Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

# Scrutability of Proactive Intelligent Personal Assistants

Jovan Jeromela

Supervisor: Prof. Owen Conlan

15 March 2026

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

# Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I agree that the Trinity College Library may lend or copy this dissertation on request. I further agree for this dissertation to be made publicly available on the University's open access institutional repository, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: _____ Date: _____

# Acknowledgements

Foremost, I am deeply grateful to Prof. Owen Conlan for his meritorious mentorship. It was a privilege to work and grow as a researcher and a person under Owen's wise and empathetic guidance. I thank Prof. Eelco Herder and Prof. Gaye Stephens for taking the time to assess the dissertation and for the enriching viva voce collegially moderated by Prof. Siobhán Clarke. My gratitude also goes to all the collaborators I've had during the past years, particularly Jasmina Gajcin, Matt Murtagh White, Dipto Barman, and Alok Debnath. The feedback and advice from Professors Ivana Dusparic, Benjamin R. Cowan, Declan O'Sullivan, Dave Lewis, Vincent Wade, Judy Kay, Judith Masthoff, Robin Burke, and other kind researchers I met have significantly improved my work. I am thankful to the Interactive AI Group at IBM Dublin for a very insightful research internship. Thanks to Janet Choi, Clare Lee, and all the administrative staff for their effort and support. I extend warmest well-wishes to lab colleagues, especially Anaïs Claire Murat, Rameez Qureshi, Monique Munarini, Harshvardhan J. Pandit, and Victoria Wiegand, as well as the righteous activists of the Postgraduate Workers' Organisation, foremost among them Jeffrey Sardina and Claudia Peroni. PhD research is scientific work, and it is wrong to mislabel it as a study under full-time instruction to bar employee protections.

Ich bedanke mich recht herzlich bei allen Vortragenden an der TU Wien, die mich auf diese wissenschaftliche Herausforderung vorbereiteten. Großes Dankeschön vor allem meinen Betreuern: Prof. Peter Knees, Prof. Gernot Salzer und Prof. Christian Fermüller. Ein weiterer Dank gilt allen lieben Kollegen für die Zusammenarbeit und Freundschaft, besonders David Gradinariu, Nemanja Stefanović, Ivana Franović, Aleksandar Pavlović und Michael B. Aigner. Ich bin Marko Sirka und Özgür Durmuş dankbar für die Unterstützung meiner ersten Karriereschritte.

Хвала драгој нани, Кати Марјановић, за сву љубав и пажњу које чине основу мог живота. Захвалан сам и својим родитељима, Ани и Драгољубу, као и брату Лазару Јеромели, зато што су увек били ту да пруже сигурност и охрабрење. Срђану и Весни Марјановић те Душку и Тини Јеромела, Амрики Мудрић и проф. Милошу Марјановићу благодарим на породичној подршци. Захваљујем и новосадским просветарима – пре свега проф. Милошу Рацковићу, Александру Турукалу, Татјани Вукадиновић, Гордани Делић, Марији Васић, проф. Ивици Бошњаку, Лидији Лалевић, Љубици Хичман и Снежани Мандић – на педагошком доприносу и значајном охрабрењу. Захваљујем Јасмини Ђурић и Маријану Матијевићу на вишегодишњој бризи. И, наравно, хвала Марку Продановићу, Александру Ивковићу, Селени Пепић, Стефану Вељи, Владимиру Винцану, Зорки Ћировић, Игору Москаљу, Александри Пајић, Дајани Крстић и свим другим пријатељима који су били уз мене. Још једном хвала Јасмини Гајчин, за трајан позитиван допринос мом животу.

It takes a village to write a dissertation.

*This dissertation has emanated from research conducted with the financial support of Research Ireland Centre for Research Training in Artificial Intelligence under Grant number 18/CRT/6223. This work was supported by the ADAPT Centre, funded by Research Ireland Research Centres Programme (Grant 13/RC/2106_P2).*

# Abstract

In just over a decade, Intelligent Personal Assistants (IPAs) have become ubiquitous, with Google Assistant, Apple Siri, and Amazon Alexa now integral components of smartphones, smart speakers, dashboards, and other internet-connected devices. More recently, advances in artificial intelligence have catalysed a transformation of IPAs from mainly command-driven programs assisting the user with discrete, straightforward tasks to systems that attempt to parse the user's intent, act proactively, and delegate in communications with others.

Evolving capabilities amplified concerns regarding trust calibration and privacy, as IPAs remained *inscrutable*, by not facilitating active user investigation of the underlying models and control of the resulting personalisation. The concept of scrutability was initially defined by Judy Kay [1999] with its importance widely recognised in user modelling research. Still, the concept is rooted in legacy systems, such as hypertext editors and single-domain recommender systems. This dissertation explores how the principles of scrutability may be extended to facilitate effort-efficient user investigation and control of models powering IPAs that operate proactively, provide personalised advice, and interact with others on the user's behalf.

Adopting a human-centric approach, we reanalysed data from studies on the causes of non-use of IPAs and expectations of future IPAs to identify user-desired interaction principles and use cases. Following the reanalyses and a literature review, we focused our further exploration on time management assistance, a complex and paradigmatic yet specific problem space. We then organised a multidisciplinary expert discussion to outline a pertinent IPA interaction framework, related modelling and interaction design challenges, and potential new scrutability principles. The insights from the expert discussion were then embedded in Scrutable Intelligent Personal Assistant for Time Management (SIPA4TM), our scrutable-by-design calendaring assistant that served as the basis for our user study with thinking-aloud interactions and focus group discussions, which has led to the final set of scrutability principles and design guidelines.

The first major contribution of the dissertation is a human-assistant interaction paradigm that links scrutability to user trust through *prospect calibration*, extending the meaning of scrutability to include supporting users in understanding and controlling how IPA features may affect their self-concept, social relations, and required effort. Second, we contribute a set of principles and design guidelines that operationalise scrutability for proactive IPAs as an integrated system-level capacity, rather than a set of model-level architectural requirements. The first minor contribution is the design and implementation of SIPA4TM, developed to enable potential expansion and reuse. Finally, in this dissertation, we contribute a consolidated landscape of user and expert desiderata and concerns regarding personalised, proactive, and delegative time management assistants.

# Contents


**1 Introduction** **1**

1.1 Research Question . . . 3

1.1.1 Research Objectives . . . 4

1.2 Research Methodology . . . 4

1.3 Research Contributions . . . 7

1.3.1 Prospects as a Paradigm for Understanding Interaction-Shaping User Expectations of IPAs . . . 7

1.3.2 Principles of Prospects-Based Scrutable Proactive IPAs . . . 8

1.3.3 Scrutable Intelligent Personal Assistant for Time Management . . . 9

1.3.4 User Attitudes and Desires from Proactive IPAs for Time Management 9

1.4 Dissertation Overview . . . 9

**2 Background and Related Work** **13**

2.1 Proactive Intelligent Personal Assistants . . . 14

2.2 Time Management Assistance . . . 16

2.3 Scrutability . . . 17

2.4 Foundation Models . . . 20

2.5 Agentic Systems . . . 21

2.6 Chapter Summary . . . 24

**3 Devising a Desired Human-Assistant Interaction Framework** **25**

3.1 Reanalysing Raw Data from Prior Studies . . . 25

3.1.1 Infrequent Users' Experiences of Intelligent Personal Assistants . . . 26

3.1.2 Users' Envisioned Dialogues with Perfect Voice Assistant . . . 28

3.1.3 Reanalysis Method . . . 30

3.1.4 Key Findings of the Reanalysis . . . 30

3.1.5 Envisioned Scrutable Interaction Diagram and User Modelling . . . 32

3.2 Expert Discussion on Proactive Time Management Assistant . . . 34

3.2.1 Motivation . . . 34
3.2.2 Discussion Participants . . . 34
3.2.3 Discussion Structure . . . 35
3.2.4 Data Processing and Analysis . . . 35
3.2.5 Thematic Analysis . . . 36
3.2.6 Proposed Interaction Diagram . . . 41
3.2.7 Scrutability and Explainability . . . 42
3.2.8 Stages . . . 43
3.2.9 Summary of the Findings . . . 45
3.3 Devising the Desired Design . . . 46
3.3.1 Interaction Challenges . . . 46
3.3.2 User, Context, and Risk Modelling . . . 49
3.4 Chapter Summary . . . 51

**4 SIPA4TM: Scrutable Intelligent Personal Assistant for Time Management 53**
4.1 Scope of the Assistant . . . 53
4.2 SIPA4TM Implementation Process . . . 55
4.2.1 Implementation Architecture . . . 62
4.2.2 Maintainability and Extendibility . . . 69
4.2.3 Takeaways from the Implementation . . . 70
4.3 User Study Protocol . . . 75
4.3.1 Phases of the User Study . . . 75
4.3.2 Chosen Tasks . . . 78
4.3.3 Initial Events and User Model Values . . . 80
4.4 Ethical Considerations and Approval . . . 81
4.5 Chapter Summary . . . 82

**5 User Study and Evaluation 85**
5.1 Task-Based User Study . . . 85
5.1.1 Research Collaborators . . . 86
5.1.2 Participant Selection . . . 87
5.2 Participant Demographics and Surveyed Attitudes . . . 87
5.3 Verbal Protocol Analysis . . . 91
5.3.1 Interaction Tasks Summary . . . 92
5.3.2 Evaluation Approach . . . 93
5.3.3 Overview of Selected Statements . . . 94
5.4 Logs-Based Quantitative Grounding . . . 103
5.4.1 Outliers . . . 106

