection, positioning, purpose, move, expressive behavior, and adaptation to agreement. The person's choice of role answers the first critical question of each scheme with the chosen role's weights, combined with the person's own empathy and challenge preferences, a computation of support-challenge level selecting between APS and APCT, and the utterance is made from a template by slot-filling from the person's model and the expert content from the knowledge base. As Paper I's results showed a close split between support and challenge preferences from the participants, a qualitative pilot with nine domain experts assessing the prototype agents was conducted.

The empirical study spanned four weeks engaging eight participants, five residents and three VIP nurses. Collected data was analyzed thematically across agent's appearance, its sociality, its behavior, and the expectations of the people using it. There were some findings that were key: For one, it was established that support and challenge desire from a person is not a static, stable property of that person. The nurses preferred an empathic companion, consistent with how they conduct Motivational Interviewing sessions. The other non-medical participants wanted some "push" and this desire shifted from time to time. The CQs of APS and APCT therefore have to be answered per person and on a rolling basis. Another finding was that *support* can represent at least two different things. For example, one participant wanted affirmation to a decision they had already taken and did not like being reminded of what to do. Another participant experienced value-based statements as mirroring what they had themselves valued. Though these are very different, they are both represented by a number on a single scale, meaning that one number is too vague. Finally, the presence of an avatar shifted the app from being an instrument to being a community member in activity theoretical terms. For instance, one participant named the avatar, and another described having "ghosted" the agent in the same way a person "ghosts" a person rather than some tool. Another participant compared the app to another app they were using which lacked an agent in it, and described themselves as being their own coach in the other app.

The penultimate case described in the previous paragraph is a scenario where the formalization is important. The participant had set the Alter Ego agent to maximum challenge and found that, at the first follow-up, that the challenge being delivered was not enough. The same participant had disengaged by the second follow-up. Since the dialogue activity is formalized, this is not recorded as the participant simply disliking the agent but it is localized to a failure point to renegotiate the agent's position after disagreement (Rule R7 of the Dialogue Rules), and to a breakdown at the Rules and Community nodes of the activity system. Being able to pinpoint the exact place where the building block has failed instead of saying the interaction failed is a more robust way of approaching a solution to a breakdown.

A visual of the argument-based dialogue activity mapped onto the activity

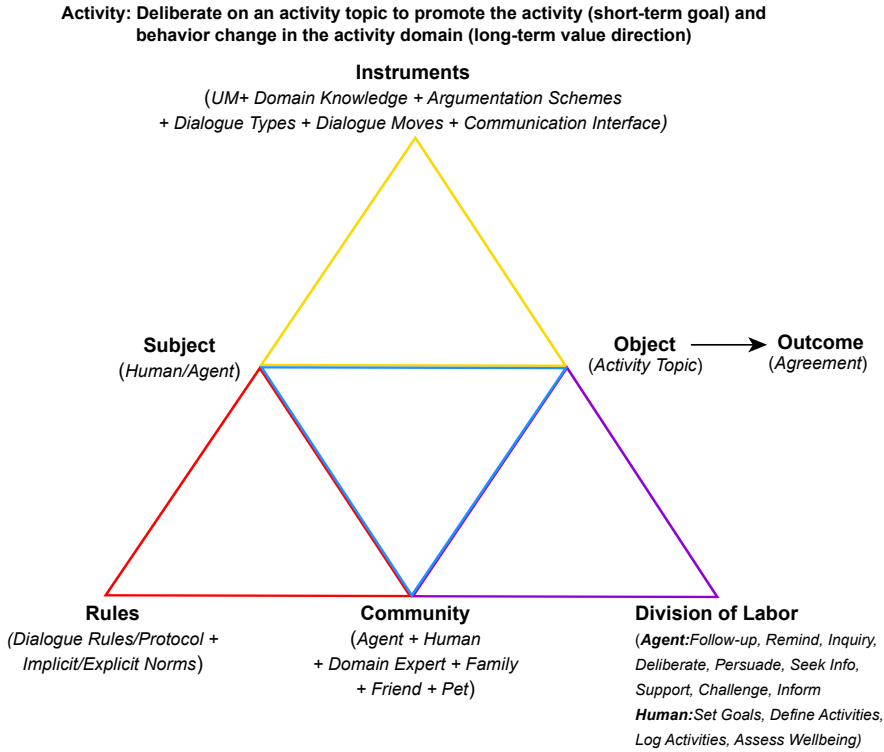


Figure 5.3: The argument-based dialogue activity of Paper IV modeled in Engeström's activity system. The object is the dialogue topic and the outcome is an agreements on how to handle it. The instruments are the user model, the domain knowledge, the argument schemes, the dialogue types and moves; the rules are the dialogue protocol together with the implicit and explicit norms; the division of labor separates that he person does from what the agent does.

system can be seen in Figure 5.3.

5.6 Paper V

Kilic, K., Kaelin, V.C., Mendez, J., and Lindgren, H. (2027) Human-Swarm Teaming for Personalized Behavior Change Support: An In-The-Wild Feasibility Study. ***Submitted***

In Paper V the focus of research expanded from the dyad between the person and *an* agent, towards the person and the *team*. Papers I and IV established that different people want different roles from a companion, and that the same person wants different things at different moments. Paper I and IV address these issues by having one agent shift its role. In Paper V, the roles were instead given to separate agents. The purpose was to explore how participants experienced the shift between roles when each role is carried by its own agent, and whether a shift that appears as a hand-off from one agent to another could be more transparent to the person than a single agent changing its behavior. This approach also is similar to how behavior change is often supported in practice, where an interdisciplinary team contributes different expertise rather than one professional "putting on different hats".

Human-AI Teaming (HAT) allows us to bring this into the digital realm. For the agents to count as members of a team and not tools, they need to have distinct roles, enough autonomy to act, and interdependence within the team it is a member of. The person who the behavior change is being imposed on also needs to be a member of the team to work together for the joint objective. Paper V's first contribution is the swarm architecture which realizes this. This architecture is the descendant of *StarCoach* documented in Paper II. It covers physical activity, recovery, nutrition, and alcohol. Four specialized agents taken from the preferences of the participants in Paper I are carried over into the swarm: the empathic companion, the challenging coach, the medical domain expert, and the neutral assistant. A triage agent orchestrates the routing between the agents initially and a RAG agent grounds the agents' utterances in the Swedish national guidelines. The architecture of the swarm can be seen in Figure 4.6.

Furthermore, the agents' roles are directly mapped onto BCTs and their MoAs which makes the agents specialized in their approach to behavior change. Instead of having a single agent carrying all of the BCTs, each agent is given a cluster of them. The challenging agent works with action planning, commitment, anticipated regret, and incompatible beliefs; the expert with credible source and information about health consequences; the companion with social support and verbal persuasion about capability; and the assistant with goal review and monitoring. A small set of techniques are shared throughout the whole team of agents.

In Paper V, we also introduce four coordination metrics that are computed from the system logs based on the dialogue between the person and the swarm. They are communication frequency, division of labor efficiency, division of labor balance, and efficiency ratio. Communication frequency tells about how much the person engaged with the swarm, and the three other metrics describe

how the agent divided the work amongst each other. Since all metrics are computed from what happened in the dialogue and not from what participants reported later, they present a glimpse into the team's coordination that is independent of the interviews and questionnaires, although the scores of efficiency and efficiency ratio are derived from the participants' reactions.

The swarm is tested in an in-the-wild feasibility study running for four weeks with eighteen participants. Many of the participants regarded the swarm as a team with role clarity, leadership, and explicitness of instruction rated well. The roles were also apparent to the participants. The study was formative, establishing a role differentiated team that can be measured, run, and experienced as a team during the course of the study. Paper V asks whether a team can hold a joint position coherently, and whether the person can see and take part in that team's activity as a member.

5.7 Paper VI

Kilic, K., Kaelin, V., and Lindgren, H. (2026) Exploring Recovery Activities in Collaboration with a Digital Companion: A Case Study of Aime. ***Submitted***

Paper VI moves the focus to supporting the person's recovery through human-agent collaboration. Conversational agents are increasingly used in technologies for stress and recovery, yet little is known about how people perceive and team up with such agents in daily use, or what it would take for them to keep using one. Paper VI addresses this with *Aime*, a mobile app for wellbeing through recovery activities.

The first contribution is Aime itself and the design decisions that came from a participatory design process with experts in cognitive psychology, stress rehabilitation, occupational therapy, and AI (Section 4.4). Following Activity Theoretical principles, Aime is organized around activities.

The next contribution is the eight propositions formed from the cases and checked across the cases [271]. The paper concluded that teaming was something the agent did for the person rather than with them. For two cases, the interview accounts and teaming items improved when they had chosen or created roles that fit them better, which suggests that some likability items miss what makes the agent fit the person. Each case reinterpreted the EMA questions in their own way and filled them out in different ways such as one case filling all of them out in nightly batches. In addition, the agent also erred by misreading some scales, and in one case told the person that it was concerned about the person based on the misread scales. In an agent-supported app, the scale is a part of the interface between the person and agent, and it so happens that both parties can misread it. When asked how they would fit the app into their lives over a longer period, none of the cases described continued use as daily, uninterrupted use. The difference between the cases were the conditions set out by each of them for continued use. Case 1 saw it as returning to the app periodically whereas Case 2 saw it more as a launchpad he would outgrow

as new habits started to take hold. Case 3 tied continued use to his own discipline, and Case 4 to a revised app and a clearer working-year routine. So, none of the cases saw the app as the end point. Paper VI therefore concludes that designing for breaks, periodic check-ins, and contexts such as deadlines, may matter more than designing for sustained, daily engagement.

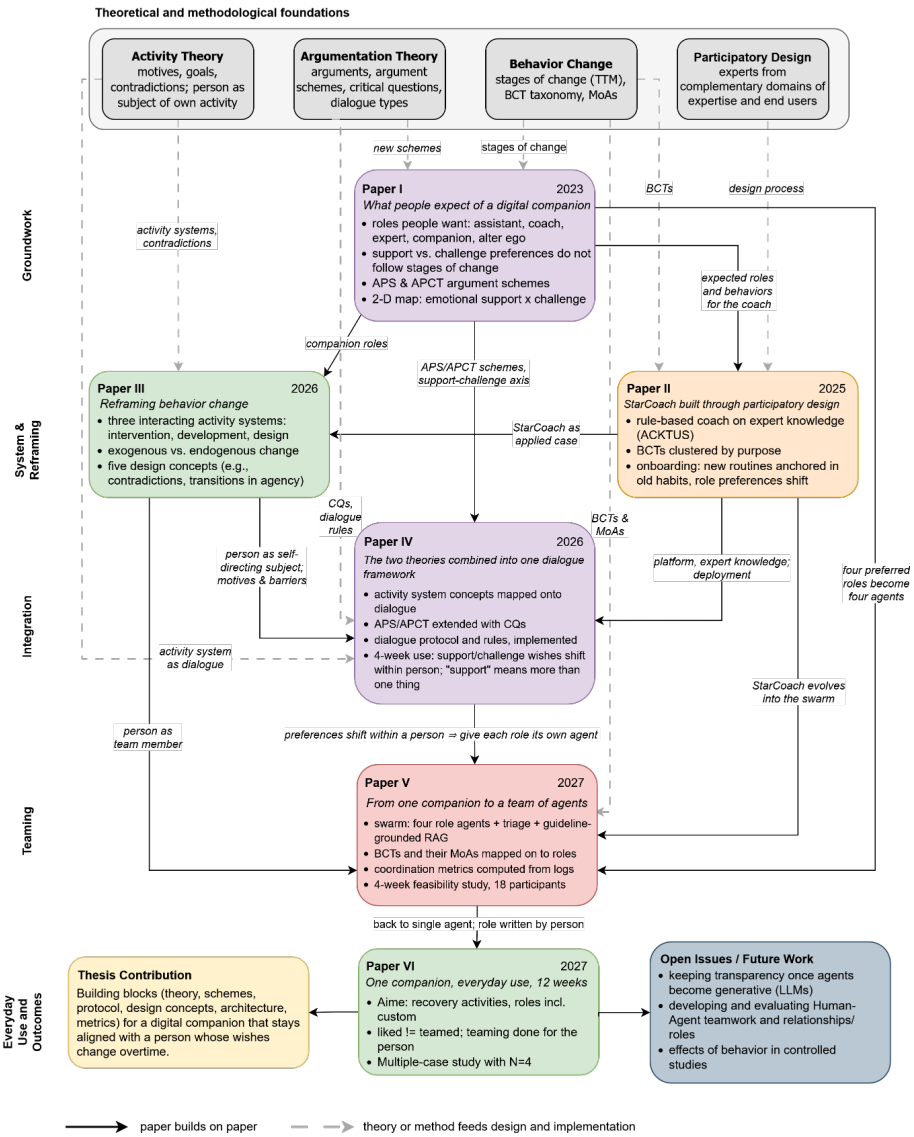


Figure 5.4: How the included papers build on each other. Dashed arrows show where a theory or methods feeds into the design and implementation of a paper. Solid arrows show where a paper builds on an earlier one.

Chapter 6

Discussion

This chapter takes the findings discussed in earlier chapters of this thesis as a whole and discusses the contributions beyond each individual paper. It discusses five challenges: aligning with a dynamic target, beyond a single scale of support/challenge,what changed when moving to GenAI, leading the team versus being a member, and habits and relations, that run through the thesis, each with its own section in this chapter. The chapter concludes by tracing back to the research questions set out in Chapter 1.

6.1 Aligning with a dynamic target

There are patterns that appeared during the studies of this thesis. The participants in the study of Paper I most often opted for the assistant and coach roles. The participating experts who used the prototype, however, preferred the empathic companion. When participants of Paper IV used the app (Paper IV), the preferences they reported shifted over the course of the study and some asked for a coach that would push harder and hold them accountable (Paper II). In Paper V seven of the 18 participants selected only the Assistant role while two who only chose the Challenging role. But the logs show that some of the participants who selected the Assistant role were pushing back on just informational replies. They called the suggested activities boring or replied that they would prefer to do the activity in their own way, basically ignoring the utterances. In Paper VI, the shifts were visible in the logs. One participant moved through roles four times in a month, from Expert to the Companion and the Alter ego and finally to a custom role they settled on. The role that fit was different for each person; a context-aware custom agent for one person, and a challenging Alter ego for another [134]. In comparison, a third person tried three different roles within a short time frame and found no agent that fit. It can be understood essentially that what the people say they want before usage, what role they choose when given a default, and what they actually want

or need are three different things. Aligning an agent with one would be misaligned with the other two. This exemplifies, from an activity-theoretical point of view, a natural part of developing new ways for the human to interact and use technology; consequently, technology should align with the human's adoption process. In activity-theoretical terms, the person initially gives the new tool or way of doing things conscious thought (what role to choose, changing, finding a way to make the tool work, etc.). Once this has been automatized, the relationship, collaboration, and interaction has taken shape.