5.4.2 Statistical Significance Testing . . . 106
5.5 Thematic Analysis . . . 109
5.5.1 Prospects . . . 112
Time and Effort . . . 113
Self-Concept . . . 115
Social Standing . . . 117
5.5.2 Human-Assistant Interaction . . . 118
Capability . . . 121
Data Use . . . 123
Scrutability and Control . . . 128
5.5.3 Trust . . . 136
Decision to use Agentic Features . . . 138
5.6 Desired Time Management . . . 141
5.6.1 Action-Ready Advice . . . 141
5.6.2 Personalised Assistance . . . 142
5.6.3 Non-Disruptive Delegation . . . 143
5.7 Chapter Summary . . . 145

**6 Designing for Scrutable Interaction 147**
6.1 Prospects-Based Framing of Desired IPAs . . . 147
6.1.1 Situating Prospects . . . 148
6.1.2 Properties . . . 158
6.1.3 Related Concepts in Human-Computer Interaction . . . 163
6.2 Principles of Scrutability . . . 166
6.2.1 System Level . . . 168
6.2.2 Prospects-Grounded Design . . . 169
6.2.3 Non-Prescriptiveness . . . 170
6.2.4 Integrated Overlays . . . 172
6.2.5 Uncertainty Signalling . . . 174
6.2.6 Cross-Model Transferability . . . 174
6.3 Roles of Scrutability . . . 175
6.3.1 Proactive Support . . . 175
6.3.2 Meaningful Constraints . . . 180
6.3.3 Prospects Alignment . . . 183
6.4 Chapter Summary . . . 186

**7 Discussion 189**

7.1 Scrutability of Proactive IPAs: From a Technology-Prescribing Goal to a Prospects-Based Capacity . . . 189
7.1.1 Problems Resolved through Scrutability . . . 193
7.1.2 Definition of Scrutable Proactive IPAs . . . 198
7.2 Agentic Interpretability . . . 200
7.2.1 Iterative Dialogue . . . 200
7.2.2 Proactivity . . . 201
7.2.3 Mental Models and User Models . . . 202
7.2.4 Feasibility . . . 203
7.2.5 The Role and the Desired Outcomes . . . 205
7.3 Limitations and Future Work . . . 206
7.3.1 SIPA4TM, Voice Assistants, and LLM-Based Agents . . . 206
7.3.2 Qualitative Approach . . . 207
7.3.3 Axes of Prospects . . . 208
7.3.4 All or Nothing? . . . 210
7.4 Chapter Summary . . . 211

**8 Conclusion** **213**
8.1 Fulfilment of Research Question and Objectives . . . 213
8.1.1 Research Objective 1: Literature Review . . . 214
8.1.2 Research Objective 2: User and Expert Expectations . . . 215
8.1.3 Research Objective 3: Interaction Paradigm . . . 215
8.1.4 Research Objective 4: Scrutable-by-Design IPA . . . 217
8.1.5 Research Objective 5: Principles of Scrutability . . . 218
8.2 Research Contributions . . . 221
8.2.1 Publication Output . . . 222
8.3 Limitations and Future Work . . . 224
8.4 Final Remarks . . . 224

**A1 Links to Further Materials** **257**

**A2 Reanalysis of Prior Studies** **259**

**A3 Expert Discussion** **263**
A3.1 Moderator's Presentation . . . 263
A3.1.1 Prologue . . . 263
A3.1.2 Discussion Preparation [25m] . . . 265
A3.1.3 Discussion [65m] . . . 265
A3.1.4 Outlook . . . 266

A3.2 Use Cases and Scenarios . . . 267

**A4 SIPA4TM-Based User Study 269**

A4.1 Informed Consent Form . . . 269
A4.2 User Study Resources . . . 274
A4.2.1 Week at a Glance . . . 274
A4.2.2 User Study Tasks . . . 276
A4.2.3 Locations, Priorities, and Commute Times . . . 276
A4.3 User Study Protocol . . . 277
A4.3.1 First Phase . . . 277
A4.3.2 Second Phase . . . 278
A4.3.3 Third Phase . . . 282
A4.3.4 Additional Information . . . 285
A4.4 Thinking-Aloud Themes and Codes . . . 285

**A5 Publications 293**

**A6 Other Academic Activities 301**

# List of Figures


3.1 Diagram of the envisioned IPA interaction design, as published in [Jeromela, 2022] . . . 33
3.2 An illustration of the outputs of the expert discussion and the method taken . 36
3.3 The four extracted themes and their subtopics; scrutability is seen as an overarching meta-theme of the discussion . . . 37
3.4 Diagram of the envisioned proactive time management assistant's internal design and its interaction with the user, arising from the expert discussion. The ellipses showcase data modelled and diamond actions performed by the assistant. 41
3.5 The proposed onboarding and control in stages: the user may "unlock" additional functionalities by permitting more data access and IPA autonomy. The experts envisioned stages as concentric and linear, yet they might conceivably "branch out", and the user might choose not to utilise one of the "inner" stage's functionalities. . . . 44
3.6 An overview of all findings from the expert discussion . . . 46

4.1 A coloured diagram from the expert discussion described in Subsection 3.4; yellow elements are researched through corresponding features of SIPA4TM; the grey elements are out of scope of implementation and are addressed through questions in the participant discussion protocol . . . 54
4.2 A screenshot of SIPA4TM with its three main components visible: the calendar view on the left-hand side; the assistant component in the upper-right corner; and the control and context component in the lower-right corner . . . 57
4.3 Three screenshots assembled to showcase the three levels of explanations: the first level is a short text-based explanation, the second is a visualisation with more details, and the final level exposes user model fields relevant for the scrutinised suggestion. The explanations are shown in the assistant component, and the user controls the scrutiny and may enter the user model to change the underlying values through the context component underneath the assistant component. . . . 61

4.4 The view of the permissions control, displayed in the assistant component for the three stages, visualised by stars and described through a high-level description 61
4.5 The assistant's contextual user model grows with the selected permissions control stage: for the lowest stage (★☆☆), user preferences are not saved or considered, at the middle stage (★★☆) the assistant only tracks the user preferred sleep pattern and commute times (the latter is not displayed), and at the highest stage (★★★) the assistant also tracks durations for different event categories . . . . . 62
4.6 System architecture of SIPA4TM, utilising Vue 3 with Pinia frontend, Flask backend with scrutability pipeline and rule-based assistant modules, and an SQLite database storing user model, suggestions, and events. . . . . 63

5.1 Aggregated results of prior experience with e-calendars and AI tools . . . . . 90
5.2 Aggregated attitudes on the interest in AI, the importance of explanations, and importance of privacy . . . . . 92
5.3 JT test p values before and after Holm-corrections above respective box plot visualisations of totals sums grouped by the four survey questions on the interest in AI, importance of privacy, importance of explanations, and the number of data access requests sent . . . . . 108
5.4 An outline of our thematic analysis approach for the focus group discussions; returning arrows indicate iterative improvement of previous steps . . . . . 110
5.5 The preliminary themes, after the conclusion of step 3 of the thematic analysis 112
5.6 The final themes, after the conclusion of steps 4 and 5 of the thematic analysis 112

6.1 Radar chart illustrating how user's trust towards the IPA (quantified as the area of the inner triangle) emerges from users' perceived prospects across three interrelated axes: anticipated time and effort, self-concept actualisation, and expected social consequences; concentric rings indicate perceived harm (negative values down to $-1$), ambivalence (0) and likely benefit (positive values up to 1) for each axis . . . . . 148
6.2 A value function, as described by Tversky and Kahneman [1992], which is concave for gains above the reference point (here, but not in general, at 0), convex for losses (here, $\alpha = \beta = 0.88$), and steeper for losses than gains (here by factor $\lambda = 2.25$) . . . . . 159
6.3 A conceptual diagram of the proposed principles . . . . . 173

# List of Tables


3.1 The names (and descriptive names, as in [Jeromela and Conlan, 2023b]) of the four identified themes and the main points contained within them . . . . . 38

4.1 Design challenges mapped to implementation and study protocol . . . . . 56
4.2 Scrutiny Levels and Explanation Types for Each Suggestion Category in SIPA4TM 58
4.3 Descriptions and proposed actions for all SIPA4TM suggestion types in the final implementation . . . . . 64
4.4 Summary of software design patterns used in the implementation of SIPA4TM 69
4.5 Tasks for the second part of the user study (interaction with SIPA4TM) . . . . 79

5.1 Aggregate demographic profile of the study participants ($N = 22$) . . . . . 88
5.2 Demographic characteristics of individual participants . . . . . 89
5.3 Individual results for AI tool usage (eCal – eCalendar, Oth. – Other AI tools), having a human personal assistant (H.PA), attitudes (AI – interest in AI, Priv. - Importance of Privacy, Expl. - Importance of Explanations), and whether they have ever requested their data (R.D.) for each participant (Part.) . . . . 91
5.4 Overview of Thinking Aloud Findings: themes, statement counts (SC), and participant coverage . . . . . 95
5.5 Logs quantification: Opened suggestion explanation levels, stage explanations, user model views, and percentage of actions done under each stage . . . . . 104
5.6 Expressed usage of LLMs, Voice Assistants, and eCalendars services . . . . . 113
5.7 Preferences for interaction modality across focus groups and participants . . . 119
5.8 Numbers of groups, codes, and participants and lists of participants coded in each point of the Data Use theme . . . . . 124
5.9 Numbers of groups, codes (refs), and participants and lists of participants coded for each stance on having the IPA delegate for them . . . . . 138

6.1 Overview of principles that served as the starting point and our proposed principles for proactive assistants . . . . . 168

6.2 Contrasting overview of roles and their main aspects or design guidelines, for scrutable user models and for scrutable IPAs . . . . . 176

A2.1 Selected statements from the study presented in Cowan et al. [2017]. The ID follows the example from the paper and consists of the focus group and participant number, as shown in presented transcripts. . . . . 259
A2.2 Selected (parts of) dialogues from the study presented in Völkel et al. [2021]. The ID corresponds to the name of the scenario followed by the row number (0–204) of the corresponding participant in the provided dataset. Text in *italics* are the IPAs' statements. . . . . 260

A4.1 Distance/Time Matrix between Locations (in minutes) . . . . . 277
A4.2 Interaction Findings Codes: themes, subthemes, statements of interest, statement count (SC), and participant count (PC) . . . . . 286

# 1 Introduction

Intelligent Personal Assistants (IPAs) are computer systems capable of assisting users in tasks varying from controlling the devices they are installed on, question-answering and providing requested information, to supporting various skill development [Islas-Cota et al., 2022, de Barcelos Silva et al., 2020]. Apple's introduction of Siri as a standard feature of iPhones in 2011, followed by development of competing systems including Google Assistant and Amazon Alexa, made IPAs practically omnipresent [Luria et al., 2020, de Barcelos Silva et al., 2020]. These systems are now routinely integrated in personal computers, smartphones, smartwatches, smart TVs, and smart speakers [Porcheron et al., 2018, Clark et al., 2019b, Islas-Cota et al., 2022]. Such general-purpose assistants typically facilitate services including retrieving information, controlling internet-of-things devices, setting alarms, saving reminders, and creating calendar events [Këpuska and Bohouta, 2018, Porcheron et al., 2018]. These systems are offered both as primary interaction mechanisms of ad-hoc devices (e.g. smart speakers) and features of more universal products (like smartphones). Smaller companies and purpose-assembled teams have also developed their own assistive systems, often focusing on particular goals (e.g. creating an IPA that is fully open-source [Campagna et al., 2017] or requires only limited computing power [Vashistha et al., 2019]) or assisting specific use cases and target groups (e.g. supporting the elderly [Costa et al., 2018], aiding language learners [Pham et al., 2018].

Other terms used to denote these assistants in the literature include: virtual assistants, digital assistants, smart assistants, conversational assistants, and (when that is a supported interaction modality) voice assistants [Chung et al., 2017, Tulshan and Dhage, 2018]. Similarly, instead of – or in addition to – the word assistant, the concepts of agent and voice interface are sometimes used [Chung et al., 2017, de Barcelos Silva et al., 2020, Sarikaya, 2017]. In our approach, we adopted the term *Intelligent Personal Assistant* because it emphasises that the primary role of the system is to provide assistance to the user in achieving their explicit or implicit goals by providing personalised services.

A complex, yet paradigmatic use case for specialised IPAs is time management. Indicatively, CALO, which later spun out into Siri, was developed for time management of knowledge

workers [Berry et al., 2017, Yorke-Smith et al., 2012]. More recently, Calendar.HELP by Cranshaw et al. [2017] enabled meeting negotiation delegation capability via email. In the last few years, advances in Artificial Intelligence (AI) furthered capabilities of chatbots to include providing personalised advice to users, blurring the line between previously mostly separate research domains of chatbot interaction design and human-assistant interaction [Bommasani et al., 2022, Bin Akhtar, 2024]. Most recently, the so-called AI agents, that is, increasingly autonomous and proactive systems capable of acting on the user's behalf, further thinned the line with what users envisioned as "perfect" IPAs of the future [Völkel et al., 2021] and what the emerging technology promises to achieve [Brachman et al., 2025, Wu et al., 2024].

Yet, proactivity of commercial assistants remains rather limited and realised primarily in the form of various smart-home routines or notifications with information deemed relevant for the user at the moment [Sun et al., 2016, 2017, Zargham et al., 2022]. These contemporary limitations are in sharp contrast with user expectations. Namely, prior work [de Barcelos Silva et al., 2020, Islas-Cota et al., 2022, Sun et al., 2016, Luria et al., 2020] foretells that future IPAs are to be *proactive* and more *personalised*, which would enable the assistants to predict and perform actions most appropriate for the domain, the user's context, and the user's intent. Similarly, the participants in a survey on the expectations of a perfect future assistant by Völkel et al. [2021] anticipated the system would need to be readily aware of changes in the surrounding.

These foreseen advanced assistants would ineluctably rely on complex and dynamic models, as they would necessitate representing not only the user but also their social relationships, mental and physical state, and the relevant broader context. Yet, even with comparatively less complex IPAs, the lack of trust has been recognised as one of the main reasons for not using virtual assistants [Luger and Sellen, 2016, Cowan et al., 2017]. We can thus assume that human-assistant interaction in the future would include communicating how the IPA understood the user's desiderata and how it can minimise the likelihood of negative effects for the user and other humans the IPA interacts with on the user's behalf.

Therefore, to make the anticipated more feature-rich proactive IPAs enticing to users, it is imperative to devise a mechanism through which users could understand what information the system collects and how it is then processed to provide befitting personalised services. Our guiding premise is that a manner to achieve this is by making the intelligent personal assistants *scrutable*. Scrutability is a term representing the ability of a user to understand, interact with, adapt, and study or *scrutinise* a personalised model [Kay and Kummerfeld, 2013]. This concept was introduced by Judy Kay [1999] in her doctoral dissertation. The word was chosen over related concepts in literature, such as *open* or *transparent*, because it emphasises the need for the user to actively engage with the system and the embedded user

model [Kay and Kummerfeld, 2013].

The leitmotiv of this dissertation is to investigate the confluence of the research fields of IPAs and scrutability, that is, to identify and address challenges arising when considering the principles and design guidelines of scrutable personalisations with the desired capabilities of proactive IPAs. Moreover, we investigate how principles of scrutability can be adapted to support human-assistant interaction, leading to a greater sense of rapport and trust than contemporary opaque, task-driven IPAs. In particular, we bound our investigation to IPAs for time management – a desired feature [Völkel et al., 2021] that may assist the user in better utilising their time [Cranshaw et al., 2017] and in organising collaborations [Toxtli et al., 2018].

## 1.1 Research Question

The research question reads as follows:

*How can principles of scrutability be extended to facilitate effort-efficient control of proactive intelligent personal assistants for time management?*

The research question explicitly assumes that IPAs in the future would be *proactive*, which means that they would anticipate the user's intent and suggest or perform appropriate time management actions without prior user prompts.

Such assistants would likely rely on complex and dynamic models. The central aspiration is to investigate how such models could be *scrutable*, allowing the user to actively study and understand how these models enable and shape personalisation.

Therefore, by *principles of scrutability*, we understand system design choices and interaction design considerations that must be met for the IPA's models to be considered scrutable.

We hold that there are both nuances and significant differences that have to be considered when deliberating about extending principles of scrutability to IPAs. These differences stem from questions concerning *proactivity* (When/why to initiate interaction? How do the user model and the context relate to the interruption design?), *delegation* (When/how to delegate for the user?), and the *broader effect* (How do the user model and the assistant's actions affect others?). Therefore, we see the need for the principles of scrutability already addressed by Kay and Kummerfeld [2013] to be *extended to facilitate* these novel challenges.

The envisioned IPA would have to reason about the provenance of modelled information and its applicability in similar situations. Namely, the assistant must possess a confidence model that balances the trade-off between potential social and financial costs of undesired

actions caused by a misunderstanding on the one hand and distracting or frustrating the user by asking additional questions on the other. Consequently, investigating the *effort-efficient control* of scrutable IPAs and its effect on the cognitive load would be a further point of investigation.

Similarly, with the increasing complexity and capabilities of both the primary task-solving capabilities of the assistants as well as their communication prowess, the importance of calibrating trust increases. Scrutiny may thus offer a mechanism for the user to better align their mental model of the assistant's capabilities, while the control would allow them to tune the assistant's user model towards more apt assistance.

Rather than focusing on general-purpose IPAs, we narrow our investigations of scrutability to *time management* assistance. On the one hand, this bounds the features that ought to be implemented and limits the relevant user and situational context to be modelled. On the other hand, in designing our assistant for the main user study, we can rely on the rich literature on behavioural change support and calendaring.

### 1.1.1 Research Objectives

To provide an answer to our research question, we aspire to fulfil the five research objectives (RO) enumerated below.

RO1 Undertake a literature review of state of the art in Intelligent Personal Assistants (IPAs), human-assistant interaction, time management assistance, and scrutability.

RO2 Understand user and expert expectations regarding future capabilities of IPAs, especially as they relate to proactivity, time management, and scrutability.

RO3 Formulate an interaction paradigm for proactive IPAs for time management that enables users to scrutinise the system's underlying models and accurately assess its capabilities and assumptions.

RO4 Model, develop, and evaluate a scrutable-by-design IPA for time management based on the defined interaction paradigm.

RO5 Distil the findings from the IPA interaction evaluations into a set of principles and design guidelines for scrutable proactive IPAs.

## 1.2 Research Methodology

To tackle the challenges formulated in RO1, RO2, RO3 and RO4, our work followed the Design-Based Research (DBR) approach [Easterday et al., 2014]. This research approach had

originally been devised for education research [Anderson and Shattuck, 2012], but has since been applied in the areas of human-centred computing as well [Amyrotos, 2021, for instance]. As explained by Easterday et al. [2014], DBR foresees an iterative process of *focusing* on defining the problem, *understanding* the existing solutions, *defining* goals and assessments, *conceiving* a sketch of a solution (which in our case is a set of principle-candidates, interaction diagrams, and iterations of the scrutable-by-design IPA incorporating them), and *building* or implementing the solution (the final set of scrutability principles and design guidelines). Returning to previous phases when that is meaningful due to the knowledge gained in later phases is integral to the DBR, which is very advantageous given the partially unpredictable nature of interventions in human-centred research [Amyrotos, 2021]. Thereby, our initial literature review and associated understanding of user and expert expectations guided the preliminary interaction paradigm, which informed the scrutable IPA development. As the interaction paradigm and the scrutable IPA are evaluated through an expert discussion and a user study, our understanding of the expectations of IPAs and relevant literature evolved, ultimately leading to further refinement of the proposed interaction paradigm and scrutability principles.