Decision making research have posited for a long time that preferences are built at the moment they are asked and not from some stable, pre-existing inner core [157]. The literature on AI alignment is slowly coming to this conclusion as well. Carroll et al. [41] argue that approaching preferences as something fixed does not mean that that alignment method is neutral, in fact, they argue that it rewards systems that shift preferences to satisfy the easiest target. Furthermore, Franklin et al. [85] say that to be able to focus on alignment correctly, a new dedicated multidisciplinary science for preference change is needed. Zhi-Xuan et al. [275] go even further by asserting that preferences are the wrong target for alignment, and propose systems to be aligned with normative standards that fit their social role instead. The domain of behavior change is a tricky one since change, in terms of behavior change, is something that is a part of the goal in what the person wants and not a side effect. SDT describes successful change as the internalization of motivation [218], which in simplest terms, in our case, means that a companion that "works" will be able to support the person in the process of internalization. But this "just do what the person prefers" is then not a well defined instruction for a behavior change companion, just like finding a new song on Spotify based on only "what the person prefers" is not a good way to find a new song one enjoys.

So, this thesis does not have a single answer to this conundrum but several approaches towards a solution. Paper I explored whether the participants' stated preferences for empathy and challenge matched their TTM stage. The descriptive results of Paper I showed no consistent pattern; desired empathy was low to neutral at every stage, and the preference for challenge was almost split in half. However, there was a difference in preferences; there was a clear desire among participants that they were not interested in an empathic agent for stress and recovery activities, but more people wished to have a challenging coach when it comes to physical activity and even more when it comes to stress, which is an interesting find [136]. This means the participants, regardless of TTM, had somewhat consistent opinion of behavior. Moreover, it indicates that stage by itself could not be used to calibrate the agent, and we are aware that the transtheoretical model has been criticized before for its stage matching [266]. The design choice was taken to let the person decide about level of empathy and challenge, in order to also study how participants perceived such assessment. But the inconsistency finding indicates that handing it over to the person alone is not enough as a choice made in the beginning might change after a week. In a study by Hawkins et al. [102] on tailoring messages, the "done

once" type of tailoring relying on a single baseline, and updated tailoring which involves continuous reassessment are distinct, and, on a similar tangent, just-in-time interventions update continuously [186]. Either way, both adapt what is delivered and when based on sensing the person's state at the time. These two approaches, despite their usefulness and wide adaptations, do not address the relational stance of the agent nor ask about this since these approaches typically do not involve an agent.

The argumentation framework allows this problem to find a location in this complex web. The CQs of APS and APCT asking whether the person wants support or challenge, and in what form, do not have a static answer for all people or for all time. APS and APCT have be asked again and again, turning the support and challenge setting to a recurring move in the dialogue. What the person sets is an instruction on how they want to be treated when their momentary reaction pushes back in the future, which in Franklin et al. [85]'s terms is "meta-preference" and in Elster [69]'s terms a "precommitment" (Ulysses wanting to be tied to the mast before hearing the sirens is a rational past decision overriding the future irrational Ulysses). Looking at it from this way, Paper IV's P4 is not a person who has changed their mind; they made a standing request to be pushed hard, the agent pushed less than asked for, and since the system did not offer a move to raise the level of challenge, the choice made by P4 was to abandon ship or "ghost" the agent.

Psychotherapy calls this phenomenon an *alliance rupture*, which is a breakdown in the collaboration between a client and therapist [74]. A client moving away from the work is a withdrawal rupture and how well these ruptures are repaired is a predictor of how well the treatment goes [74]. The "alliance" itself rests on factors such as agreeing about goals, tasks, and forming an emotional bond [28]. The CQs of APS and APCT address the task part of the agreement, basically how the two parties will work together toward the goal. "Ghosting" the agent is a withdrawal rupture in Eubanks, Muran, and Safran [74]'s words, and rule R7 of the dialogue protocol introduced in Paper IV shows where the repair move belongs. The authority of the change is still vested in the person, nevertheless the prompt to revisit the transitions of agency (Paper III) can and should come from the agent. Such an approach also offers a solution to the influence issue raised by [41]. A companion that challenges will shift the person's desire for challenge, and the question of which shifts are acceptable cannot be decided in advance. The position that this thesis takes is that a shift is acceptable when it is made visible and agreed to in the dialogue itself, a question of contestability in and of itself [106].

Such a repair when it works, is shown in Paper VI. A participant in Paper VI initially rejected an activity the Alter ego proposed, told the agent that it did not fit, and then got a new proposal, as an example. Two months later the same person described the exchange as a "co-creation" and that the agent was "appropriately challenging" [134]. The challenge by the agent was disputed and the agent adjusted, and ultimately, the relationship survived. The difference from P4 is the system, where the generative agent could take a

counter-argument it had no template for and respond to it, though nothing in Aime guaranteed such a move. It was due to the language model's flexibility and did not come from a rule, which means the system could have, if it wanted, continue to push. The repair move that rule R7 places in the protocol should be designed in and not left to up to the language model's possible response.

6.2 Beyond a single scale of support and challenge

Two weights, ω_s and ω_c, represent the agent's stance in StarCoach and combine into a single support-challenge level (Chapter 4 and Paper IV). The single level was implemented due to the agent having to choose between APS and APCT, with the person setting the levels directly, and also allowed for the domain experts to be able to inspect it, which were requirements from the participatory design process (Paper IV). But this came with a cost of having to merge the values of Paper I's framework which had two separate dimensions for support and challenge. Paper IV established that support means different things to people and Paper V found participants preferring persuasive replies while resisting informational ones. This thesis proposes that a support or challenge move varies along the axes of at least three interdependent dimensions. The first axis is *Direction*, whether the move affirms (support) or creates tension (challenge) comes from behavior change theory and evidence based studies on how to support an individual [177, 219]. The second axis is *Grounds*, whose reasons are carrying the move, the person's own motives or an external authority (national guidelines or expert), supplied by argumentation theory [260]. The final axis is *level*, which layer of the person's activity the move addresses, whether the motive behind it (why), a goal-directed action (what), or a routine operation (how and when) [150], is provided by activity theory [150]. A single scalar collapses all three dimensions (Figure 6.1).

The vocabulary introduced in Paper III, for instance, falls into the grounds axis. The endogenous and exogenous distinction as well as the support-challenge distinction ran, until this chapter, without being truly connected. In this chapter, it is argued that they are two different axes which are not interchangably used. As a supporting example to this claim, a challenge can be endogenous. When StarCoach suggests a walk to the person, telling them it is time for their walk and that it would give them more energy, it is creating tension using only the person's own motive, which is also called developing discrepancy in MI [182]. Also, a challenge can be exogenous like when the agent cites the national guidelines for limits on alcohol consumption. Support is split in the same way into affirming a choice the person made and reassuring them with what experts recommend. SDT would predict that the grounds matter just as

[2]In Generation II, the person's preference was logged into the StarCoach as a single support-challenge level (Equation 1, Paper IV). The equation in Paper IV chooses the direction of a move only, hence it cannot show a preference about the grounds or level.

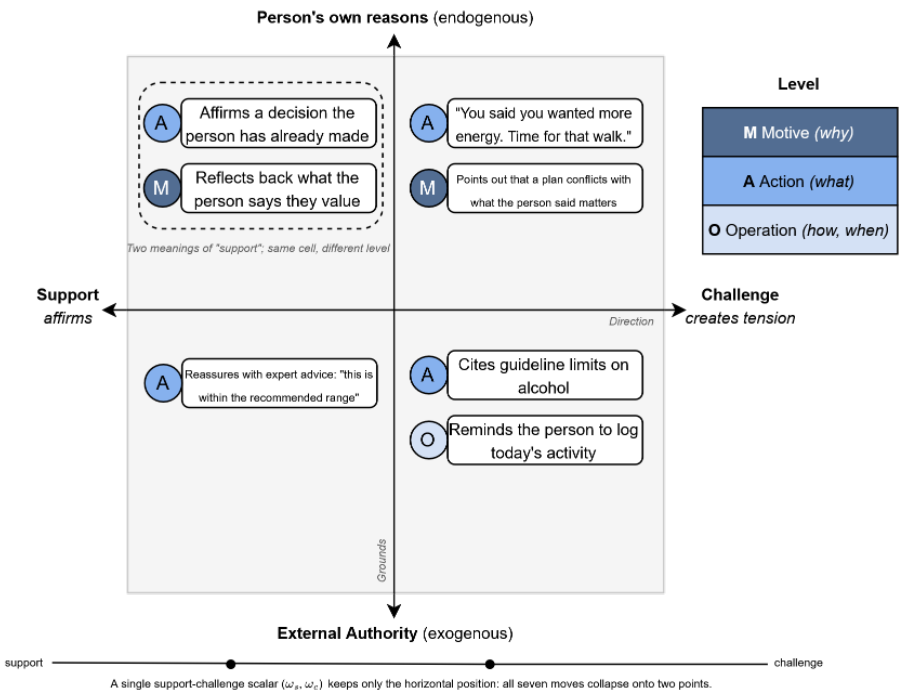


Figure 6.1: Three dimensions which a supportive and challenging move varies: Direction (horizontal), Grounds (vertical), and the level of the activity it addresses (badge). The examples given are from the four generations of systems. A single support-challenge scalar keeps only the horizontal position, so moves that differ in grounds and level become indistinguishable[2].

much as the direction because the exact same pressure can feel supportive or controlling dependent on whether it connects to the person's own values [254]. When looking at the pushback to the Assistant and Challenging agent whose utterances were mostly guideline based in Paper V, from this point of view, it can be stated that the participants resisted exogenous grounds delivered at the level specific food and exercise choices.

The level axis in Figure 6.1 shows two meanings of support in Paper IV that the scalar could not represent. A participant wanted the agent to affirm decisions they had already made and did not like being reminded of what to do whereas the other participant valued replies that acted as a mirror to what they already cared about. So, both are supportive and are both grounded in the person, but the first operates at the level of actions and the second at the level of motives. The reminders the first participant did not want operate at a lower level, the operations level, as this participant considered it under their own control. A similar observation occurs in Paper I where participants read a neutral agent as empathic when the scenario involved a person in distress. Direction is partly assigned by the receiver depending on the situation which is why a number set beforehand cannot correctly capture the expectation in a particular situation.

Generation IV replaced the scalar by allowing role descriptions to be given in natural language, which the person could also do. The custom role, one participant wrote in Paper VI, was a psychotherapist who carried the participant's background. The agent's turns always referred to the participant's context and offer their native language, and the participant valued how the agent adjusted to her communication style and current situation [134]. A written role can hold all three axes at once; how much to push, on whose grounds, and at which level of the activity, something the scalar could not do itself. The cost of this is that the position and behavior is no longer something the system can check or compute. Since there is no ω to compare against what the agent actually did, whether the agent kept to its role could only be judged through going over the conversation afterwards.

A little detour from the thread but a worthwhile one, as a criticism for the current literature on how sycophancy is discussed, it tries to and tends to measure sycophancy as something in the direction axis even if the model and the user agree that it should not [205, 230]. This three axis approach grounded in theory shows that an agent can still remain sycophantic even when disagreeing, for instance, through challenging the person on some random operations while not once touching the motive they are avoiding, and can try to resist sycophancy while agreeing through affirming a motive the person has based on external pressure. Alignment in these digital companions is decided on all three axes, grounds and level included. The Blumeai swarm separates its agents by direction (Companion and Challenging Coach) and partly by grounds (the Medical Domain Expert speaks partly for external authority), but level is not as clearly represented in the three generations of the apps and future work on this dimension in the architecture could prove highly useful.

6.3 What changed when moving to genAI?

Reduced traceability in Generation III is discussed as a limitation in Chapter 7. Here we outline what exactly was reduced and what remained because just the word "transparency" is ambiguous. To define transparency, it is made up of three properties: *Provenance*, knowing where the content comes from; *Inspectability*, being able to see why a move was made; and *Contestability*, the person's ability to dispute a move and the system's ability to adjust in response [4, 106]. In addition, previous research on recommender systems agree but call contestability scrutability instead while stating that transparency and scrutability are different goals as well [250]. According to Leofante et al. [149], it is Contestability that is the most important and requires systems to be able to exchange reasoning. Computational argumentation provides an opportunity to be able to do this [149].

These three properties were apparent, at least technically, in the first two generations. The content delivered was through expert authored knowledge and each utterance was traceable through its schemes and CQs. Counter-arguments were admissible moves in the protocol. In practice, however, the contestability was not a large feature. The support and challenge parameters were set and stayed fixed until the person changed them through the settings screen. The transparency and rule-based utterances were in fact transparent as can be to the researchers and domain experts who could audit each statement but this auditing was not a feature readily available to the person themselves unless, i.e., they did not know about the scheme, as an example, nor did the agent proclaim it during the conversation. Now, this is a big plus for systems deployed in health contexts but it should also be noted that the person who was using these systems might have had even more of a say if they had access to this information during conversations, though the benefits are conjecture at this point.

Generation III's purpose was not to address this issue. Therefore, it did not resolve this issue. Provenance was not as clear as it was in Generations I and II even though retrieval from the national guidelines were used as grounding for what the agents said [154]. Also, as Bender et al. [18] states, it cannot be fully guaranteed that each sentence that is generated through an LLM will stay faithful to what was retrieved. Due to the nature of LLMs, inspectability was also lost at the single sentence level since it is currently not possible to ask an LLM to "explain itself" as the explanations of an LLM might not be the reason the output was produced [253]. The change that majorly occurred was the movement of decision about stance from the agent to the structure of the team. Handing off to the Challenging Coach from the Empathic Companion is, in a way, the same decision as instantiating APCT instead of APS. The meta-level schemes remained in the generation as routing, and the transfer notice showed the person which position the agent was taking at that moment. In Generation II, this reasoning happened in the reasoning engine which was not visible to the person. So it can be said that in Generation III, traceability was

reduced in favor of visibility which could arguably be more important to the person using the system.

Two things did not carry over to the genAI version (Generation III) though they were taken care of in a different way. The first way was renegotiation. Routing took the classifier's reading of the message with the person's preference weights but the swarm and the person did not revisit those weights once the conversation took off. The swarm solved the issue it was set out to do which was to have a single agent transfer to a more helpful agent between incompatible roles, yet it did not solve the breakdown situation outlined in Paper IV; more like it relocated it. The second was the drift to no position. Paper V highlights that the neutral assistant attracted workload even though specialized agent did better on the same turns. Looking through the argumentative lens, this means that the Assistant is the agent that takes no position. Still, a system falling back to it by default is not a neutral system per se; it is answering the question to support or challenge with "neither" which for a person who wants some push is still a form of misalignment. In Generation IV, the Assistant was again the default role, and none of the cases in Paper VI ended up using it most [134].

Lastly, using genAI altered what participatory design can do, by definition. In Generation II, the multidisciplinary expert team reviewed the knowledge base and also co-authored what the agent would say. When it comes to Generation III, this was not possible as no one can co-author utterances that do not exist until run time. The things that stakeholders could decide now were the roles, the boundaries of those roles, the corpus the agents could retrieve from, and the rules that decide who speaks when. Therefore, participation moved from authoring to governing behavior, which is a crossroads of participatory design and responsible AI in itself [23, 159]. Generation IV took this idea a bit further. With the custom role, governing the agent was no longer reserved to only the design team since the person could also create their own agent (within boundaries). Participation thus reaches from the expert team to the person themselves, appropriation in Paper III's sense. The expert team still decided the boundaries that the person could not overwrite (the activity bank and safety filter). Through the app generations, this thesis points to a practical solution for combining the two approaches, one argumentation-based dialogue research is moving toward: the choice of position and its renegotiation belong in an explicit, symbolic layer that can be inspected and also contested, while the wording can be provided by a genAI model [43].

6.4 Team vs. Leader

Participants of Paper V were asked to place themselves into the team when they rated the swarm team with the Teamwork Assessment Scale (TAS) [133], and several of the participants placed themselves in the lead (Chapter 4). Such a result is a nice fit into the activity theoretical account in which the person holds the motive and the agents hold the commitments derived from said motive. It

also fits Paper III's description of the person being the architect of their own change. From the point of autonomy, this is a great result.

However, it does create some peculiar tensions that were not expected initially. The team in our use case also then has the task of challenging its leader. Edmondson [67] and Morrison [185]'s works on human-human teams have shown that speaking to someone with more power is a risky move and depends on how open that person is to the talk, and also whether the team feels it safe to disagree. Approaching this from a human-agent team perspective, the risk is on the challenger in lieu of the agent delivering it as the leader can simply route around the challenger if they so choose, since the human leader is in control. In Blumeai, the person's preference weights can limit the Challenging Coach's share of the conversation, which could lead to the person not addressing it at all. The swarm architecture gives the leader the power to both grant the challenging agent a voice or to stifle it, and without an agreement of the kind discussed in Section 6.1. Simply put, a team being led by the person will most likely drift toward agreement if this is what the person seeks, or toward challenge if this is what they want. The first CQ of APCT, whether the agent is in a position to create tension, is a matter of voice in this case. The answer must come from what the person allowed in advance and not from a momentary approval, and at the same time the person keeps the right to take away the grant.

As mentioned in Chapter 4, we will also assess the applicability of the Teamwork Assessment Scale (TAS) [133]. The TAS had some matching strengths in measuring teamwork for our use case. It has a structure that addresses leadership, roles, cooperation, and it was adjustable. It also allowed the person to rate themselves as a part of the team. Finally, it also allows for comparison to human teams. However, there are also weaknesses that must be discussed.

This scale was developed and validated as an observer instrument with physicians rating a team of medical students in simulated ward rounds. In the validation of the instrument, the ratings that were given by students were not as reliable as those given by the physicians [133]. Paper V used the TAS as a self-reporting tool from a *member of the team*, where validity of the instrument is not as concrete. The items of the TAS also assume that other team members can observe each other, which in our case is somewhat true as the swarm agent could observe one another. But there is also the fact that the person was speaking to one agent at a time despite there being multiple agents, so the person was also rating agents they did not converse with implicitly, which, in a way, is also how some teamwork is conducted. So, some of the items in the TAS that were rated by the participant could have been more a rating of how visible the coordination was rather than how good it was.

The log-based metrics in Paper V answer a completely different question than the TAS. They were used to show the division of labor between the agents and whether it was done in a balanced and effective way, which is a type of coordination the person could not observe unless they kept track of the transfers. From this, it could be elicited that coordination the person could

not see might not count as teamwork from the person's angle, a principle also known as "observability" in coactive design [119]. This would agree, then, with Paper V's observation that the agent answering matters more to the person that the coordination happening behind the scenes. These limitations were faced due to scales of human-AI teaming just very recently appearing [9], and even then it measures a person working with one agent. A team of several agents led by the person still has no suitable instrument at the time of writing this thesis. The studies in this thesis suggest that such an instrument needs items on the person's own role in the team, on the observables, and on whether members of the team were allowed to disagree.

6.5 Habits and relations

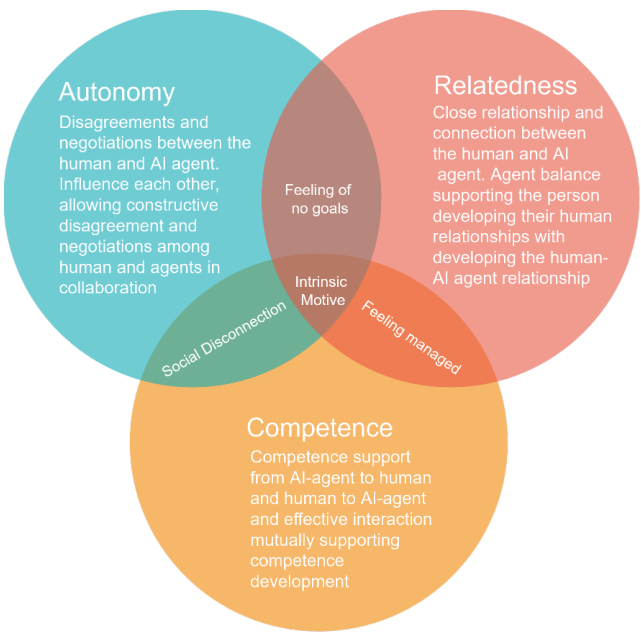


Figure 6.2: Three basic needs of intrinsic motivation according to Ryan and Deci [218]. The intersections between two of three needs lead to diminished motivation and well-being [218]; all three needs are required for intrinsic motivation.

The onboarding phase of Paper II and IV have a feature that should be discussed as well. The app featured in these papers centers around the activities a person has defined and chosen to do to create new habits, and so the companion does not enter empty "space". If it does, it suggests to create an activity so that it can function. In Paper II, participants made StarCoach a

part of their daily routine by adding habituated activities to it. The star profile was perceived as very important; it represented an existent desire of wanting a "better" version of themselves in a visible shape. In activity-theoretical terms, a tool becomes part of an activity as the actions of using it, which first require attention, are automatized into operations. In this process, operations that are already running are attached [129]. Studies on forming habits make a similar point from a different angle by stating that new behavior takes hold when it is tied to cues and habits that already exist in daily life [91, 148].

Moreover, the addition of the avatar as a persona changes what the agent becomes. People respond to computers that present social cues [188, 213], relational behavior keeps people engaged [21], and people form companion-like relationships with chatbots [236]. The results in Paper IV agree with these findings where participants, for example, named their avatar. In this thesis, we also see another side of this relationship. When the agent becomes a member of the person's community node, leaving that node take a social form, and ghosting is how people end relationships they do not want to argue about. The addition of an avatar persona adds a potential experience of relatedness, which is one of the fundamental human needs specified by SDT [218], and it also brings relational exits with it. A companion simply cannot have only the joining of a community, because joining also implies leaving which the companion must be designed to align with.

The relationship between the person and agent that supports behavior in terms of SDT can be seen in Figure 6.2. It is not the agent, per se, who should deliver the relatedness nor the competence but more imperative is how the agent can promote relatedness and competence. On one hand, the agent acts as a social actor but, on the other hand, there is the contradiction of replacing human relationships which can lead to negative consequences. Competence is developed in the onboarding phase; the agent can end up as the person's "buddy", but it is a question of how the person developed competence in relation to the agent but also competence in tracking their own behavior change trajectory in the process that is interesting. The person is in control of the relationship (autonomy) but autonomy is also a control of managing behavior change. So, the question that is tackled in this thesis is ultimately what role the agent has in promoting autonomy, competence, and relatedness.

These taken together directs the onboarding. The terms of the relationship, also how the agent will challenge and how the person can alter that, should be negotiated in the beginning or the first weeks of what Paper IV calls the forming phase in line with team development theory [252]. The sleep function of StarCoach coming from design sessions and appropriation in Paper III, is an example of an exit that is not a rupture (Section 6.1). In other words, the person can remove the agent from the activity for a period without cutting off the relationship, keeping the door open for future contact.

6.6 Reflections on Methodology

In the constructive approach, which was used during these studies, a methodological choice is gauged by if the artifact served the purpose of teaching something reliable about the problem it is addressing. The four phases abide by this approach, and they also indicate where the next round of studies can be refined in how the qualitative studies are conducted and how the agent strategies can be built.

A key qualitative choice we took during the studies was the timing of the questions asked. Participants were interviewed at baseline and after certain time of use. However, the thesis is arguably more concerned with things that happen within between those scheduled interview sessions such as whether the person wants support or challenge, as an example. Participants changed their coaches multiple times during theis use of StarCoach but the interviews reviewed the changes in retrospect [158]. A question after each dialogue about whether the agent was too soft or too harsh, or just right would have sampled participants' preferences during use instead of trying to recall during the interviews [233]. Another thing that could have been done is to bring the dialogue history into the interviews and ask the participants to walk through the dialogue to see whether what they did was what they said. Doing this would have let a breakdown situation such as P4's disengagement in Paper IV [135] be traced by the researchers and the system [38]. Though, bringing dialogue histories to discuss the participants' exchanges with the system could affect how they communicate with the system in the future, and also lead to awkwardness during the interview [90]. Furthermore, the interviewers and transcript coders were the same people who had developed the systems, and the participants were aware that they would meet again soon for an interview, which, according to Paper II, could have had a positive effect on adherence [158]. Bringing in another interviewer to the interviews from outside the research team, to also code the transcripts on top of the researchers, and agree with the structured codes could have strengthened the findings without affecting the qualitative rigor.

Paper VI brought both suggestions partly into practice. Every logged activity was followed by a short ecological momentary assessment, though it asked about the activity and the person's mood rather than about the agent, and the interviews were compared with each person's logs, ratings, and conversations, with discrepancies treated as findings [134]. However, momentary assessment did not always output momentary data. One participant logged 82% of her activities late at night in batches of three or more, turning momentary ratings into end-of-day recollections; another participant left the sliders describing mood at default for the first two weeks; a third participant recalled a 0-10 scale which the app did not have; and a fourth also logged in batches. Comparing the logs and the interviews shows that people described *how* they used the app accurately but misremembered *what* they had rated. Momentary assessment therefore does not replace the comparison of accounts against their logs. On

the contrary, it needs it, as well as the timestamps of when a rating was given just as much as the rating itself [233]. The questionnaires were completed just before the interviews held by members of the design team, with a single member of the team coding the conversations and the interviews. Further, a participant decided to try out the custom role after a researcher suggested it.

On the agent side, how the support and challenge were set by the participants, is a choice that can be discussed. In StarCoach, the participants were given the choice to set the support and challenge with two scales that remained static until the participant updates the values. In Blumeai, this became weights over the agents. Yes, this gave the person the autonomy and choice to compose how they are coached, which agrees with the activity-theoretical view of the person as the subject of their own change, and also kept the participant choices traceable. But, there were other paths that could have been taken, for instance, the agent could have asked at intervals whether its position and behavior was still apt for the person's state, which would have made the critical question a recurring dialogue move (Chapter 6.1). Alternatively, the system could have learned from the person's reactions (thumbs-up, thumbs-down, excuses as feedback) to better adapt, like a contextual bandit [145, 248, 255]. Such an approach could have been faster but it comes with the consequence of reducing the person's autonomy and control, which is kept to a much larger extent through asking the person. A system that can propose change and, at the same time, gives the person the power to confirm or reject, combines best of both worlds. Additionally, the position of the agent could have been more comprehensive. As an example Paper IV found that "support" meant praise to some, but being left alone to others [135], which is something hard to express in a single support-challenge scale.

Generation IV took a different avenue. In Aime, the person picks one of four predefined roles or creates the agent's role and instructions themselves, so the behavior of the agent is set by the person's own words as opposed to a scale [134]. This gave the person most control of any generation, and the one participant who created her own role valued that it adjusted to her own communication style and context. However, the agent did not initiate a conversation itself, by design, and never asked if the role still fit after a certain period of time. The participant who moved to a better fitting role did so on their own accord. The recurring check-in on the agent's position and behavior, as suggested by one participant, was still missing.

With Generation III, the positioning was moved into the LLM's routing which meant that it could no longer be traced as an argument chain. By keeping the support and challenge schemes as a planning layer, and using the LLM to only phrase the result would have kept the traceability and included the fluency of the language model, similar to what Koumi and Kakas [144] aimed to do. The swarm in Generation III was also chosen with the premise that a single agent playing all roles risks an inconsistent identity. Letting the same participants use both designs and comparing their accounts could test the premise in the same qualitative approach. Generation IV returned to a single

agent with several selectable roles but focused on a different premise (recovery) and with different participants, so it does not settle this question.

6.7 Addressing the RQs

RQ1 asks how theory and empirical insight (including the participatory design process shaping the content and design) can inform strategies for aligning a companion with a person in lifestyle change, and how these can inform the agent's knowledge and reasoning. The structure of the approach is built through theory: Activity Theory gives the agent a model of motives, goals, and contradictions. Argumentation theory gives the agent positions, grounds, and critical questions it can ask and answer. The empirical work shows that the content of alignment cannot be fixed and static when the agent is designed because, the content comes from the person. It is dynamic, fluid, and variable along the axes of direction, grounds, and level (Section 6.2). The agent's knowledge therefore needs to hold the person's instructions about how they want to be treated aside from their momentary reactions, and its reasoning needs to treat the question of whether to support or challenge as one that is iteratively asked rather than asked once and done (Sections 6.1 and 6.2).