A broad literature review was undertaken to address Research Objective 1 (**RO1**). The main findings can be seen in Chapter 2, with further literature discussed through the dissertation, in line with the DBR approach.

To get an initial insight into user and expert expectations regarding future capabilities of IPAs and address the **RO2**, we first reanalysed raw data from studies by Cowan et al. [2017] and Völkel et al. [2021]. In doing so, we focused on user expectations regarding the proactivity, personalisation, and delegative capabilities of intelligent personal assistants. The results of these reanalyses can be seen in Section 3.1. Section 3.2 reports on the results of this discussion. Based on insights from the work towards fulfilling RO1 and RO2 and the reanalysis, our first IPA interaction and modelling diagrams have been created in partial fulfilment of **RO3**. This diagram can be seen in Figure 3.1 and is discussed in Subsection 3.1.5. The reanalysis of the two prior user studies has also expanded our understanding of user expectations regarding future proactive IPAs and guided us to narrow the application domain of our future research to the domain of time management assistance.

We then devised and facilitated a multidisciplinary expert discussion to understand how they envision the role, design, and constraints of proactive time management assistants. The structure of the expert discussion was shaped by the prior study by Meurisch et al. [2017], with methodological adaptations to elicit discussion on the role of scrutability within the domain of time management. A thematic analysis of the results ensued, aiming to contribute towards **RO2** and **RO3**. Our analysis of the discussion transcripts has revealed a range of

opinions on appropriate use cases and usability concerns, as well as an emphasis on appropriate data stewardship and the disclosure of capabilities. Further results are discussed in Section 3.2, and the ensuing diagram of the envisioned IPAs and their interaction design is shown in Figure 3.4.

The work thus far was then concretised by implementing Scrutable Intelligent Personal Assistant for Time Management (SIPA4TM), our scrutable-by-design calendar assistant, which objectifies the fulfilment of our **RO4**. This iterative development process is described in Chapter 4, which also describes how our findings to that point and previously published guidelines, including by Kay and Kummerfeld [2019] and Shneiderman [2022], have influenced our design decisions. SIPA4TM was implemented to log user interactions and thus provide a quantitative record of actual user scrutiny. SIPA4TM was also implemented with software engineering design principles in mind to ensure extensibility and maintainability, thereby offering a framework for future studies.

A three-stage task-based study with 22 users based on SIPA4TM was then facilitated, representing the main source of insight regarding user experience and expectations on scrutable proactive assistance. The data collected during the user study was used for a mixed-method analysis of user interest in scrutinising the system's models while fulfilling tasks regarding the next week's schedule. SIPA4TM offered advice for time management at both the daily and weekly scale, and the participants could scrutinise any advice as well as inspect and control the underlying user model powering the personalisation. Finally, participants discussed their opinions on SIPA4TM as well as IPAs, chatbots, proactivity, privacy, and the role of explanations more broadly. The findings from the focus group discussions and the thinking-aloud interaction protocol were then qualitatively analysed, while the interaction logs were reported and quantitatively analysed in Chapter 5. The results of these analyses were then synthesised and contrasted with expert discussion-stemming findings to constitute the final elements of **RO2** in Chapter 6.

The same chapter also contains a synthesis of the results into a novel paradigm of prospects of IPA use. This paradigm explains a crucial finding of our user study that scrutability and user control are principally only desired when the users anticipate they would lead to a decrease in time and effort needed for their future task fulfilment, actualisation of their self-concept [Burns, 1980], or an improvement in their social standing. The term *prospect* is borrowed from behavioural economics, and in Chapter 6 we propose evidence that users' deliberation on prospects along the three axes – time and effort, self-concept, and social standing – follows the properties originally devised by Tversky and Kahneman [1992] for the wealth value function in Prospect Theory. This theoretical framework that bridges user interaction and the feeling of trust towards the IPA while situating scrutability, represents the final fulfilment of **RO3**.

Ultimately, all findings from the first four research objectives are distilled into a set of principles of scrutability and design principles fulfilling **RO5**. As we elaborate in Chapter 6, the principles are based on the strong indication from our user study that user expectations of IPAs have evolved along advancing technology that is based primarily on opaque underlying algorithms that do not admit foundation-level architectural requirements prescribed by Kay and Kummerfeld [2013]. Moreover, our analysis indicates that the desire for scrutability is situated in prospects rather than existing per se. With that in mind, our scrutability principles foresee realisation of scrutability at the level of the system, not user model. We then revisit each principle by Kay and Kummerfeld [2013] and reframe it with new principles for scrutable proactive IPAs that do not prescribe the underlying model type or realisation, but rather position scrutability as a conditionally desired capacity framed within broader human-centred design. Finally, in Chapter 7 we provide a discussion that argues that this conceptual shift results in a notion of scrutability that still resolves the interaction problems devised by Kay and Kummerfeld [2013], while offering a pathway for a realisation of the concept as a design goal and conceptual framing in the current period of much more capable, yet architecturally opaque models.

## 1.3 Research Contributions

There are two major and two minor research contributions presented in this dissertation.

### 1.3.1 Prospects as a Paradigm for Understanding Interaction-Shaping User Expectations of IPAs

The first major contribution (C1) is a novel IPA interaction paradigm for the envisioned proactive IPAs for time management based on prospects. Through prospects, we formalise a conceptual bridge between user interaction with the IPA and the trust it elicits.

The term *prospect* stems from the Prospect Theory in behavioural economics, originally devised by Kahneman and Tversky [1979]. It emphasises that, when interacting with complex and capable IPAs, users' actions are guided by considerations regarding time and effort, self-concept actualisation (whether using it in that way is in line with their internalised ideal self), and the effect on their social standing. We have found indications that such framing of user prospects in HCI relates to several properties of the value function in Prospect Theory. For instance, prospects are reference-dependent and status quo-anchored, as the results of using an IPA are internally compared with the current way of resolving tasks. Moreover, we have found indications of diminishing sensitivity towards gain in terms of explanation seeking, where users seem to be satisfied with more shallow justifications of IPA's behaviour, with further exploration

being comparatively less attractive. Interestingly, we also found an indication that there is reduced sensitivity to marginal pain, especially in terms of privacy.

This paradigm of prospects suggests that incorporating scrutability into the interaction design process can help users calibrate their trust levels relative to the system's actual task-solving capabilities. Crucially, it also reframes scrutability from a goal per se, to a conditionally desired capability with the broader goal of prospects-alignment, whereby the user's estimated prospects of using (a feature of) the proactive IPA better match the actual capability-based prospects of use. We also connect the paradigm of prospects to concepts that affect user trust and willingness to use, such as propensity, reliability, and affordances.

### 1.3.2 Principles of Prospects-Based Scrutable Proactive IPAs

Following the nomenclature used by Kay and Kummerfeld [2013], we present a set of principles that answer the question of *how* to create scrutable proactive IPAs, constituting our second major contribution (C2). These principles emerged from two core ideas from the analysis of the user studies. The first core idea behind our proposed principles is that scrutability of IPAs should be a *non-prescriptive*, *system-level* capacity, rather than a model-level architectural property. Second, scrutability should be seen as a *prospects-grounded* alignment tool within the broader framework of human-centred AI, rather than a value-bearing goal in itself. Moreover, the principles of *integrated overlays* and *uncertainty signalling* state that the user should be enabled to scrutinise their IPA in the same manner they interact for other purposes and that IPA should indicate to the user the level of uncertainty in its responses. Lastly, with the principle of *cross-model transferability*, we encapsulate the idea that scrutinising an IPA should recover not only the saved data but also broader modelling insights.

We also provide design guidelines that enable the scrutable assistant to act in accordance with the identified principles of prospect-based scrutability and thus to verify its assumptions and elucidate its features and requirements, especially during the onboarding of new users, when *prospects* are most malleable. These guidelines stem from three roles for scrutable IPAs: *prospects alignment*, *offering meaningful constraints*, *proactively supporting* user scrutiny and control. We then argue that following these principles and guidelines still results in an AI system that addresses the problems that Kay and Kummerfeld [2013] identified as a motivation for scrutability. Finally, we contribute an outlook for the role and realisation of both prospects and scrutability in the emerging era of agentic systems powered by foundation models.

### 1.3.3 Scrutable Intelligent Personal Assistant for Time Management

Our first minor contribution (c1) is a Scrutable Intelligent Personal Assistant for Time Management (SIPA4TM), a desktop web application for calendar assistance. SIPA4TM was implemented to embody both the design guidelines for scrutable user models formulated by Kay and Kummerfeld [2013, 2019] and ideas from an expert discussion on anticipated proactive time management assistants that we had previously facilitated and analysed. The assistant facilitates scrutiny of the proactive suggestions at three levels of detail as well as a mode of explicitly controlling the underlying user model. Similarly, as done by Tolmeijer et al. [2021], the primary mode of interaction for the participant is through on-screen buttons, one of which allows them to start a multi-layered iterative process of scrutiny. In Chapter 7, we then offer avenues for potential further development and research based on SIPA4TM, whose architecture was implemented to be extendible and reusable.