RQ2 asks how supporting and challenging behavior can be formalized and implemented so that dialogue stays aligned with motives, readiness, and shifting preferences. APS and APCT with their CQs, coupled with the dialogue protocol make failures traceable to a particular rule [135] and also places decision making in the hands of the party that should be making the decision in that situation. The single scalar of support-challenge is not enough and the renegotiation rule must be implemented as a move the person and agent can make. Formalizing it is not enough alone, interaction design and implementation must be based on the formalization. In a generative system, the choice of position can be carried by the architecture itself which keeps the agents' stance visible to person even if wording cannot be as traceable as a rule-based system (Section 6.3).

RQ3 asks how people experience collaborating with such a companion, and with a team of agents, in daily life, and what design implications follow. Most participants in our studies treat the companion as a social actor and, when it is a team, as a team that they belong to and often lead. Each person is different and want and expect different things from the agents and, naturally, they want different things over time, often more push than initially asked for. The design implication that follow are (Sections 6.1 - 6.5):

- to make renegotiation a part of dialogue;
- to represent stance along the three axes of direction, grounds, and level;
- to conform with the person's consent to be challenged without removing their ability to withdraw from being challenged;
- to make team coordination visible to the person;
- to design exits that do not break relationship;

- to design for periodic use rather than daily use, with pauses and check-backs; and
- to expand participatory design beyond co-authoring content and interaction design to governing roles, and boundaries once the agents become generative.

Chapter 7

Limitations and Ethical Concerns

In this chapter is outlined the limitation, ethical considerations, and future work of this thesis.

7.1 Limitations

As stated previously, behavior change to improve health is the application domain of this thesis. However, the studies presented in this thesis do not measure sustained behavior change as an effect of an intervention in a longitudinal study. The empirical studies that are presented are exploratory, formative, or feasibility oriented. The questionnaires and formative participatory design sessions with domain experts presented in Paper I, for instance, lay the requirements but do not evaluate an intervention [136]. Following that, the StarCoach implementation studied in Papers II and IV are focusing on the onboarding phase without a comparable control condition [135, 158]. Furthermore, the study of Blumeai in Paper V was designed as a feasibility study and also has no control condition while reporting on descriptive statistics only [137]. We defer a randomized control trial to future work as it requires a large sample or participants and a time frame beyond the time of the writing of this thesis. The contribution of this thesis is, therefore, in terms of theory, frameworks, formalizations, design implications, and feasibility study results.

The largest limitation of this thesis was participant sample composition. All the samples of the studies are small and self-selected through Prolific, snowball sampling, or convenience recruitment, biased towards recruitment of interested and engaged participants. As an example, the StarCoach deployment recruited no participants at low readiness for change, which is typical of volunteer studies since volunteers have already taken some action if they chose to take part in such a study, but this also means that this thesis does not explore how a person

who is not yet motivated would experience these systems. Overall, the people who took part in the studies are people who are already inclined to engage which can skew some results.

Moreover, in a similar tone to the first limitation, most studies included in this thesis cover the first four weeks of use, the onboarding phase, in which a person decides whether the technology has a place in their lives [135, 137, 158], with Paper VI being an exception with 12 weeks of use with some participants. This time is very important as this is where the positioning of the agent matters the most because no relationship between the agent and the person yet exists to act as a fallback to. What happens when that relationship is fully established in TTM stages such as maintenance, is not explored in this thesis. The claims about AI alignment in this thesis is more focused on that key initial time where relationships are made between the dynamic duo of human and AI.

It can also be argued that the most important empirical finding of this thesis is that a person's preference for support or challenge is dynamic; it differs between different people and shifts within the person over time (Paper IV) [135]. Yet, although this is a qualitative finding based on limited number of participants, it can be assumed that this is prevalent in a larger population and needs to be addressed.

7.2 Ethical Considerations

The empirical studies presented in Papers II, IV, and V were conducted under the ethical approval granted by the Swedish Ethical Review Authority (Dnr 2019-02924), with written informed consent obtained from each participant [135, 137, 158]. The study in Paper VI was reviewed and approved separately (Dnr 2025-06073-01). Following the ethical approval, all data is stored securely with access limited to strictly the research team.

While research studies are following ethical approval, there are ethical concerns relating to AI-based behavior change systems and persuasive technology aimed at nudging users to change behavior, envisioned to be used in daily life beyond controlled scientific studies. In Paper V's ethical considerations section, it outlines the risk of a system challenging a person through using behavior change techniques such as anticipated regret or the confronting of inconsistencies. The APCT scheme to create tension from Paper I, for example, and every following generation of the system embodies and delivers tension to a person in daily use [136]. The thing that changes across the studies is what stands behind the challenge when it is delivered not whether the system challenges a person. In generation I and II, every challenge statement is a template populated from content provided by and validated by domain experts. It should be noted, however, that the CQs delivered in Paper IV are aimed to specify the conditions for challenge, in particular, putting the person in charge of deciding and not the domain experts. In generation III, the same class of statement is generated by an LLM at inference time even if the utterance is

grounded in a RAG version of the Swedish national guidelines. It says that these can be experienced as coercive if mistimed, and the Blumeai study did not implement a mechanism to exclude the Challenging agent from the agent team, or adjust its intensity, just the option to not use the agent team [137]. To address this ethical concern about using GenAI and commercialized LLMs, the participants were informed before entering the study, and they were provided the choice to not use the LLM-based agents when using the apps in the studies. Furthermore, a reminder was provided each time the person entered the LLM-based agent module that it was using an LLM, fo the person to be able to choose to not use the agent module.

A major ethical concern in persuasive technologies and behavior change systems is this idea of coercion and the risks that follow it. Using AI systems for health is at par with the AI Act's high-risk AI systems [52] and the use of GenAI on sensitive data, which in turn is a concern at par with GDPR [25][1]. The applications in this thesis fall into the high-risk category of the EU AI Act. As research tools, they are governed and controlled by the ethical approval as well as under Article 2 of the EU AI Act. Though it is important to note that these apps would fall under different regulation if they were to transition from research to "market", which would include obtaining a CE marking, undergoing compliance checks, and monitoring. This has consequences for future work, affecting potentials of leading to something that is actually used in practice and daily use.

Paper III's activity-theoretical framework names transparency, data ownership, and designing against manipulation and the erosion of autonomy as ethical priorities for personal health technologies, with references to the risk genAI system pose [127]. A later paper, Paper V's ethical considerations naming agency, control, oversight, and accountability, is an instance of the concerns raised in Paper III. Considering these two papers jointly shows that the theoretical work considered the ethical problem before the systems caught up, which is usually the other way around in real world technologies. Further, the participatory design in Papers II and IV thoroughly explored ethical considerations relating to AI-based behavior change systems, which is the methodology that incorporates the value-based design promoted in AI research [159].

[1] "...data sent to the OpenAI API is not used to train or improve OpenAI models": https://developers.openai.com/api/docs/guides/your-data

Bibliography

[1] Icek Ajzen. "The theory of planned behavior". In: *Organizational Behavior and Human Decision Processes*. Theories of Cognitive Self-Regulation 50.2 (1991), pp. 179–211. DOI: 10.1016/0749-5978(91)90020-T.

[2] Zeynep Akata et al. "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence". In: *Computer* 53.8 (2020), pp. 18–28. DOI: 10.1109/MC.2020.2996587.

[3] Khalid M. Alameer, Sjir Uitdewilligen, and Ute R. Hülsheger. "What Are the Active Ingredients in Recovery Activities? Introducing a Dimensional Approach." In: *Journal of Occupational Health Psychology* 28.4 (2023), pp. 239–262. DOI: 10.1037/ocp0000354.

[4] Kars Alfrink et al. "Contestable AI by Design: Towards a Framework". In: *Minds and Machines* 33.4 (2023), pp. 613–639. DOI: 10.1007/s11023-022-09611-z.

[5] Anees Alyafei and Raul Easton-Carr. "The Health Belief Model of Behavior Change". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2026. PubMed ID: 39163427.

[6] Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. "Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior". In: *Frontiers in Artificial Intelligence* 8 (2025), p. 1622292. DOI: 10.3389/frai.2025.1622292.

[7] Ingela Aronsson et al. ""Recovery activities are needed every step of the way"—exploring the process of long-term recovery in people previously diagnosed with exhaustion disorder". In: *BMC Psychology* 12.1 (May 2024), p. 248. DOI: 10.1186/s40359-024-01756-z.

[8] Katie Atkinson and Trevor Bench-Capon. "Practical reasoning as presumptive argumentation using action based alternating transition systems". In: *Artificial Intelligence*. Argumentation in Artificial Intelligence 171.10 (2007), pp. 855–874. DOI: 10.1016/j.artint.2007.04.009.

[9] Christiane Attig et al. *Measuring Successful Cooperation in Human-AI Teamwork: Development and Validation of the Perceived Cooperativity and Teaming Perception Scales*. Version Number: 1. 2026. DOI: 10.48550/ARXIV.2604.24461. URL: https://arxiv.org/abs/2604.24461.

[10] Christiane Attig et al. "Understanding successful human–AI teaming: The role of goal alignment and AI autonomy for social perception of LLM-based chatbots". In: *Computers in Human Behavior: Artificial Humans* 7 (2026), p. 100246. DOI: 10.1016/j.chbah.2025.100246.

[11] Albert Bandura. *Self-efficacy: The exercise of control*. Self-efficacy: The exercise of control. Pages: ix, 604. New York, NY, US: W H Freeman/Times Books/ Henry Holt & Co, 1997. ISBN: 978-0-7167-2626-5.

[12] Albert Bandura. *Social foundations of thought and action: A social cognitive theory*. Social foundations of thought and action: A social cognitive theory. Pages: xiii, 617. Englewood Cliffs, NJ, US: Prentice-Hall, Inc, 1986.

[13] Christoph Bartneck et al. "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots". In: *International Journal of Social Robotics* 1.1 (2009), pp. 71–81. DOI: 10.1007/s12369-008-0001-3.

[14] Jayalakshmi Baskar et al. "A Multipurpose Goal Model for Personalised Digital Coaching". In: *Agents and Multi-Agent Systems for Health Care*. Ed. by Sara Montagna et al. Cham: Springer International Publishing, 2017, pp. 94–116. ISBN: 978-3-319-70887-4. DOI: 10.1007/978-3-319-70887-4_6.

[15] Amit Baumel, Theresa Fleming, and Stephen M Schueller. "Digital Micro Interventions for Behavioral and Mental Health Gains: Core Components and Conceptualization of Digital Micro Intervention Care". In: *Journal of Medical Internet Research* 22.10 (2020), e20631. DOI: 10.2196/20631.

[16] T. Bench-Capon. *Value Based Argumentation Frameworks*. Version Number: 1. 2002. DOI: 10.48550/ARXIV.CS/0207059. URL: https://arxiv.org/abs/cs/0207059.

[17] T. J. M. Bench-Capon and Paul E. Dunne. "Argumentation in Artificial Intelligence". In: *Artificial Intelligence*. Argumentation in Artificial Intelligence 171.10 (2007), pp. 619–641. DOI: 10.1016/j.artint.2007.05.001.

[18] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922.

[19] Daniel Berdichevsky and Erik Neuenschwander. "Toward an Ethics of Persuasive Technology". In: *Communications of the ACM* 42.5 (1999), pp. 51–58. DOI: 10.1145/301353.301410.

[20] Erwin P. Bettinghaus. "Health promotion and the knowledge-attitude-behavior continuum". In: *Preventive Medicine* 15.5 (1986), pp. 475–491. DOI: 10.1016/0091-7435(86)90025-3.

[21] Timothy W. Bickmore and Rosalind W. Picard. "Establishing and maintaining long-term human-computer relationships". In: *ACM Transactions on Computer-Human Interaction* 12.2 (2005), pp. 293–327. DOI: 10.1145/1067860.1067867.

[22] Allison A M Bielak, Jacqueline A Mogle, and Martin J Sliwinski. "Within- and Between-Person Differences in Activity Factor Structure: Results From an Ecological Momentary Assessment Study". In: *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 80.5 (2025), gbaf017. DOI: 10.1093/geronb/gbaf017.

[23] Abeba Birhane et al. "Power to the People? Opportunities and Challenges for Participatory AI". In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '22. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 1–8. DOI: 10.1145/3551624.3555290.

[24] Gallus Bischof, Anja Bischof, and Hans-Jürgen Rumpf. "Motivational Interviewing: An Evidence-Based Approach for Use in Medical Practice". In: *Deutsches Ärzteblatt International* 118.7 (2021), pp. 109–115. DOI: 10.3238/arztebl.m2021.0014.

[25] European Data Protection Board. *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*. 2024.

[26] Susanne Bødker. *Through the Interface: A Human Activity Approach to User Interface Design*. 1st ed. Boca Raton: CRC Press, 2021. ISBN: 978-1-003-06397-1. DOI: 10.1201/9781003063971.

[27] Susanne Bødker et al. *Participatory Design*. Synthesis Lectures on Human-Centered Informatics. Cham: Springer International Publishing, 2022. ISBN: 978-3-031-01107-8. DOI: 10.1007/978-3-031-02235-7.

[28] Edward S. Bordin. "The generalizability of the psychoanalytic concept of the working alliance." In: *Psychotherapy: Theory, Research & Practice* 16.3 (1979), pp. 252–260. DOI: 10.1037/h0085885.

[29] Cedric Bousquet and Divà Beltramin. "Advantages and Inconveniences of a Multi-Agent Large Language Model System to Mitigate Cognitive Biases in Diagnostic Challenges". In: *Journal of Medical Internet Research* 27 (Jan. 2025), e69742. DOI: 10.2196/69742.

[30] Ryan L. Boyd and David M. Markowitz. “Artificial Intelligence and the Psychology of Human Connection”. In: *Perspectives on Psychological Science* 21.2 (Jan. 2026), pp. 192–220. DOI: 10.1177/17456916251404394.

[31] Jeffrey M. Bradshaw et al. “Dimensions of Adjustable Autonomy and Mixed-Initiative Interaction”. In: *Agents and Computational Autonomy*. Ed. by Gerhard Goos et al. Vol. 2969. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 17–39. ISBN: 978-3-540-22477-8. DOI: 10.1007/978-3-540-25928-2_3.

[32] Michael E. Bratman. “Shared Cooperative Activity”. In: *The Philosophical Review* 101.2 (1992), p. 327. DOI: 10.2307/2185537.

[33] Virginia Braun and Victoria Clarke. *Thematic analysis: a practical guide*. London ; Thousand Oaks, California: SAGE, 2022. ISBN: 978-1-4739-5323-9.

[34] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa.

[35] John Brooke. “SUS – a quick and dirty usability scale”. In: *Usability Evaluation in Industry*. Ed. by Patrick W. Jordan et al. CRC Press, Jan. 1996, pp. 189–194. ISBN: 978-0-429-15701-1. DOI: 10.1201/9781498710411.