### 1.3.4 User Attitudes and Desires from Proactive IPAs for Time Management

Our second minor contribution (c2) is a compiled landscape of user attitudes towards a personalised, proactive, and delegative time management assistant. In recent years, primarily due to advances in natural language processing enabled by foundation models [Bommasani et al., 2022], perceptions of possibilities, concerns, and opportunities in the domain of AI-based assistance have evolved. This dissertation thus offers a timely overview of related attitudes, as expressed by both domain experts and non-expert users. These user-driven findings outline both the desired development directions and the pitfalls to avoid.

## 1.4 Dissertation Overview

This dissertation contains eight chapters and is structured as follows.

Chapter 1 introduced the motivation for exploring the role of scrutability – i.e. the ability of the user to actively investigate and study how an adaptive system is personalised to them – for proactive Intelligent Personal Assistants (IPAs). We then indicated our research question, research objectives, and methodology, thereby emphasising the human-driven approach and qualitative insight-driven method.

Chapter 2 provides a literature review in the domains of proactive intelligent personal assistants, time management assistance, human-assistant interaction, scrutability and

explainability, as well as a brief overview of currently emerging agentic systems and the underlying technology of foundation models.

Chapter 3 presents the results of our reanalysis of user studies by Cowan et al. [2017] and Völkel et al. [2021] and the results of our expert discussion in which we elicited and analysed experts' predictions and concerns on anticipated proactive IPAs for time management. The chapter also includes our diagrams of the envisioned human-IPA interaction based on the reanalysis and the expert discussion. The chapter concludes by summarising envisioned modelling and interaction challenges to guide the development of proactive time management assistants.

Chapter 4 presents the design and implementation of Scrutable Intelligent Personal Assistant for Time Management (SIPA4TM), a desktop application we developed to serve as the focal point of the dissertation's main user study. In the chapter, we also elaborate on practical findings from the iterative implementation process. Finally, the chapter elaborates on the user study protocol for the main user study and describes how SIPA4TM and the protocol were co-designed to provide insight into the modelling and interaction challenges introduced in the previous chapter.

Chapter 5 presents the main user study of the dissertation, in which 22 participants were invited to an in-person survey, followed by a thinking-aloud task-based interaction with SIPA4TM, and concluding with semi-structured focus group discussions. The chapter starts by providing demographic information about the participants and the methodological grounding for our facilitation and analysis approach. The findings and evaluations are also presented in the chapter, including an analysis of logs that indicated a weakly significant relationship between self-reported past GDPR data retrieval invocations and logged assistant scrutiny. A qualitative analysis of a thinking-aloud interaction transcript has revealed a tendency to scrutinise SIPA4TM directly, through probing behaviour, rather than the explicit channels. Finally, an extensive thematic analysis of the focus group discussions has identified three main themes: interaction, trust, and prospects, which connect the previous two themes and comprise what participants expect the outcome of the interaction with SIPA4TM would be. These prospects were identified as relating to time and effort, self-concept actualisation, and the effect on their social standing. The chapter closes with a summative overview of desired time management assistance expressed by the participants, which has revealed a set of highly context-dependent and prospects-sensitive considerations.

Chapter 6 synthesises our findings regarding the prospects into a paradigm for understanding the role of scrutability and broader user interaction in shaping trust towards

the IPAs. We then relate our proposed paradigm to related concepts in human-computer interaction and the properties of the value function in Prospect Theory in behavioural economics, from which the term was taken. We then distil findings from our entire research into a six system-level prospects-based principles for scrutable proactive IPAs. Finally, the chapter introduces a set of design guidelines that serve to steer the development of scrutable IPAs.

Chapter 7 argues that our reframing of scrutability from a user-model-level, top-down design paradigm to a systems-level capacity whose desirability is conditioned on user prospects still resolves the problem presented as the motivation for enabling scrutability in the seminal paper by Kay and Kummerfeld [2013]. We propose and discuss a unified definition of system-level prospect-based scrutability for proactive IPAs, then place it in the context of the emerging concept of "agentic interpretability" by Kim et al. [2025]. The chapter concludes by discussing the limitations of SIPA4TM, our user study approach, and the dissertation as a whole, and by proposing potentially interesting avenues for further research.

Chapter 8 concludes the dissertation by summarising how each of the research objectives was fulfilled, by answering the research question, and by reiterating the contributions of the dissertation to the state of the art. The dissertation ends with a short recapitulation of the limitations and future work, followed by closing remarks.

After each chapter, we provide a brief summary.

# 2 Background and Related Work

This chapter presents related research, i.e. the background literature of this dissertation. Specifically, during the initial phase of our work towards the dissertation, we focused on defining the research scope and refining our research question. For that purpose, an extensive survey of related work has been undertaken and then continuously expanded. Our research question is at the confluence of three research areas: 1) proactive Intelligent Personal Assistants (IPAs) and human-assistant interaction; 2) time management and behavioural change support; and 3) scrutability. Notably, we chose to focus on time management assistance, as it is a complex yet bounded use case of IPAs, with the expectation that it would enable us to investigate paradigmatic user perspectives, while narrowing the research scope from that of general-purpose IPAs[1]. Our review of literature relating to these three areas is presented in sections 2.1, 2.2, and 2.3 below. This background literature has constituted the basis for our reanalysis and the expert discussion (Chapter 3) as well as for the implementation of SIPA4TM, our Scrutable Intelligent Personal Assistant for Time Management (Chapter 4).

As we elaborate in Sections 2.4 and 2.5, during the research work for this dissertation and the implementation of SIPA4TM, capabilities of AI-based systems powered by the so-called foundation models have increased dramatically, significantly shifting user expectations (as evaluated in Chapter 5) and thus necessitating a theoretical reframing of scrutability (as proposed in Chapter 6 and discussed in Chapter 7). Therefore, the last two sections of this chapter provide background to this concurrent process. Nevertheless, it is important to emphasise that this dissertation has not scoped to investigate the algorithmic realisation or interpretability of foundation models or the currently evolving agentic systems.

[1]The research direction was informed in part by the feedback and advice received from many prominent researchers in the field, including Judy Kay, Judith Masthoff, Peter Brusilovsky, Robin Burke, Dave Lewis, Banjamin R. Cowan, Vincent Wade, Peter Knees, Piek Th.J.M. Vossen, Thomas Baier, and Kieran Fraser. In particular, the feedback and further references received during the UMAP '22 Doctoral Consortium [Bellogin et al., 2022, DC track chaired by Giovanni Stilo and Rita Orji] were of paramount importance for choosing time management assistance as a meaningful use case grounding for further research. We thank everyone who has so kindly offered their time and guidance and thus ultimately contributed to this dissertation.

## 2.1 Proactive Intelligent Personal Assistants

*Intelligent Personal Assistants (IPAs)* have become ubiquitous over the past decade and a half, with prominent examples such as Siri, Google Assistant, and Alexa integrated in smartphones, home controllers and desktop computers [Clark et al., 2019b, Islas-Cota et al., 2022]. The designation *intelligent personal assistant* became prevalent in the literature, as it declares that the core of its functionality is personalisation powered by artificial intelligence, as elaborated by Chung et al. [2017]. Other adjectives often used to identify these assistants include virtual, digital, smart, and voice-enabled [Tulshan and Dhage, 2018, de Barcelos Silva et al., 2020, Chung et al., 2017, Sarikaya, 2017].

In their systematic review of IPAs, de Barcelos Silva et al. [2020] state that "future research could deal with the *proactivity and personalization* challenges when developing intelligent personal assistants". Sarikaya [2017] divides IPA support into proactive (anticipatory, system-initiated) and reactive (user-initiated) assistance. Sarikaya [2017] then references the work by Isbell and Pierce [2005] on action responsibility sharing between a user and a proactive adaptive system, to propose that proactive assistance operates on a proactivity continuum ranging from the user having to do the task manually; through having the assistant suggest actions to the user; and to the full autonomy of the assistant to decide and perform actions. To achieve the latter stages on that continuum, the assistant has to model the helpfulness of actions and user desires [Sarikaya, 2017, Schiaffino and Amandi, 2004]. Notably, Shneiderman [2020] argued that a higher degree of automation should not be seen as achievable only at the expense of user control, and proposed a two-dimensional framework for Human-Centred AI. One dimension in the framework is human control and the other is computer automation, and Shneiderman [2020] argues that reliable, safe, and trustworthy AI is high in both of these dimensions. Völkel et al. [2021] studied how users envision a perfect future IPA, and reported that proactivity, personalisation, and smart scheduling were among the desired features, with further findings described in Section 3.1.

One of the most well-known *early proactive assistants* was Clippy, which offered situational advice in Microsoft Office '97 [Whitworth, 2005]. Whitworth [2005] suggested that Clippy ultimately failed (and miserably so, as per Maedche et al. [2016]) because it had not followed the four rules of polite computing: respect user choice; disclose yourself; offer useful choices; and remember past choices. Moreover, as Kay and Kummerfeld [2013] note, Clippy was not scrutable either, as Word users could neither see, correct, nor control the underlying user model. Similarly, Xiao et al. [2003] hypothesised that the usefulness and relevance of the offered proactive assistance might explain why participants in their study reacted positively to their WoZ prototype, while having a unanimous negative opinion of Clippy.

A *new wave of work on proactive assistance* then commenced with the proliferation of agentic systems from around 2022 onwards [Bommasani et al., 2022], which we discuss separately in Section 2.5. A commercial example of a proactive system at the early cusp of this wave is Xiaolce, which was originally developed by Microsoft as an empathetic chatbot on Chinese social media [Zhou et al., 2020]. The system is proactively suggesting reminders for actions deemed relevant, and it may nudge the user to change behaviour estimated to be disruptive.

Similarly, before the most recent proliferation of agentic systems discussed in Section 2.5, there were few works on enabling assistants to directly represent or *delegate* for their users in interactions in natural language with other people [de Barcelos Silva et al., 2020, Islas-Cota et al., 2022]. One notable exception is Google Duplex whose advertised functionality included calling to book appointments at hairdressers' and restaurants [Leviathan and Matias, 2018]. Yet the system was controversial from the launch due to ethical concerns [O'Leary, 2019] and has to date only seen a limited distribution [Mrkšić, 2020]. Relatedly, Zheng et al. [2022] surveyed works in the area of polyadic assistance encompassing human-to-assistant, human-to-human, and human-to-group interactions. Their study shows that polyadic conversational agents influence how people communicate and maintain relationships, and that designing them ethically requires careful attention to privacy and social boundaries, along with improved methods for testing usability and fostering trust.

Even before most recent advances in natural language processing that we discuss in Section 2.4, users of commercial IPAs found it difficult to *accurately estimate the system's actually capabilities*, as reported for both frequent [Luger and Sellen, 2016] and infrequent IPA users [Cowan et al., 2017]. In the latter case, Cowan et al. [2017] also identified the fear of negative social reactions when using voice to interact with an IPA in public as one of the barriers to broader usage. Wanting to address these and other shortcomings of current voice-based IPAs, Viswanathan et al. [2022] designed a WoZ prototype that proactively continues conversations and offers explanations. Their study found that these interaction elements led to a positive interaction experience. Similarly, Kulesza et al. [2013] studied how explanation properties influence end users' mental models. According to their findings, explanation completeness was more important than soundness and has led to participants investing more effort in understanding the agents.

Lastly, a qualitative analysis by Doyle et al. [2019] found that user *perceptions of humanness* are multidimensional and affected by vocal qualities and perceptions of IPA capabilities. Similarly, Kim et al. [2018] reported evidence supporting the hypothesis that speech, embodiment, and social behaviours may increase the user's confidence and trust in IPAs. The relation between cognitive load, on the one hand, and an appropriate multimodal interface, on the other, was investigated through a digital tabletop game by Zimmerer et al. [2022].

## 2.2 Time Management Assistance

The use cases of both contemporary [de Barcelos Silva et al., 2020, Tulshan and Dhage, 2018, for instance] and envisioned future IPAs [Völkel et al., 2021, Meurisch et al., 2020] are very broad. Based on the prior research and feedback received as part of the doctoral consortium at the UMAP '22 conference [Bellogin et al., 2022], we decided to narrow our research to time management assistance. Help in managing time is a desired feature of IPAs [Völkel et al., 2021] that has the potential to save the user's time [Cranshaw et al., 2017] and facilitate collaboration [Toxtli et al., 2018].

The idea of utilising IPAs to help the user organise their time is not new. For instance, scheduling meetings and helping knowledge workers to organise their time was one of the primary motivations for CALO [SRI International, 2020], a framework for a cognitive assistant, whose spin-out became Apple's Siri [Berry et al., 2017]. Yorke-Smith et al. [2012] presented CALO's generic framework for proactive goal generation and deliberation. The main parts of the framework are a task manager (through which the proactivity is realised), a to-do manager that aids the user in tracking their tasks, and a calendar manager, that is also capable of delegating for the user in meeting negotiations via email [SRI International, 2020]. Moreover, Berry et al. [2017] presented how CALO could automate meeting scheduling.The CALO user would enter their guidance and control system restrictions through declarative commands [Morley and Myers, 2004].

A more recent scheduling system is Calendar.help, a micro-services framework built for hybrid intelligence that delegates meeting scheduling negotiations via email. [Cranshaw et al., 2017]. Furthermore, Shakshuki and Hossain [2014] described a meeting scheduler capable of delegating bilateral and multilateral negotiations. Regarding individual time management more broadly, Sezgin et al. [2020] conducted a survey on the use of IPAs in behavioural health interventions. While the results indicated that the IPAs might be a promising aid, the authors noted that works in the area remain limited in number and scope. On the other hand, Ptakauskaite et al. [2018] investigated how stress management apps support reflection and behaviour change and noted that integrated calendar views might help the user to gather insight from their data and to identify when to perform a stress management activity.

The situations in which IPA proactivity was desirable were investigated by Zargham et al. [2022]. In their study, the participants gave the highest median ratings for appropriateness and usefulness to two groups of use cases related to time management ("reminder" and "health risk"). Furthermore, in the aforementioned large online study in which Völkel et al. [2021] asked the participants to imagine conversations with a perfect IPA, the participants frequently expected the assistant to help them manage their time. The participants in their study also

envision that the IPA would help them in multitasking, yet Edwards et al. [2019] uncovered that IPAs such as Alexa can be disruptive towards the primary task of the user.

## 2.3 Scrutability

As first defined by Kay [1999], scrutability denotes the ability of the user to delve into and study or *scrutinise* the system's user model personalised for them. It should also be noted that Tintarev and Masthoff [2022] influentially, yet somewhat divergently and more narrowly defined scrutability as the ability of the user to correct the modelled information. The definition by Tintarev and Masthoff [2022] is common in publications on recommender systems [Wang et al., 2018, He et al., 2015, Balog et al., 2019, among others] likely due to the definition appearing in the widely-referenced Recommender Systems Handbook by Ricci et al. [2022]. Contrastingly, the view of scrutability by Kay and Kummerfeld [2013] is more prominent in other works on adaptive systems [Brusilovsky, 2022], including in particular for personalised systems in education [Conlan, 2005, Khosravi et al., 2022, Czarkowski, 2006]. In this dissertation, we generally follow the broader view of scrutability by Kay and Kummerfeld [2013] until Chapter 6, in which we formalise our own adaptation of the concept for prospective proactive IPAs. Notably, Kay and Kummerfeld [2013, cf. pp. 6, 22] affirm the significance of the ability to correct and *control* the model, but consider scrutability primarily as an *active process* of the user's understanding of how a user model enables an application to work for, and in conjunction with, them.

This active user effort aimed at gaining an understanding of how the system works and how the system's personalisation may be tuned most effectively for their use is what differentiates the term from related but more algorithm-centric concepts [Kay, 2021a]. These related concepts in human-agent systems were surveyed by Rosenfeld and Richardson [2019]. They defined *explainability* as "the ability for the human user to understand the agent's logic". Rosenfeld and Richardson [2019], in reference to Guidotti et al. [2018], also noted that *transparency* refers to systems whose decision-making is easily and directly graspable by the user. In contrast, in their review of explainable artificial intelligence methods, Hanif et al. [2023] note that explainability has varying definitions and promote that by Saeed and Omlin [2023]: "Explainability aims to bridge the gap between the algorithm and human interpretation of the decision-making process. It is capable of enhancing trust by answering the how and why of the system". Hanif et al. [2023] stress that the lack of explanations raises ethical concerns, limits robustness, and ultimately diminishes user trust and restricts deployment in vulnerable domains. Hanif et al. [2023] then offer their overview of terms related to explainable artificial intelligence, thereafter highlighting *interpretability* as, in their view, a more user-centred term than explainability as it emphasises human-intelligibility. Still, Hanif et al. [2023] caution that over-reliance on human

evaluators can lead to systems that are persuasive rather than understandable.

In this dissertation, we join Khosravi et al. [2022] in observing the two focal differences between scrutability and explainability. The first is that the term scrutability underlines the aforementioned requirement for the user to invest their time and effort in scrutinising the system for the answers they seek. The second is the acknowledgement that, for contemporary complex systems, the user "may gain an understanding of just some aspects but this may fall short of the breadth of understanding implied by terms like transparency and explainability" [Khosravi et al., 2022]. Further accepting the reasoning by Kay [2021a], we see explainability as a more "techno-centric" concept, whereas scrutability is more focused on the user's active understanding-seeking.

Kay and Kummerfeld [2013] also define roles and principles of designing scrutable user modelling and personalisation, which arise from solving problems that relate to *privacy*, *invisibility* (i.e. the fact that it is often impossible to determine if and to what extent the interface differs from that of another user due to personalisation), *errors in user model* (which are exacerbated by the fact that there is often no explicit way for the user to fix them), *wasted user models* (i.e. the data that is captured by various devices but not utilised meaningfully for personalisation), and *the lack of control* of the user's personal information that is integrated into models.

To address these problems, Kay and Kummerfeld [2013, pp. 21–24] define three main roles for scrutable user models that underline why scrutability is valuable for the user. These roles are: 1) *control of the modelling process*; 2) *validity* (allowing the user to validate or rectify modelled information, thus tackling the problem of errors in user model); and 3) *re-purposing* (of the user model for navigation, influence, learning, improving usability, and programmer accountability). Concretely, controlling the modelling process addresses the problems of invisibility, privacy, and wasted data, by allowing the user to see all (potential) uses of the user model within an application, as well as by facilitating the reusing of data across multiple personalised applications. Controlling the model also includes allowing the user to understand the inflow of data, that is, to see where each piece of information is coming from. This, alongside controlling the reasoning of the user model – that is, investigating the processes of entailing new information from existing data – also contributes to addressing the problem of invisibility. Concerning the role of re-purposing, Kay and Kummerfeld [2013] state that scrutable user models serve five sub-roles: navigation (whereby the user model acts as an interface to an information space); influence (scrutable outputs such as recommendations are more likely to be accepted by the user); learning (scrutable user models serve as a mirror that prompt user reflection, as was also shown in the dissertations by Conlan [2005] and Czarkowski [2006]); usability (as scrutable models support personal, comparative, and long-term usability by allowing the user to understand and control the underlying model and thus to complete their tasks

successfully and recover from errors); and programmer accountability (designing scrutable systems necessitates that the programmers to remain mindful of the end-user). Re-purposing thus addresses the problems of invisibility and wasted user models.