[36] Christine Bundy. “Changing behaviour: using motivational interviewing techniques.” In: *Journal of the Royal Society of Medicine* 97.Suppl 44 (2004), pp. 43–47. PubMedCentral: PMC1308798.

[37] M. D. Cabana et al. “Why don’t physicians follow clinical practice guidelines? A framework for improvement”. In: *JAMA* 282.15 (Oct. 1999), pp. 1458–1465. DOI: 10.1001/jama.282.15.1458.

[38] J. Calderhead. “Stimulated Recall: A Method for Research on Teaching”. In: *British Journal of Educational Psychology* 51.2 (1981), pp. 211–217. DOI: 10.1111/j.2044-8279.1981.tb02474.x.

[39] James Cane, Denise O’Connor, and Susan Michie. “Validation of the theoretical domains framework for use in behaviour change and implementation research”. In: *Implementation Science* 7.1 (2012), p. 37. DOI: 10.1186/1748-5908-7-37.

[40] Rachel N Carey et al. “Behavior Change Techniques and Their Mechanisms of Action: A Synthesis of Links Described in Published Intervention Literature”. In: *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* 53.8 (Oct. 2018), pp. 693–707. DOI: 10.1093/abm/kay078.

[41] Micah Carroll et al. “AI alignment with changing and influenceable reward functions”. In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. ICML’24. Vienna, Austria: JMLR.org, 2024, pp. 5706–5756.

[42] Charles S. Carver and Michael F. Scheier. “Control theory: A useful conceptual framework for personality–social, clinical, and health psychology”. In: *Psychological Bulletin* 92.1 (1982), pp. 111–135. DOI: 10.1037/0033-2909.92.1.111.

[43] Federico Castagna et al. “Computational Argumentation-based Chatbots: A Survey”. In: *Journal of Artificial Intelligence Research* 80 (2024), pp. 1271–1310. DOI: 10.1613/jair.1.15407.

[44] Cristiano Castelfranchi and Rino Falcone. “Towards a theory of delegation for agent-based systems”. In: *Robotics and Autonomous Systems*. Multi-Agent Rationality 24.3 (1998), pp. 141–157. DOI: 10.1016/S0921-8890(98)00028-1.

[45] Daniela Castillo and Lawrence Farrugia Caruana. “Unveiling Customer Expectations of Chatbot Interactions: A Systematic Literature Review and Research Agenda”. In: *International Journal of Human–Computer Interaction* 42.14 (2026), pp. 10767–10794. DOI: 10.1080/10447318.2025.2588815.

[46] Francesca Cellina et al. “A Large Scale, App-Based Behaviour Change Experiment Persuading Sustainable Mobility Patterns: Methods, Results and Lessons Learnt”. In: *Sustainability* 11.9 (2019), p. 2674. DOI: 10.3390/su11092674.

[47] Chalaguine Lisa A. and Hunter Anthony. “A Persuasive Chatbot Using a Crowd-Sourced Argument Graph and Concerns”. In: *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2020. DOI: 10.3233/FAIA200487.

[48] Carlos Chesñevar et al. “Towards an argument interchange format”. In: *The Knowledge Engineering Review* 21.4 (2006), pp. 293–316. DOI: 10.1017/S0269888906001044.

[49] Jah Yeon Choi et al. “A Smartphone App (AnSim) With Various Types and Forms of Messages Using the Transtheoretical Model for Cardiac Rehabilitation in Patients With Coronary Artery Disease: Development and Usability Study”. In: *JMIR Medical Informatics* 9.12 (2021), e23285. DOI: 10.2196/23285. PubMed ID: 34878987.

[50] K Christoffersen and D Woods. “How to make automated systems team players”. In: *Advances in Human Performance and Cognitive Engineering Research*. Vol. 2. Elsevier, 2002, pp. 1–12. ISBN: 978-0-7623-0748-7. DOI: 10.1016/S1479-3601(02)02003-9.

[51] Philip R. Cohen and Hector J. Levesque. “Teamwork”. In: *Noûs* 25.4 (1991), p. 487. DOI: 10.2307/2216075.

[52] European Commission. *Artificial Intelligence in Healthcare - Public Health - European Commission*. 2026.

[53] Lauren E. Connell et al. "Links Between Behavior Change Techniques and Mechanisms of Action: An Expert Consensus Study". In: *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* 53.8 (2019), pp. 708–720. DOI: 10.1093/abm/kay082.

[54] Sunny Consolvo, David W. McDonald, and James A. Landay. "Theory-Driven Design Strategies for Technologies That Support Behavior Change in Everyday Life". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston MA USA: ACM, 2009, pp. 405–414. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518766.

[55] Edward Meredith Cope and John Edwin Sandys. *The Rhetoric of Aristotle*. UK: Legare Street Press, 2023. ISBN: 978-1-02-037932-1.

[56] Kristijonas Čyras et al. *Argumentative XAI: A Survey*. 2021. DOI: 10.48550/ARXIV.2105.11266.

[57] Rachel Davis et al. "Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review". In: *Health Psychology Review* 9.3 (2015), pp. 323–344. DOI: 10.1080/17437199.2014.941722.

[58] Edward L. Deci. *Intrinsic Motivation*. Boston, MA: Springer US, 1975. ISBN: 978-1-4613-4446-9. DOI: 10.1007/978-1-4613-4446-9.

[59] Edward L. Deci and Richard M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. en. Boston, MA: Springer US, 1985. ISBN: 978-1-4899-2273-1. DOI: 10.1007/978-1-4899-2271-7.

[60] Matt Dennis, Judith Masthoff, and Chris Mellish. "Adapting Progress Feedback and Emotional Support to Learner Personality". In: *International Journal of Artificial Intelligence in Education* 26.3 (2016), pp. 877–931. DOI: 10.1007/s40593-015-0059-7.

[61] Sebastian Deterding et al. "Advancing Self-Determination Theory via computational modelling: the case of competence and optimal challenge". In: *Motivation and Emotion* 50.1 (Feb. 2026), pp. 80–99. DOI: 10.1007/s11031-025-10156-8.

[62] E. Diener et al. "Intensity and frequency: dimensions underlying positive and negative affect". In: *Journal of Personality and Social Psychology* 48.5 (1985), pp. 1253–1265. DOI: 10.1037//0022-3514.48.5.1253.

[63] Emmanuelle Dietz, Antonis Kakas, and Loizos Michael. "Argumentation: A Calculus for Human-Centric AI". In: *Frontiers in Artificial Intelligence* 5 (2022). DOI: 10.3389/frai.2022.955579.

[64] Elon University Imagining the Digital Future Center. *The Rise of AI Companions*. 2026. URL: https://imaginingthedigitalfuture.org/reports-and-publications/the-rise-of-ai-companions/.

[65] Leonard Dung. “Current cases of AI misalignment and their implications for future risks”. In: *Synthese* 202.5 (Oct. 2023), p. 138. DOI: 10.1007/s11229-023-04367-0.

[66] Phan Minh Dung. “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games”. In: *Artificial Intelligence* 77.2 (1995), pp. 321–357. DOI: 10.1016/0004-3702(94)00041-X.

[67] Amy Edmondson. “Psychological Safety and Learning Behavior in Work Teams”. In: *Administrative Science Quarterly* 44.2 (1999), pp. 350–383. DOI: 10.2307/2666999.

[68] Niklas Elmqvist et al. *Participatory AI: A Scandinavian Approach to Human-Centered AI*. arXiv:2509.12752 [cs.HC]. June 2026. DOI: 10.48550/arXiv.2509.12752.

[69] Jon Elster. *Ulysses and the Sirens: studies in rationality and irrationality*. Rev. ed. Cambridge paperback library. Cambridge [Cambridgeshire] ; New York : Paris: Cambridge University Press ; Editions de la Maison des sciences de l’homme, 1984. ISBN: 978-0-521-26984-1.

[70] Y. Engeström. “Activity theory as a framework for analyzing and redesigning work”. In: *Ergonomics* 43.7 (July 2000), pp. 960–974. DOI: 10.1080/001401300409143.

[71] Yrjö Engeström. “Expansive Learning at Work: Toward an activity theoretical reconceptualization”. In: *Journal of Education and Work* 14.1 (Feb. 2001), pp. 133–156. DOI: 10.1080/13639080020028747.

[72] Yrjö Engeström. *Learning by expanding: an activity-theoretical approach to developmental research*. Second edition. Cambridge: Cambridge University Press, 2015. ISBN: 978-1-107-07442-2. DOI: 10.1017/CBO9781139814744.

[73] Malin Eriksson et al. “Health Behavioural Change - the Influence of Social-Ecological Factors and Health Identity”. In: *International Journal of Qualitative Studies on Health and Well-Being* 20.1 (2025), p. 2458309. DOI: 10.1080/17482631.2025.2458309. PubMed ID: 39885779.

[74] Catherine F. Eubanks, J. Christopher Muran, and Jeremy D. Safran. “Alliance rupture repair: A meta-analysis”. In: *Psychotherapy* 55.4 (2018), pp. 508–519. DOI: 10.1037/pst0000185.

[75] Mark D. Faries. “Why We Don’t “Just Do It””. In: *American Journal of Lifestyle Medicine* 10.5 (2016), pp. 322–329. DOI: 10.1177/1559827616638017.

[76] Martin Fishbein. “An integrative model for behavioral prediction and its application to health promotion”. In: *Emerging Theories in Health Promotion Practice and Research*. Ed. by Ralph J. DiClemente, Richard Crosby, and Michelle C. Kegler. 2nd ed. Jossey-Bass, 2009. Chap. 8, pp. 214–234. ISBN: 978-0-470-17913-0.

[77] Frank Flemisch et al. “Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations”. In: *Cognition, Technology & Work* 14.1 (2012), pp. 3–18. DOI: 10.1007/s10111-011-0191-6.

[78] B. J. Fogg. *Persuasive technology: using computers to change what we think and do*. The Morgan Kaufmann series in interactive technologies. Amsterdam ; Boston: Morgan Kaufmann Publishers, 2003. ISBN: 978-1-55860-643-2.

[79] B.J. Fogg and Clifford Nass. “Silicon Sycophants: The Effects of Computers That Flatter”. In: *International Journal of Human-Computer Studies* 46.5 (1997), pp. 551–561. DOI: 10.1006/ijhc.1996.0104.

[80] Bj Fogg. “A behavior model for persuasive design”. In: *Proceedings of the 4th International Conference on Persuasive Technology*. Claremont California USA: ACM, Apr. 2009, pp. 1–7. ISBN: 978-1-60558-376-1. DOI: 10.1145/1541948.1541999.

[81] Bj Fogg. “Persuasive computers: perspectives and research directions”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’98*. Los Angeles, California, United States: ACM Press, 1998, pp. 225–232. ISBN: 978-0-201-30987-4. DOI: 10.1145/274644.274677.

[82] Folkhälsomyndigheten. *Statistics on mental health in Sweden - Din psykiska hälsa*. 2024. URL: https://dinpsykiskahalsa.se/artiklar/other-languages/engelska/statistics-on-mental-health-in-sweden/.

[83] Försäkringskassan. *Sjukfrånvaro i psykiatriska diagnoser*. Tech. rep. Försäkringskassan, 2020.

[84] John Fox et al. “Argumentation-Based Inference and Decision Making–A Medical Perspective”. In: *IEEE Intelligent Systems* 22.6 (2007), pp. 34–41. DOI: 10.1109/MIS.2007.102.

[85] Matija Franklin et al. *Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI*. Version Number: 2. 2022. DOI: 10.48550/ARXIV.2203.10525.

[86] Gabriel Freedman et al. “Argumentative Large Language Models for Explainable and Contestable Claim Verification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.14 (2025), pp. 14930–14939. DOI: 10.1609/aaai.v39i14.33637.

[87] Iason Gabriel. “Artificial Intelligence, Values, and Alignment”. In: *Minds and Machines* 30.3 (Sept. 2020), pp. 411–437. DOI: 10.1007/s11023-020-09539-2.

[88] Andrew Gambino, Jesse Fox, and Rabindra Ratan. “Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm”. In: *Human-Machine Communication* 1 (2020), pp. 71–86. DOI: 10.30658/hmc.1.5.

[89] Benjamin Gardner. "A review and analysis of the use of 'habit' in understanding, predicting and influencing health-related behaviour". In: *Health Psychology Review* 9.3 (2015), pp. 277–295. DOI: 10.1080/17437199.2013.876238.

[90] Susan M. Gass et al. *Stimulated Recall Methodology in Second Language Research*. 0th ed. Routledge, 2013. ISBN: 978-1-135-66978-2. DOI: 10.4324/9781410606006.

[91] Peter M. Gollwitzer. "Implementation intentions: Strong effects of simple plans". In: *American Psychologist* 54.7 (1999), pp. 493–503. DOI: 10.1037/0003-066X.54.7.493.

[92] Floriana Grasso, Alison Cawsey, and Ray Jones. "Dialectical Argumentation to Solve Conflicts in Advice Giving: A Case Study in the Promotion of Healthy Nutrition". In: *International Journal of Human-Computer Studies* 53.6 (2000), pp. 1077–1115. DOI: 10.1006/ijhc.2000.0429.

[93] Mary L. Greaney et al. "Social Support for Changing Multiple Behaviors: Factors Associated With Seeking Support and the Impact of Offered Support". In: *Health Education & Behavior: The Official Publication of the Society for Public Health Education* 45.2 (Apr. 2018), pp. 198–206. DOI: 10.1177/1090198117712333.

[94] Joan M. Greenbaum and Morten Kyng, eds. *Design at work: cooperative design of computer systems*. Hillsdale, N.J: L. Erlbaum Associates, 1991. ISBN: 978-0-8058-0611-3.

[95] Barbara J. Grosz and Sarit Kraus. "Collaborative plans for complex group action". In: *Artificial Intelligence* 86.2 (1996), pp. 269–357. DOI: 10.1016/0004-3702(95)00103-4.

[96] Greg Guest, Arwen Bunce, and Laura Johnson. "How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability". In: *Field Methods* 18.1 (2006), pp. 59–82. DOI: 10.1177/1525822X05279903.

[97] Frida Lundin Gurné et al. "Seeking lifestyle counselling at primary health care centres: a cross-sectional study in the Swedish population". In: *BMC Primary Care* 24 (Mar. 2023), p. 74. DOI: 10.1186/s12875-023-02035-3.

[98] Dylan Hadfield-Menell et al. "Cooperative Inverse Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016.