Kay and Kummerfeld [2013, pp. 25–27, 37–38] propose four principles for building scrutable user models: 1) *parallel design*; 2) *first-class citizen*; 3) *context-based interpretation*; 4) *client side*. Parallel design necessitates that, through the whole process of designing personalised systems, both the interface issues and the user model representation and reasoning are considered together. A scrutable recommender system proposed by Glass et al. [2008] and a service-oriented adaptation by Koidl and Conlan [2008] are cited as examples of where this principle was applied. First-class citizen denotes a vision of a centralised user model that is shared among all personalised applications that the user uses and that is scrutinised and controlled through a unique interface. Context-based interpretation refers to the goal of enabling the user to understand how their user model and the contained data would be used in different contexts or applications. A cited example of the work that attempts to achieve this vision is SASY by Czarkowski [2006], a scrutability interface which would show only the parts of the model relevant to the current context and application. Lastly, through the client-side principle, Kay and Kummerfeld [2013] highlight the belief that the first-class citizen user models should be stored in a manner which guarantees full user access and control over the stored data. Thus, the principles of first-class citizens and client-side explicitly call for a paradigm shift of the current common practice, where user models are mostly application-specific and stored on servers inaccessibly for the user, especially if they are not tech-savvy and cannot use API calls to scrape the server for their own modelled data.

In their summative work, Kay and Kummerfeld [2013] also revisit several prior user studies and applications that showcase how these principles of scrutability can be applied in practice. Examples include a framework for creating personalised web pages, a program for creating personalised learning paths, and a personalised indoor locator interface. Additionally, IoTum – presented by Kummerfeld and Kay [2017] and evolved from the *um* toolkit [Kay, 1994] and the *Personis* user model Kay et al. [2002] – is a framework for user modelling for IoT applications. The application can *tell* evidence to IoTum, *ask* it for interpretations, or *listen* for triggering actions. Decisively, IoTum facilitates scrutability and accountability by maintaining provenance information. Notably, our proposed framework of a scrutable proactive IPA summarised in Section 3.1, envisioned not only exchanging data with applications in a similar manner but (proactively) controlling the IoT devices and corresponding with other people or their IPAs as well. This would necessitate extending the framework to allow for scrutability and control of the broader effects of the IPA's models and actions. This requires extending the map of the personalisation-related foundational problems and scrutability principles that arise therefrom.

## 2.4 Foundation Models

The years during the work on this dissertation were, as Herder et al. [2024] put it, a "period of upheaval", with several technological breakthroughs enabling a more precise personalisation and a further integration of online and real-world realities. Enabling this change in human-computer interactions is increased computing power, advanced cloud services, and generative artificial intelligence, notably that powered by large language models (LLMs) [Herder et al., 2024]. LLMs are an instance of a class of models which a group of researchers at Stanford University [Bommasani et al., 2022] termed foundation models, to emphasise both their central role in many downstream tasks as well as incomplete character. This dissertation does not investigate foundation models directly, but given their influence on both contemporary research trends and user expectations (which we discuss further in Chapter 7), in this and the following section we provide an overview of foundation models and adjacent research, especially in relation to human-computer interaction, explainability, and pathways towards realisation of a greater degree of proactivity and delegation through agentic systems.

As explored by Bommasani et al. [2022] in their collaborative report, foundation models are not conceptually new and rely on the decades-old ideas of deep neural networks and self-supervised learning. The key difference that led to their significance is that foundation models are trained at scale and subsequently adapted across a broad range of downstream tasks, such that a single model family may underpin heterogeneous deployments and stakes. This paradigm shift foregrounds both expanded capability and amplified risk: the same general-purpose backbone can be re-contextualised into applications with markedly different failure costs, while inheriting systemic concerns such as reliability, bias, and misuse. In everyday life, the paradigm shift is most frequently encountered through natural language processing, enabling conversational interfaces that dominate standard benchmarks. Technically, the behaviour of foundation-model chat interfaces is governed by token-level generation under finite context and is underpinned by the transformer network developed by Vaswani et al. [2017]. [Bommasani et al., 2022]

Kadavath et al. [2022], from an LLM-maker Anthropic, have argued that language models are well-calibrated on questions in various domains, as their tests have shown that LLMs can indicate whether their own previously generated answers are correct, as well as that models can be trained to predict the probability of being able to correctly answer a question without a specific prior answer. Moreover, Cosseddu et al. [2026] showcase building of an LLM-based system that references all produced output to a verified knowledge base, thus decreasing challenges that relate to verbose and convincing, yet incorrect output ("hallucinations"). Still, behavioural tests highlight that apparently competent generation may coexist with brittle underlying competencies, as shown by Rane et al. [2024], who have shown systematic failure

patterns alongside partial success on conceptually simple counting tasks as well as indicated how incorporating unusual adjectives for objects counted can significantly affect the performance of tested foundation models of varying sizes. Similarly, Liu et al. [2024] identify the "Lost in the Middle" phenomenon, demonstrating that even long-context models can exhibit degraded performance when relevant information is located in the middle of an input. And even Kadavath et al. [2022] qualified their well-calibrated argument by reporting a lower performance on calibration in new tasks. Such findings situate conclusions that cast doubt on LLM's self-checking capabilities [Yu et al., 2024] and describe "prompting LLMs [as] an art in itself" [Broy et al., 2025].

Crucially, complex capabilities of foundation models are said to *emerge*, that is they are implicitly induced rather than explicitly contracted, so even the developers of foundation models "lack a clear understanding of how they work, when they fail, and what they are even capable of" [Bommasani et al., 2022]. Amini et al. [2023] provide an overview of text generation from language models, emphasising that modelling and decoding decisions (e.g., the choice of sampling strategy) influence the resulting coherence and diversity, and thus become part of the user experience. As elaborated by Ivanova et al. [2021], one approach to understanding the internal computations and representational structures underlying the success of foundation models is probing, in which an auxiliary model is trained to access internal representations and predict predefined target features, such as parts of speech. Ivanova et al. [2021] situate probing practices relative to neuroscientific inference, highlighting both promise and pitfalls in drawing conclusions about internal structure from behavioural tests. Hewitt and Liang [2019] caution that probe accuracy can reflect probe capacity rather than representational structure, motivating control tasks to assess selectivity. Hewitt et al. [2021] then extend this idea by measuring usable information beyond a baseline via conditional probing. Alternatively, the approach of *mechanistic interpretability* aims for reverse-engineering the detailed computations performed by transformers [Elhage et al., 2021]. Showcasing such an approach, Elhage et al. [2021] propose a mathematical framework for analysing transformer "circuits" in small models as a step toward reverse engineering how algorithmic structure can be implemented by attention-based components. As Bereska and Gavves [2024] conclude in their review paper, mechanistic interpretability has the potential to prevent catastrophic failures of increasingly potent yet inscrutable AI models and call for standards-setting (metrics, benchmarks, and tests) and concepts-clarifications (by integrating existing literature and terminology).

## 2.5 Agentic Systems

The last few years have seen a resurgence of research on proactivity and delegation within the context of *agents* and *agentic systems* [Mihale-Wilson, 2025]. The underlying idea is to use

LLMs as drivers of systems capable of acting via APIs, enabling environmental interactions and multi-step plan completion [Yao et al., 2023]. Yao et al. [2023] operationalised this shift as an interleaving of reasoning steps and task-specific actions, and introduced the ReAct model capable of gathering external information and maintaining an explicit progression trace of intermediate intent. While these traces are presented as more interpretable and diagnosable, they are generated via prompting rather than reflecting the underlying computation. Complementarily, Schick et al. [2023] showed that tool use can be treated as a learnable behaviour by training their Toolformer model to choose when and which APIs to call as part of its output. Among the tools included in the model was a calendar, although its use was limited to determining the current date. HuggingGPT, presented by Shen et al. [2023], illustrated an explicit controller–executor pattern in which ChatGPT is used to understand user intent, parse it into actionable tasks and then select and invoke expert models hosted on Hugging Face that solve these tasks. Shen et al. [2023] argue that such use of LLMs as controllers of various AI tools offers a promising path for complex autonomous task-solving in various domains.

Several similar systems have also been built, typically utilising LLMs for agent planning [Shen et al., 2023]. The research space also includes diverging use or related concepts with surveys such as that by Zhang et al. [2026], providing a unified vocabulary. They formally define *generalisable frameworks* as underlying methods, structural paradigms, or interaction logic intended to transfer across scenarios, which can then be used to build an *agent*. To illustrate this conceptual separation, Zhang et al. [2026] classify ReAct by Yao et al. [2023] as a generalisable framework, rather than an agent. The term agent itself, however, remained overladen and inconsistently used [Botti, 2025]. Wang et al. [2024] consolidate the literature into a unified view of LLM-based autonomous agents by separating core modules (e.g., perception/observation, memory, planning, tool use, and action) from the diverse application contexts in which these modules are deployed. Similarly, Chowa et al. [2026] emphasises the transition "from language to action", framing agents as tool users and sequential decision-makers whose performance depends on prompting, feedback, and increasingly on explicit memory and planning scaffolds.

This *modular view* also supports a principled discussion of how agent behaviour is improved: Du et al. [2026] distinguish optimisation strategies that change model parameters (e.g., fine-tuning and approaches based on reinforcement learning) from parameter-free strategies (e.g., prompting and retrieval augmentation), which is useful for delimiting what kinds of "learning" a deployed assistant can realistically exhibit. Zheng et al. [2026] pushes this modular decomposition forward into a roadmap for lifelong learning agents, thereby providing an appropriate bridge between LLM-based agents and IPAs in reference to the prior work by Li et al. [2024]. Similarly, Wu et al. [2024] argues for building agentic systems as compositions of conversational agents with differentiated roles (e.g., planner, coder, critic, and human-in-the-loop). Notably,

such architectures also facilitate specialised subsystems for information extraction and verification. For example, Chen et al. [2025] presents the HEAR framework that explicitly couples structured extraction with agentic reasoning and cross-modal verification protocols.