[99] Emmanuel Hadoux and Anthony Hunter. "Comfort or Safety? Gathering and Using the Concerns of a Participant for Better Persuasion". In: *Argument & Computation* 10.2 (2019), pp. 113–147. DOI: 10.3233/AAC-191007.

[100] Martin S. Hagger and Nikos L.D. Chatzisarantis. "An Integrated Behavior Change Model for Physical Activity". In: *Exercise and Sport Sciences Reviews* 42.2 (2014), pp. 62–69. DOI: 10.1249/JES.0000000000000008.

[101] Charles L. Hamblin. *Fallacies*. Socorro, NM: Advanced Reasoning Forum, 1970. ISBN: 978-1-938421-66-2.

[102] Robert P. Hawkins et al. "Understanding tailoring in communicating about health". In: *Health education research* 23.3 (2008), pp. 454–466. DOI: 10.1093/her/cyn004.

[103] Eric B. Hekler et al. "Mind the Theoretical Gap: Interpreting, Using, and Developing Behavioral Theory in HCI Research". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris France: ACM, 2013, pp. 3307–3316. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2466452.

[104] Monique Hennink and Bonnie N. Kaiser. "Sample sizes for saturation in qualitative research: A systematic review of empirical tests". In: *Social Science & Medicine* 292 (2022), p. 114523. DOI: 10.1016/j.socscimed.2021.114523.

[105] Kristin E. Heron and Joshua M. Smyth. "Ecological Momentary Interventions: Incorporating Mobile Technology into Psychosocial and Health Behaviour Treatments". In: *British Journal of Health Psychology* 15.Pt 1 (2010), pp. 1–39. DOI: 10.1348/135910709X466063. PubMed ID: 19646331.

[106] Tad Hirsch et al. "Designing Contestability: Interaction Design, Machine Learning, and Mental Health". In: *DIS. Designing Interactive Systems (Conference)* 2017 (2017), pp. 95–99. DOI: 10.1145/3064663.3064703.

[107] Godfrey M. Hochbaum. "Public participation in medical screening programs; a socio-psychological study - National Library of Medicine Institution". In: *Public Health Service Publication* 572 (1958), pp. 1–14. URL: https://catalog.nlm.nih.gov/discovery/fulldisplay/alma9969295340667601NLM_INST:01NLM_INST.

[108] Wilhelm Hofmann, Malte Friese, and Reinout W. Wiers. "Impulsive versus reflective influences on health behavior: a theoretical framework and empirical review". In: *Health Psychology Review* 2.2 (2008), pp. 111–137. DOI: 10.1080/17437190802617668.

[109] Eric Horvitz. "Principles of mixed-initiative user interfaces". In: *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*. Pittsburgh, Pennsylvania, United States: ACM Press, 1999, pp. 159–166. ISBN: 978-0-201-48559-2. DOI: 10.1145/302979.303030.

[110] Human Clarity Institute. *AI Companionship & Human Connection 2025 (Dataset, n = 501)*. 2025. DOI: 10.5281/ZENODO.18025716. URL: https://zenodo.org/doi/10.5281/zenodo.18025716 (visited on 09/18/2026).

[111] Jing Wen Hung et al. “Parasocial relationships with artificial intelligence (AI): A systematic review of benefits and risks”. In: *Computers in Human Behavior: Artificial Humans* 8 (May 2026), p. 100323. DOI: 10.1016/j.chbah.2026.100323.

[112] Anthony Hunter. “Computational Persuasion with Applications in Behaviour Change”. In: *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2016. DOI: 10.3233/978-1-61499-686-6-5.

[113] Anthony Hunter. “Towards a Framework for Computational Persuasion with Applications in Behaviour Change”. In: *Argument & Computation* 9.1 (2018), pp. 15–40. DOI: 10.3233/AAC-170032.

[114] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. “An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study”. In: *JMIR mHealth and uHealth* 6.11 (2018), e12106. DOI: 10.2196/12106.

[115] Erica James et al. “Comparative efficacy of simultaneous versus sequential multiple health behavior change interventions among adults: A systematic review of randomised trials”. In: *Preventive Medicine* 89 (Aug. 2016), pp. 211–223. DOI: 10.1016/j.ypmed.2016.06.012.

[116] Karen A. Jehn. “A Multimethod Examination of the Benefits and Detriments of Intragroup Conflict”. In: *Administrative Science Quarterly* 40.2 (1995), p. 256. DOI: 10.2307/2393638.

[117] Nick R. Jennings. “Commitments and conventions: The foundation of coordination in multi-agent systems”. In: *The Knowledge Engineering Review* 8.3 (1993), pp. 223–250. DOI: 10.1017/S0269888900000205. (Visited on 09/11/2026).

[118] Jiaming Ji et al. “AI Alignment: A Contemporary Survey”. In: *ACM Computing Surveys* 58.5 (Nov. 2025), 132:1–132:38. DOI: 10.1145/3770749.

[119] Matthew Johnson et al. “Coactive Design: Designing Support for Interdependence in Joint Activity”. In: *Journal of Human-Robot Interaction* 3.1 (2014), p. 43. DOI: 10.5898/JHRI.3.1.Johnson.

[120] Matthew Johnson et al. “The Fundamental Principle of Coactive Design: Interdependence Must Shape Autonomy”. In: *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*. Ed. by Marina De Vos et al. Berlin, Heidelberg: Springer, 2011, pp. 172–191. ISBN: 978-3-642-21268-0. DOI: 10.1007/978-3-642-21268-0_10.

[121] Sara S. Johnson et al. “Coaction in multiple behavior change interventions: consistency across multiple studies on weight management and obesity prevention”. In: *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association* 33.5 (2014), pp. 475–480. DOI: 10.1037/a0034215.

[122] Marie Johnston et al. “Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis and expert consensus”. In: *Translational Behavioral Medicine* 11.5 (2020), pp. 1049–1065. DOI: 10.1093/tbm/ibaa050.

[123] Vera C. Kaelin et al. “Characteristics of Digital Twins to Visualize, Monitor and Predict Outcomes on Participation in Meaningful Activities”. In: *Frontiers in Psychiatry* 16 (2026). DOI: 10.3389/fpsyt.2025.1694966.

[124] Daniel Kahneman. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux, 2011. ISBN: 978-0-374-27563-1.

[125] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. “Coercion and Deception in Persuasive Technologies”. In: *TRUST@AAMAS*. 2018. URL: https://api.semanticscholar.org/CorpusID:52066920.

[126] Maurits Kaptein et al. “Personalizing Persuasive Technologies: Explicit and Implicit Personalization Using Persuasion Profiles”. In: *International Journal of Human-Computer Studies* 77 (2015), pp. 38–51. DOI: 10.1016/j.ijhcs.2015.01.004.

[127] Victor Kaptelinin, Kaan Kilic, and Helena Lindgren. “Behavior change beyond intervention: an activity-theoretical perspective on human-centered design of personal health technology”. In: *Frontiers in Digital Health* 8 (2026), p. 1754399. DOI: 10.3389/fdgth.2026.1754399.

[128] Victor Kaptelinin, Kari Kuutti, and Liam Bannon. “Activity theory: Basic concepts and applications: A summary of a tutorial given at the east west HCI95 conference”. In: *Human-Computer Interaction*. Ed. by Gerhard Goos et al. Vol. 1015. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 189–201. ISBN: 978-3-540-60614-7. DOI: 10.1007/3-540-60614-9_14.

[129] Victor Kaptelinin and Bonnie A. Nardi. *Acting with technology: activity theory and interaction design*. 1. MIT Press paperback ed. Acting with technology. Cambridge, Mass. London: MIT Press, 2009. ISBN: 978-0-262-51331-9.

[130] Helen Keller. *Script for Helen Keller and Anne Sullivan Macy's Vaudeville Performances*.

[131] Michael P. Kelly and Mary Barker. “Why is changing health-related behaviour so difficult?” In: *Public Health* 136 (2016), pp. 109–116. DOI: 10.1016/j.puhe.2016.03.030.

[132] Mohsen Khosravi et al. “A systematic review of the limitations of large language models in generating healthcare content”. In: *PLOS Digital Health* 5.4 (2026), e0001354. DOI: 10.1371/journal.pdig.0001354.

[133] Jan Kiesewetter and Martin R. Fischer. “The Teamwork Assessment Scale: A Novel Instrument to Assess Quality of Undergraduate Medical Students’ Teamwork Using the Example of Simulation-based Ward-Rounds”. In: *GMS Zeitschrift für Medizinische Ausbildung* 32.2 (2015), Doc19. DOI: 10.3205/zma000961.

[134] Kaan Kilic, Vera Kaelin, and Helena Lindgren. “Aime: Wellbeing through Recovery Activities with Digital Companions for Cognitive Support”. Submitted. 2027.

[135] Kaan Kilic and Helena Lindgren. “To Challenge or Support: Managing Behavior in Dialogues for Promoting Health Using Argument Schemes with a Digital Companion for Human-AI Teaming”. In: *User Modeling and User-Adapted Interaction* (2026), pp. 1–68. Submitted.

[136] Kaan Kilic et al. “Argument-based human–AI collaboration for supporting behavior change to improve health”. In: *Frontiers in Artificial Intelligence* 6 (Feb. 2023). DOI: 10.3389/frai.2023.1069455.

[137] Kaan Kilic et al. “Human-Swarm Teaming for Personalized Behavior Change Support: An In-The-Wild Feasibility Study”. submitted. 2027.

[138] Jonathan Kim et al. “Limitations of large language models in clinical problem-solving arising from inflexible reasoning”. In: *Scientific Reports* 15 (2025), p. 39426. DOI: 10.1038/s41598-025-22940-0.

[139] Yubin Kim et al. “Capable language models can outgrow the benefits of collaboration”. In: *Nature Machine Intelligence* 8.7 (2026), pp. 1157–1172. DOI: 10.1038/s42256-026-01268-y.

[140] G. Klien et al. “Ten challenges for making automation a ”team player” in joint human-agent activity”. In: *IEEE Intelligent Systems* 19.6 (2004), pp. 91–95. DOI: 10.1109/MIS.2004.74.

[141] Juliane Kloidt and Lawrence W. Barsalou. “Successful Outcomes, Fulfilment, and Positive Affect: Establishing Facets of Eustress Across Real-Life Challenging Situations”. In: *Social Indicators Research* 184.1 (2026), p. 20. DOI: 10.1007/s11205-026-03909-6.

[142] Gerjo Kok et al. “A Taxonomy of Behaviour Change Methods: An Intervention Mapping Approach”. In: *Health Psychology Review* 10.3 (2016), pp. 297–312. DOI: 10.1080/17437199.2015.1077155.

[143] Nadin Kökciyan et al. “A Collaborative Decision Support Tool for Managing Chronic Conditions”. In: *Studies in Health Technology and Informatics* 264 (2019), pp. 644–648. DOI: 10.3233/SHTI190302. PubMed ID: 31438003.

[144] Adamos Koumi and Antonis Kakas. “LLM-COGNICA: Towards Reliable Agent Decision Making”. In: *Artificial Intelligence Applications and Innovations*. Ed. by Ilias Maglogiannis et al. Vol. 793. Cham: Springer Nature Switzerland, 2027, pp. 87–101. ISBN: 978-3-032-30801-6. DOI: 10.1007/978-3-032-30801-6_7.

[145] Harsh Kumar et al. “Using Adaptive Bandit Experiments to Increase and Investigate Engagement in Mental Health”. In: *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence* 38.21 (2024), pp. 22906–22912. DOI: 10.1609/aaai.v38i21.30328. PubMed ID: 38666291.

[146] Florian Künzler et al. “Exploring the State-of-Receptivity for mHealth Interventions”. In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3.4 (2019), p. 140. DOI: 10.1145/3369805. PubMed ID: 36159209.

[147] Kari Kuutti. “Activity Theory as a Potential Framework for Human-Computer Interaction Research”. In: *Context and Consciousness*. Ed. by Bonnie A. Nardi. The MIT Press, 1995, pp. 17–44. ISBN: 978-0-262-28041-9. DOI: 10.7551/mitpress/2137.003.0006.

[148] Phillippa Lally et al. “How are habits formed: Modelling habit formation in the real world”. In: *European Journal of Social Psychology* 40.6 (2010), pp. 998–1009. DOI: 10.1002/ejsp.674.

[149] Francesco Leofante et al. “Contestable AI Needs Computational Argumentation”. In: *Proceedings of the TwentyFirst International Conference on Principles of Knowledge Representation and Reasoning*. Hanoi, Vietnam: International Joint Conferences on Artificial Intelligence Organization, 2024, pp. 888–896. ISBN: 978-1-956792-05-8. DOI: 10.24963/kr.2024/83.

[150] A. N. Leont’ev. *Activity, consciousness, and personality*. Englewood Cliffs, N.J.: Prentice-Hall, 1978. ISBN: 978-0-13-003533-2.

[151] A. N. Leont’ev. “The Problem of Activity in Psychology”. In: *Soviet Psychology* 13.2 (Dec. 1974), pp. 4–33. DOI: 10.2753/RPO1061-040513024.

[152] A. N. Leont’ev and Mike Cole. *The development of mind: selected works*. OCLC: 424541132. Pacifica, CA: Marxists Internet Archive, 2009. ISBN: 978-0-9805428-6-8.

[153] Jeffery A. Lepine, Nathan P. Podsakoff, and Marcie A. Lepine. “A Meta-Analytic Test of the Challenge Stressor–Hindrance Stressor Framework: An Explanation for Inconsistent Relationships Among Stressors and Performance”. In: *Academy of Management Journal* 48.5 (2005), pp. 764–775. DOI: 10.5465/amj.2005.18803921.

[154] Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474.

[155] Han Li and Renwen Zhang. “Algorithmic accommodation: linguistic alignment in human-AI relational engagement”. In: *Communication and Change* 2.1 (2026), p. 11. DOI: 10.1007/s44382-026-00032-5.

[156] Qian Li et al. “Human-like conversational agents as social partners: a scoping review of socioaffective mechanisms, well-being outcomes, risks and governance in the post-Turing era”. In: *Frontiers in Artificial Intelligence* 9 (July 2026). DOI: 10.3389/frai.2026.1810097.

[157] Sarah Lichtenstein and Paul Slovic, eds. *The Construction of Preference*. 1st ed. Cambridge University Press, 2006. ISBN: 978-0-511-61803-1. DOI: 10.1017/CBO9780511618031.

[158] Helena Lindgren and Kaan Kilic. “Participatory design and evaluation of digital coaching for improving health—the star multicomponent lifestyle intervention”. In: *Frontiers in Digital Health* 7 (June 2025). DOI: 10.3389/fdgth.2025.1600535.

[159] Helena Lindgren, Kristina Lindvall, and Linda Richter-Sundberg. “Responsible design of an AI system for health behavior change—an ethics perspective on the participatory design process of the STAR-C digital coach”. In: *Frontiers in Digital Health* 7 (Mar. 2025). DOI: 10.3389/fdgth.2025.1436347.

[160] Helena Lindgren and Saskia Weck. “Contextualising Goal Setting for Behaviour Change – from Baby Steps to Value Directions”. In: *Proceedings of the 33rd European Conference on Cognitive Ergonomics*. Kaiserslautern Germany: ACM, 2022, pp. 1–7. ISBN: 978-1-4503-9808-4. DOI: 10.1145/3552327.3552342.

[161] Helena Lindgren and Chunli Yan. “ACKTUS: A Platform for Developing Personalized Support Systems in the Health Domain”. In: *Proceedings of the 5th International Conference on Digital Health 2015*. Florence Italy: ACM, 2015, pp. 135–142. ISBN: 978-1-4503-3492-1. DOI: 10.1145/2750511.2750526.

[162] Helena Lindgren et al. “The STAR-C Intelligent Coach: A Cross-Disciplinary Design Process of a Behavior Change Intervention in Primary Care”. In: *Studies in Health Technology and Informatics* 273 (2020), pp. 203–208. DOI: 10.3233/SHTI200640. PubMed ID: 33087613.

[163] Edwin A. Locke and Gary P. Latham. “Building a practically useful theory of goal setting and task motivation: A 35-year odyssey”. In: *American Psychologist* 57.9 (2002), pp. 705–717. DOI: 10.1037/0003-066X.57.9.705.

[164] Matthew Lombard and Kun Xu. “Social Responses to Media Technologies in the 21st Century: The Media Are Social Actors Paradigm”. In: *Human-Machine Communication* 2 (2021), pp. 29–55. DOI: 10.30658/hmc.2.2.

[165] Pippa Lord, Mary Atkinson, and Holly Mitchell. *Mentoring and Coaching for Professionals: A Study of the Research Evidence*. United Kingdom, 2008. URL: https://www.nfer.ac.uk/media/henbjb0b/mentoring_and_coaching_for_professionals_a_study_of_the_research_evidence.pdf.

[166] Joseph B. Lyons et al. “Human–Autonomy Teaming: Definitions, Debates, and Directions”. In: *Frontiers in Psychology* 12 (2021), p. 589585. DOI: 10.3389/fpsyg.2021.589585.

[167] J. D. Mackenzie. “Question-Begging in Non-Cumulative Systems”. In: *Journal of Philosophical Logic* 8.1 (1979), pp. 117–133. DOI: 10.1007/BF00258422.

[168] Aikaterina Manoli et al. “Digital Companionship: Overlapping Uses of AI Companions and AI Assistants”. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Barcelona Spain: ACM, 2026, pp. 1–25. ISBN: 979-8-4007-2278-3. DOI: 10.1145/3772318.3791331.

[169] Marta M. Marques et al. “The Behaviour Change Technique Ontology: Transforming the Behaviour Change Technique Taxonomy v1”. In: *Wellcome Open Research* 8 (2024), p. 308. DOI: 10.12688/wellcomeopenres.19363.2.

[170] Judith Masthoff. “Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers”. In: *Personalized Digital Television: Targeting Programs to Individual Viewers*. Dordrecht: Springer Netherlands, 2004, pp. 93–141. ISBN: 978-1-4020-2164-0. DOI: 10.1007/1-4020-2164-X_5.

[171] Judith Masthoff, Floriana Grasso, and Jaap Ham. “Preface to the Special Issue on Personalization and Behavior Change”. In: *User Modeling and User-Adapted Interaction* 24.5 (2014), pp. 345–350. DOI: 10.1007/s11257-014-9151-1.

[172] Irene Mazzotta, Fiorella de Rosis, and Valeria Carofiglio. “Portia: A User-Adapted Persuasion System in the Healthy-Eating Domain”. In: *IEEE Intelligent Systems* 22.6 (2007), pp. 42–51. DOI: 10.1109/MIS.2007.115.

[173] Peter McBurney and Simon Parsons. “Dialogue Games for Agent Argumentation”. In: *Argumentation in Artificial Intelligence*. Ed. by Guillermo Simari and Iyad Rahwan. Boston, MA: Springer US, 2009, pp. 261–280. ISBN: 978-0-387-98197-0. DOI: 10.1007/978-0-387-98197-0_13.

[174] Jack McKinlay et al. “Understanding the Process of Human-AI Value Alignment”. In: *Journal of Artificial Intelligence Research* 85 (Mar. 2026). DOI: 10.1613/jair.1.18846.

[175] Jonah Meyerhoff et al. “Meeting Young Adults’ Social Support Needs across the Health Behavior Change Journey: Implications for Digital Mental Health Tools”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (Nov. 2022), pp. 1–33. DOI: 10.1145/3555203.

[176] S. Michie et al. “Making psychological theory useful for implementing evidence based practice: a consensus approach”. In: *Quality & Safety in Health Care* 14.1 (2005), pp. 26–33. DOI: 10.1136/qshc.2004.011155.

[177] Susan Michie, Maartje M. van Stralen, and Robert West. "The behaviour change wheel: a new method for characterising and designing behaviour change interventions". In: *Implementation science: IS* 6 (2011), p. 42. DOI: 10.1186/1748-5908-6-42.

[178] Susan Michie et al. "Effective techniques in healthy eating and physical activity interventions: a meta-regression". In: *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association* 28.6 (2009), pp. 690–701. DOI: 10.1037/a0016136.

[179] Susan Michie et al. "The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions". In: *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* 46.1 (2013), pp. 81–95. DOI: 10.1007/s12160-013-9486-6.

[180] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38. DOI: 10.1016/j.artint.2018.07.007.

[181] William R. Miller. "Motivational Interviewing with Problem Drinkers". In: *Behavioural Psychotherapy* 11.2 (1983), pp. 147–172. DOI: 10.1017/S0141347300006583.

[182] William R. Miller and Stephen Rollnick. *Motivational interviewing: helping people change*. Third edition. Applications of motivational interviewing. New York, NY London: Guilford Press, 2013. ISBN: 978-1-60918-227-4.

[183] Sanjay Modgil and Henry Prakken. "The $ASPIC^+$ framework for structured argumentation: a tutorial". In: *Argument & Computation* 5.1 (2014), pp. 31–62. DOI: 10.1080/19462166.2013.869766.

[184] Sanjay Modgil et al. "The Added Value of Argumentation". In: *Agreement Technologies*. Ed. by Sascha Ossowski. Dordrecht: Springer Netherlands, 2013, pp. 357–403. ISBN: 978-94-007-5583-3. DOI: 10.1007/978-94-007-5583-3_21.

[185] Elizabeth W. Morrison. "Employee Voice and Silence". In: *Annual Review of Organizational Psychology and Organizational Behavior* 1.1 (2014), pp. 173–197. DOI: 10.1146/annurev-orgpsych-031413-091328.

[186] Inbal Nahum-Shani et al. "Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support". In: *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* 52.6 (Dec. 2017), pp. 446–462. DOI: 10.1007/s12160-016-9830-8.

[187] Clifford Nass, B.J. Fogg, and Youngme Moon. "Can Computers Be Teammates?" In: *International Journal of Human-Computer Studies* 45.6 (1996), pp. 669–678. DOI: 10.1006/ijhc.1996.0073.

[188] Clifford Nass and Youngme Moon. "Machines and Mindlessness: Social Responses to Computers". In: *Journal of Social Issues* 56.1 (2000), pp. 81–103. DOI: 10.1111/0022-4537.00153.

[189] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. "Computers Are Social Actors". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston Massachusetts USA: ACM, 1994, pp. 72–78. ISBN: 978-0-89791-650-9. DOI: 10.1145/191666.191703.

[190] Clifford Nass et al. "Can Computer Personalities Be Human Personalities?" In: *International Journal of Human-Computer Studies* 43.2 (1995), pp. 223–239. DOI: 10.1006/ijhc.1995.1042.

[191] Johan Y. Y. Ng et al. "Self-Determination Theory Applied to Health Contexts: A Meta-Analysis". In: *Perspectives on Psychological Science* 7.4 (2012), pp. 325–340. DOI: 10.1177/1745691612447309.

[192] Hien Nguyen and Judith Masthoff. "Designing Persuasive Dialogue Systems: Using Argumentation with Care". In: *Persuasive Technology*. Ed. by Harri Oinas-Kukkonen et al. Berlin, Heidelberg: Springer, 2008, pp. 201–212. ISBN: 978-3-540-68504-3. DOI: 10.1007/978-3-540-68504-3_18.

[193] Seth M. Noar and Rick S. Zimmerman. "Health Behavior Theory and Cumulative Knowledge Regarding Health Behaviors: Are We Moving in the Right Direction?" In: *Health Education Research* 20.3 (2005), pp. 275–290. DOI: 10.1093/her/cyg113. PubMed ID: 15632099.

[194] Margareta Norberg et al. "The Västerbotten Intervention Programme: background, design and implications". In: *Global Health Action* 3 (2010), pp. 1–15. DOI: 10.3402/gha.v3i0.4643.

[195] Kenneth Nowack. "Facilitating successful behavior change: Beyond goal setting to goal flourishing." In: *Consulting Psychology Journal: Practice and Research* 69.3 (Sept. 2017), pp. 153–171. DOI: 10.1037/cpb0000088.

[196] Thomas O'Neill et al. "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature". In: *Human Factors* 64.5 (Aug. 2022), pp. 904–938. DOI: 10.1177/0018720820960865.

[197] Jane Ogden. "Celebrating Variability and a Call to Limit Systematisation: The Example of the Behaviour Change Technique Taxonomy and the Behaviour Change Wheel". In: *Health Psychology Review* 10.3 (2016), pp. 245–250. DOI: 10.1080/17437199.2016.1190291. PubMed ID: 27189585.

[198] Harri Oinas-Kukkonen. "A foundation for the study of behavior change support systems". In: *Personal and Ubiquitous Computing* 17.6 (Aug. 2013), pp. 1223–1235. DOI: 10.1007/s00779-012-0591-5.

[199] Harri Oinas-Kukkonen. "Behavior Change Support Systems: A Research Model and Agenda". In: *Persuasive Technology*. Ed. by Thomas Ploug, Per Hasle, and Harri Oinas-Kukkonen. Berlin, Heidelberg: Springer, 2010, pp. 4–14. DOI: 10.1007/978-3-642-13226-1_3.

[200] Harri Oinas-Kukkonen and Marja Harjumaa. “Persuasive Systems Design: Key Issues, Process Model, and System Features”. In: *Communications of the Association for Information Systems* 24 (2009). DOI: 10.17705/1CAIS.02428.

[201] World Health Organization. *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization, 2001.

[202] Rita Orji and Karyn Moffatt. “Persuasive Technology for Health and Wellness: State-of-the-Art and Emerging Trends”. In: *Health Informatics Journal* 24.1 (2018), pp. 66–91. DOI: 10.1177/1460458216650979. PubMed ID: 27245673.

[203] Simon Parsons, Carles Sierra, and Nick Jennings. “Agents That Reason and Negotiate by Arguing”. In: *Journal of Logic and Computation* 8.3 (1998), pp. 261–292. DOI: 10.1093/logcom/8.3.261.

[204] Chaim Perelman and Lucie Olbrechts-Tyteca. *The new rhetoric: a treatise on argumentation*. Reprinted. Notre Dame, Ind: Univ. of Notre Dame Press, 1969. ISBN: 978-0-268-00446-0.

[205] Ethan Perez et al. *Discovering Language Model Behaviors with Model-Written Evaluations*. Version Number: 1. 2022. DOI: 10.48550/ARXIV.2212.09251.

[206] Gjalt-Jorn Peters and Rik Crutzen. “Knowing What We’re Talking About: Facilitating Decentralized, Unequivocal Publication of and Reference to Psychological Construct Definitions and Instructions”. In: *Meta-Psychology* 8 (2024). DOI: 10.15626/MP.2022.3638.

[207] John L. Pollock. “Defeasible Reasoning”. In: *Cognitive Science* 11.4 (1987), pp. 481–518. DOI: 10.1207/s15516709cog1104_4. (Visited on 09/11/2026).

[208] Andrzej Porebski and Jakub Figura. “There is no such thing as conscious artificial intelligence”. In: *Humanities and Social Sciences Communications* 12.1 (Oct. 2025), pp. 1–12. DOI: 10.1057/s41599-025-05868-8.

[209] Henry Prakken. “Formal systems for persuasion dialogue”. In: *The Knowledge Engineering Review* 21.2 (2006), pp. 163–188. DOI: 10.1017/S0269888906000865.

[210] J. O. Prochaska and W. F. Velicer. “The transtheoretical model of health behavior change”. In: *American journal of health promotion: AJHP* 12.1 (1997), pp. 38–48. DOI: 10.4278/0890-1171-12.1.38.

[211] Judith J Prochaska et al. “A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study”. In: *Journal of Medical Internet Research* 23.3 (2021), e24850. DOI: 10.2196/24850. PubMed ID: 33755028.

[212] Judith J. Prochaska, Bonnie Spring, and Claudio R. Nigg. "Multiple Health Behavior Change Research: An Introduction and Overview". In: *Preventive medicine* 46.3 (Mar. 2008), pp. 181–188. DOI: 10.1016/j.ypmed.2008.02.001.

[213] Byron Reeves and Clifford Nass. *The media equation: how people treat computers, television, and new media like real people and places*. 5. [Nachdr.], paperback ed. Stanford, Calif: CSLI Publ, 1996. ISBN: 978-1-57586-052-7.

[214] Stephen Rollnick and William R. Miller. "What Is Motivational Interviewing?" In: *Behavioural and Cognitive Psychotherapy* 23.4 (1995), pp. 325–334. DOI: 10.1017/S135246580001643X.

[215] Irwin M. Rosenstock. "The Health Belief Model and Preventive Health Behavior". In: *Health Education Monographs* 2.4 (1974), pp. 354–386. DOI: 10.1177/109019817400200405.

[216] Michael A. Russell and Jamie M. Gajos. "Annual Research Review: Ecological Momentary Assessment Studies in Child Psychology and Psychiatry". In: *Journal of child psychology and psychiatry, and allied disciplines* 61.3 (2020), pp. 376–394. DOI: 10.1111/jcpp.13204.

[217] Francine Russo. *Personality Can Change from One Hour to the Next*. Sept. 2024. URL: https://www.scientificamerican.com/article/personality-can-change-from-one-hour-to-the-next/.

[218] R. M. Ryan and E. L. Deci. "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being". In: *The American Psychologist* 55.1 (Jan. 2000), pp. 68–78. DOI: 10.1037//0003-066x.55.1.68.

[219] Richard Ryan et al. "Facilitating health behavior change and its maintenance: Interventions based on Self-Determination Theory". In: *Eur. Health Psychol.* 10 (Nov. 2007).

[220] Richard M Ryan and Edward L. Deci, eds. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. Guilford Press, 2017. ISBN: 978-1-4625-3896-6. DOI: 10.1521/978.14625/28806.

[221] Eduardo Salas, Dana E. Sims, and C. Shawn Burke. "Is there a "Big Five" in Teamwork?" In: *Small Group Research* 36.5 (2005), pp. 555–599. DOI: 10.1177/1046496405277134.

[222] Margarete Sandelowski. "Sample size in qualitative research". In: *Research in Nursing & Health* 18.2 (1995), pp. 179–183. DOI: 10.1002/nur.4770180211.

[223] Nevitt Sanford. *Where Colleges Fail: A Study of the Student as a Person*. 4. print. The Jossey-Bass Series in Higher Education. San Francisco: Jossey-Bass, 1969. ISBN: 978-0-87589-003-6.

[224] P. Scerri, D. V. Pynadath, and M. Tambe. “Towards Adjustable Autonomy for the Real World”. In: *Journal of Artificial Intelligence Research* 17 (2002), pp. 171–228. DOI: 10.1613/jair.1037.

[225] Douglas Schuler and Aki Namioka, eds. *Participatory design: principles and practices*. Hillsdale, N.J: L. Erlbaum Associates, 1993. ISBN: 978-0-8058-0952-7.

[226] Ralf Schwarzer. “Modeling Health Behavior Change: How to Predict and Modify the Adoption and Maintenance of Health Behaviors”. In: *Applied Psychology* 57.1 (2008), pp. 1–29. DOI: 10.1111/j.1464-0597.2007.00325.x.

[227] Ramona Schweitzer, Stephan Schlögl, and Marco Schweitzer. “Technology-Supported Behavior Change—Applying Design Thinking to mHealth Application Development”. In: *European Journal of Investigation in Health, Psychology and Education* 14.3 (2024), pp. 584–608. DOI: 10.3390/ejihpe14030039. PubMed ID: 38534900.

[228] Isabella Seeber et al. “Machines as teammates: A research agenda on AI in team collaboration”. In: *Information & Management* 57.2 (2020), p. 103174. DOI: 10.1016/j.im.2019.103174.

[229] Divya Sharma et al. “Reimagining psychiatric care with agentic AI: promise, challenges, and a roadmap forward”. In: *npj Digital Medicine* 9.1 (2026), p. 252. DOI: 10.1038/s41746-026-02453-4.

[230] Mrinank Sharma et al. *Towards Understanding Sycophancy in Language Models*. Version Number: 4. 2023. DOI: 10.48550/ARXIV.2310.13548.

[231] Paschal Sheeran. “Intention—Behavior Relations: A Conceptual and Empirical Review”. In: *European Review of Social Psychology* 12.1 (2002), pp. 1–36. DOI: 10.1080/14792772143000003.

[232] Paschal Sheeran and Thomas L. Webb. “The intention–behavior gap”. In: *Social and Personality Psychology Compass* 10.9 (2016), pp. 503–518. DOI: 10.1111/spc3.12265.

[233] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. “Ecological momentary assessment”. In: *Annual Review of Clinical Psychology* 4 (2008), pp. 1–32. DOI: 10.1146/annurev.clinpsy.3.022806.091415.

[234] Guillermo Simari and Iyad Rahwan, eds. *Argumentation in Artificial Intelligence*. Boston, MA: Springer US, 2009. ISBN: 978-0-387-98197-0. DOI: 10.1007/978-0-387-98197-0.

[235] Molly J. Simis et al. “The lure of rationality: Why does the deficit model persist in science communication?” eng. In: *Public Understanding of Science* 25.4 (2016), pp. 400–414. DOI: 10.1177/0963662516629749.

[236] Marita Skjuve et al. “My Chatbot Companion - a Study of Human-Chatbot Relationships”. In: *International Journal of Human-Computer Studies* 149 (2021), p. 102601. DOI: 10.1016/j.ijhcs.2021.102601.

[237] Elizabeth I. Sklar and Mohammad Q. Azhar. "Explanation through Argumentation". In: *Proceedings of the 6th International Conference on Human-Agent Interaction*. Southampton United Kingdom: ACM, Dec. 2018, pp. 277–285. ISBN: 978-1-4503-5953-5. DOI: 10.1145/3284432.3284470.

[238] Noam Slonim et al. "An Autonomous Debating System". In: *Nature* 591.7850 (2021), pp. 379–384. DOI: 10.1038/s41586-021-03215-w.

[239] Kirsten Ailsa Smith et al. "Adapting Emotional Support to Personality for Carers Experiencing Stress". In: *Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP)*. Dublin, Ireland, 2015, pp. 30–39. DOI: 10.13140/RG.2.1.3898.9929.

[240] Sabine Sonnentag. "The Recovery Paradox: Portraying the Complex Interplay between Job Stressors, Lack of Recovery, and Poor Well-Being". In: *Research in Organizational Behavior* 38 (2018), pp. 169–185. DOI: 10.1016/j.riob.2018.11.002.

[241] Sabine Sonnentag, Bonnie Hayden Cheng, and Stacey L. Parker. "Recovery from Work: Advancing the Field Toward the Future". In: *Annual Review of Organizational Psychology and Organizational Behavior* 9.1 (2022), pp. 33–60. DOI: 10.1146/annurev-orgpsych-012420-091355.

[242] Sabine Sonnentag and Charlotte Fritz. "The Recovery Experience Questionnaire: Development and Validation of a Measure for Assessing Recuperation and Unwinding from Work." In: *Journal of Occupational Health Psychology* 12.3 (2007), pp. 204–221. DOI: 10.1037/1076-8998.12.3.204.

[243] Maria Spante et al. "Activity theory as a tool for reimagining work-integrated learning: Conducting contradiction analysis". In: *Refereed Proceedings of the 5th WACE International Research Symposium on Cooperative and Work-Integrated Education*. Ed. by Jenny Fleming and Karsten E. Zegwaard. Waterloo, Canada: WACE, 2024, pp. 258–263. ISBN: 978-1-7386618-2-4.

[244] F. Sperrle et al. "A Survey of Human-Centered Evaluations in Human-Centered Machine Learning". In: *Computer Graphics Forum* 40.3 (2021), pp. 543–568. DOI: 10.1111/cgf.14329.

[245] Keith Stenning and Michiel van Lambalgen. *Human reasoning and cognitive science*. Cambridge, Mass: MIT Press, 2008. ISBN: 978-0-262-19583-6.

[246] Fritz Strack and Roland Deutsch. "Reflective and impulsive determinants of social behavior". In: *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc* 8.3 (2004), pp. 220–247. DOI: 10.1207/s15327957pspr0803_1.

[247] Pedro J. Teixeira et al. “Exercise, physical activity, and self-determination theory: a systematic review”. In: *The International Journal of Behavioral Nutrition and Physical Activity* 9 (2012), p. 78. DOI: 10.1186/1479-5868-9-78.

[248] Ambuj Tewari and Susan A. Murphy. “From Ads to Interventions: Contextual Bandits in Mobile Health”. In: *Mobile Health*. Ed. by James M. Rehg, Susan A. Murphy, and Santosh Kumar. Cham: Springer International Publishing, 2017, pp. 495–517. ISBN: 978-3-319-51394-2. DOI: 10.1007/978-3-319-51394-2_25.

[249] Rosemary J. Thomas, Judith Masthoff, and Nir Oren. “Can I Influence You? Development of a Scale to Measure Perceived Persuasiveness and Two Studies Showing the Use of the Scale”. In: *Frontiers in Artificial Intelligence* 2 (2019). DOI: 10.3389/frai.2019.00024.

[250] Nava Tintarev and Judith Masthoff. “Evaluating the Effectiveness of Explanations for Recommender Systems”. In: *User Modeling and User-Adapted Interaction* 22.4 (2012), pp. 399–439. DOI: 10.1007/s11257-011-9117-5.

[251] Stephen Edelston Toulmin. *The uses of argument*. Repr. of updated ed. Cambridge: Cambridge University Press, 1958. ISBN: 978-0-521-53483-3.

[252] Bruce W. Tuckman. “Developmental sequence in small groups.” In: *Psychological Bulletin* 63.6 (1965), pp. 384–399. DOI: 10.1037/h0022100.

[253] Miles Turpin et al. *Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. Version Number: 2. 2023. DOI: 10.48550/ARXIV.2305.04388.

[254] Maarten Vansteenkiste and Kennon M. Sheldon. “There’s nothing more practical than a good theory: integrating motivational interviewing and self-determination theory”. In: *The British Journal of Clinical Psychology* 45.Pt 1 (Mar. 2006), pp. 63–82. DOI: 10.1348/014466505X34192.

[255] Yogatheesan Varatharajah and Brent Berry. “A Contextual-Bandit-Based Approach for Informed Decision-Making in Clinical Trials”. In: *Life* 12.8 (2022), p. 1277. DOI: 10.3390/life12081277. PubMed ID: 36013456.

[256] Viswanath Venkatesh, James Y. L. Thong, and Xin Xu. “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology1”. en. In: *MIS Quarterly* 36.1 (Mar. 2012), pp. 157–178. ISSN: 0276-7783, 2162-9730. DOI: 10.2307/41410412. URL: https://misq.umn.edu/misq/article/36/1/157/1470/Consumer-Acceptance-and-Use-of-Information (visited on 10/09/2026).

[257] Laura M. Vowels et al. “From Friends to Lovers: Understanding Motivations and Barriers in AI Companionship”. In: *Archives of Sexual Behavior* (2026). DOI: 10.1007/s10508-025-03375-0.

[258] Lev S. Vygotski and Eugenia Hanfmann. *Thought and Language*. Martino Fine Books, 1986. ISBN: 978-1-61427-244-1.

[259] L. S. Vygotsky. *Mind in Society: Development of Higher Psychological Processes*. Ed. by Michael Cole et al. Harvard University Press, 1978. ISBN: 978-0-674-57628-5. DOI: 10.2307/j.ctvjf9vz4.

[260] Douglas Walton. "Using Argumentation Schemes to Find Motives and Intentions of a Rational Agent". In: *Argument & Computation* 10.3 (2020), pp. 233–275. DOI: 10.3233/AAC-190480.

[261] Douglas N. Walton. *Argumentation Schemes for Presumptive Reasoning*. Hoboken: Taylor and Francis, 1996. ISBN: 978-0-8058-2071-3.

[262] Douglas N. Walton and Erik C. W. Krabbe. *Commitment in dialogue: basic concepts of interpersonal reasoning*. SUNY series in logic and language. Albany, NY: State Univ. of New York Press, 1995. ISBN: 978-0-7914-2586-2.

[263] Douglas N. Walton, Chris Reed, and Fabrizio Macagno. *Argumentation schemes*. Repr. Cambridge: Cambridge University Press, 2008. ISBN: 978-0-521-72374-9.

[264] LianHong Wang et al. "Transtheoretical Model-Based Mobile Health Application for PCOS". In: *Reproductive Health* 19.1 (2022), p. 117. DOI: 10.1186/s12978-022-01422-w.

[265] Thomas L. Webb and Paschal Sheeran. "Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence". In: *Psychological Bulletin* 132.2 (2006), pp. 249–268. DOI: 10.1037/0033-2909.132.2.249.

[266] Robert West. "Time for a change: putting the Transtheoretical (Stages of Change) Model to rest". In: *Addiction* 100.8 (2005), pp. 1036–1039. DOI: 10.1111/j.1360-0443.2005.01139.x.

[267] Kristina Wilson et al. "When It Comes to Lifestyle Recommendations, More is Sometimes Less: A Meta-Analysis of Theoretical Assumptions Underlying the Effectiveness of Interventions Promoting Multiple Behavior Domain Change". In: *Psychological bulletin* 141.2 (2015), pp. 474–509. DOI: 10.1037/a0038295.

[268] Ruth Q. Wolever et al. "A Systematic Review of the Literature on Health and Wellness Coaching: Defining a Key Behavioral Intervention in Healthcare". In: *Global Advances in Health and Medicine* 2.4 (2013), pp. 38–57. DOI: 10.7453/gahmj.2013.042. PubMed ID: 24416684.

[269] Wendy Wood and David T. Neal. "A new look at habits and the habit-goal interface". In: *Psychological Review* 114.4 (2007), pp. 843–863. DOI: 10.1037/0033-295X.114.4.843.

[270] Asma Yamani, Malak Baslyman, and Moataz Ahmed. “Are We Aligned? A Preliminary Investigation of the Alignment of Responsible AI Values Between LLMs and Human Judgment”. In: *2nd IEEE/ACM International Conference on AI-powered Software (AIware)*. IEEE Computer Society, 2025, pp. 133–142. ISBN: 979-8-3315-8269-2. DOI: 10.1109/AIware69974.2025.00022.

[271] Robert K. Yin. *Case study research and applications: design and methods*. Sixth edition. Los Angeles: SAGE, 2018. ISBN: 978-1-5063-3616-9.

[272] Laura Zaikauskaitė et al. “The Theory of Planned Behaviour doesn’t reveal ’attitude-behaviour’ gap? Contrasting the effects of moral norms vs. idealism and relativism in predicting pro-environmental behaviours”. In: *PLOS ONE* 18.11 (2023), e0290818. DOI: 10.1371/journal.pone.0290818.

[273] Youyang Zhang. “Artificial Intelligence (AI) and Virtual Companionship: A Psychological Review and Future Directions”. In: *Academic Journal of Science and Technology* 18.2 (2025), pp. 248–255. DOI: 10.54097/qmkntc12.

[274] Michelle Zhao, Reid Simmons, and Henny Admoni. “The Role of Adaptation in Collective Human–AI Teaming”. In: *Topics in Cognitive Science* 17.2 (2025), pp. 291–323. DOI: 10.1111/tops.12633.

[275] Tan Zhi-Xuan et al. “Beyond Preferences in AI Alignment”. In: *Philosophical Studies* 182.7 (2025), pp. 1813–1863. DOI: 10.1007/s11098-024-02249-w.