Despite rapid progress, *autonomy remains difficult* in areas relevant for delegative assistants: generalisable, long-horizon, tool-mediated tasks with compounding error and ambiguous intermediate state [Zhang et al., 2026, Zheng et al., 2026]. For structuring this discussion, Mohammadi et al. [2025] provides a two-dimensional taxonomy of what to evaluate (e.g., capabilities, reliability, safety, behaviour) and how to evaluate (e.g., interaction modes, benchmarks, metrics, tooling), which supports a more careful decomposition than a single aggregate "agent performance" score. As argued by Zhou et al. [2024], the driving promise of agentic systems is autonomous resolution of daily tasks, steered through interactions in natural language. To operationalise this promise, Zhou et al. [2024] introduces a benchmarking web environment that approximates the procedural complexity of everyday digital tasks. Zhou et al. [2024] also notes that LLM agents are increasingly evaluated on end-to-end task completion rather than on isolated question-and-answer competence. Another benchmark, put forward by Mialon et al. [2024], proposes a set of real-world questions that "require a set of fundamental abilities such as reasoning, multi-modality handling, web browsing, and generally tool-use proficiency". Their findings signal that tasks which are conceptually simple for humans can remain challenging for advanced agents.

Classical *interpretability debates* already caution against treating explanations as a monolith, and against conflating plausibility with faithful access to what drives behaviour [Lipton, 2018]. In agentic systems, attributing the actual reason for an outcome is further complicated because behaviour emerges from interactions among multiple components, including prompt design, memory retrieval, tool invocation, planning strategies, and environmental feedback—making attribution difficult [Xi et al., 2025, Wang et al., 2024]. In multi-agent systems, the interaction between the agents and the environment results in emergent, more complex behaviours, further complicating the understanding of both the resulting system-level risks and capabilities [Miehling et al., 2025]. Moreover, repeated runs of the same agent can yield different action trajectories, with Mehta [2026] finding that the inherent uncertainty of LLMs also transfers to LLM-based agents. Interestingly, Mehta [2026] also found that the consistency between trajectories of repeated runs and the length of trajectories indicates accuracy on the reasoning task. Relatedly, Botti [2025] tempers novelty claims by situating contemporary agentic and multi-agentic framings relative to earlier findings on agents, while still acknowledging the specific discontinuities introduced by foundation-model tool use and open-world interaction.

## 2.6 Chapter Summary

One of the earliest examples of proactive assistance was Clippy, but it was not scrutable, as users could neither investigate how its suggestions were personalised to them nor how to correct its behaviour, as reported by Kay and Kummerfeld [2013].

Scrutability, as originally defined by Kay [1999], denotes the ability of the user of an adaptive system to actively study or investigate the user model which enables personalisation. While Tintarev and Masthoff [2022] take the concept to describe systems that permit modifications of underlying user models, Kay and Kummerfeld [2013] state that user control is just one of the key problems that scrutable user models resolve.

Cowan et al. [2017] found that the gap between users' expectations and Apple Siri's actual capabilities, the dominant IPA at the time, was among the reasons users abandoned the assistant. Further research, including by Kulesza et al. [2013] and Rader and Gray [2015], also highlights the importance of the user's mental model alignment.

In their systemic survey of IPAs, de Barcelos Silva et al. [2020] anticipated that more advanced personalisation and proactivity would be among the chief anticipated future directions of research. A notable use case for IPAs is that of time management [Zargham et al., 2022, Berry et al., 2017].

Since the commencement of work on this dissertation, foundation models have enabled rapid advancement of AI-based systems [Bommasani et al., 2022], including in the domain of emerging autonomously acting AI agents [Mihale-Wilson, 2025].

# 3 Devising a Desired Human-Assistant Interaction Framework

To understand the role of scrutability in IPAs, we first needed to investigate what constituted a desired human-assistant interaction framework. We began this task by reviewing publications related to IPA design and user sentiments. Having identified studies by Cowan et al. [2017] and Völkel et al. [2021] as most relevant for prospective proactive IPAs, we then contacted the authors to gain access to their user study data repositories. Acquiring the dataset from these user studies allowed us to reanalyse them, focusing on user statements regarding proactivity, personalisation, delegation, and scrutability and to outline our preliminary human-assistant interaction framework. Details regarding our findings are presented in Section 3.1.

To further refine the framework and understand the open challenges and research perspectives in human-assistant interaction and related scientific disciplines, such as multi-agent interaction, data modelling and governance, conversational user interfaces, and explainability, we assembled a group of experts to discuss the challenges and opportunities of personalised time management assistance. The discussion was audio-recorded, transcribed, and thematically analysed, and the user-assistant interaction and modelling framework outlined by the experts was situated and expanded. Section 3.2 details these findings and their implications for the design.

Section 3.3 then synthesises the findings from our reanalyses and the expert discussion into a set of identified interaction and modelling challenges that ought to be addressed by scrutable time management assistants.

## 3.1 Reanalysing Raw Data from Prior Studies

The participants in a survey on the expectations of a perfect future assistant led by Völkel et al. [2021] anticipated that the system would need to be readily aware of changes in the physical environment the user is in. Such assistants would ineluctably rely on complex and dynamic models, as they would necessitate representing not only the user but also their social

relationships, mental and physical states, and the relevant broader context. Even for contemporary IPAs, with comparatively fewer capabilities, not trusting the system with complex tasks was recognised as a major issue by participants in the study conducted by Cowan et al. [2017]. To investigate whether users also expressed views on scrutinising the underlying models, or whether understanding how they work might mitigate such concerns, we wanted to examine the data from these studies.

This section presents the results of that reanalysis, and it was originally published in the proceedings of the 30th ACM User Modelling and Personalisation (UMAP 2022) conference and presented as part of the Doctoral Consortium track [Jeromela, 2022]. Subsections 3.1.1 and 3.1.2 summarise the reported findings of Cowan et al. [2017] and Völkel et al. [2021], respectively. Subsection 3.1.3 shortly introduces the method for our reanalysis, with Subsection 3.1.4 detailing our main findings. Finally, the resulting user-IPA interaction diagram is included and described in Subsection 3.1.5.

### 3.1.1 Infrequent Users' Experiences of Intelligent Personal Assistants

In their paper published at Mobile Human-Computer Interaction (MobileCHI), Cowan et al. [2017] investigate the experiences of infrequent users of Intelligent Personal Assistants, defined as individuals who have tried IPAs but have not incorporated them into regular practice. Their motivation was to examine a user group whose disengagement would highlight barriers to sustained use and the friction between user expectation and system reality. We obtained the original dataset used for the paper by contacting the corresponding author.

The user study had 20 participants (11 female and 9 male), all of whom were members of a university community (with 14 students and 6 "non-students"). Cowan et al. [2017] report that 14 participants were Irish, with all participants being "native or near native English speakers". The study comprised two parts, with the first being an online survey and performing six tasks using Siri, chosen as common from a prior work by Luger and Sellen [2016]. Specifically, the tasks were to find out the weather in Dublin on the following day, to get driving directions to Dublin city centre, send an SMS to a phone contact, discover the number of calories in the previous meal, set a reminder, and search for a recipe. The purpose of the tasks was to inspire the second part of the study. The second part was discussions in focus groups of 3–5 participants lasting approximately 60 minutes. The transcripts of these discussions were the main source of data for the analysis. Namely, the transcripts of the discussions were first coded and then qualitatively analysed following the six steps of inductive qualitative analysis by Braun and Clarke [2006].

The analysis identified that infrequent use is not primarily driven by a lack of awareness or novelty fatigue, but by *a persistent gap between user expectations and actual interactional experience*. Participants repeatedly described IPAs as falling short of the implied promise of natural language processing, command flexibility, and human-like understanding. Rather than perceiving assistants as adaptive collaborators, users framed them as brittle systems that require careful phrasing, accent changes, repeated command corrections, and ongoing monitoring. Cowan et al. [2017] conclude their discussion by stating that while they had identified user desire for a more conversational interaction paradigm, IPA developers should consider whether humanness is indeed the right metaphor, as it signals to the user a level of system performance beyond actual capabilities.

A further finding is the frustration with the *severe limitation on task complexity* that IPAs may successfully perform. Even users who engaged with the technology frequently did not trust the system with complex tasks, restricting their usage to largely simple, low-stakes commands such as setting timers, alarms, or checking the weather. Notably, the authors identified the tendency of the participants to see IPAs as an interface that had to be learnt and highlighted that many users lacked a clear mental model of the IPA. This led to uncertainty about the systems capabilities, how the system changes over time, and how exactly it works. Ultimately, IPAs were thus seen as having an interactional cost that outweighed their benefits, especially when compared to alternative modalities such as manual input or human assistance. Cowan et al. [2017] show that disengagement with IPAs was a result of accumulated interactional failures such as command misrecognitions, inappropriate responses, or an inability to recover gracefully from errors.

Participants in the user study expressed heightened sensitivity to *privacy, monetisation, and data permanency*, often articulating uncertainty about who has access to their data and how it might be repurposed over time. A prominent concern was misusing the data for later targeted advertising. A notable place of user discomfort was the feature of wake-word activation, which implied that the smartphone microphones were on. Interestingly, the users saw interpersonal data as more sensitive for the IPAs to access and handle in tasks than data on banking or health. Cowan et al. [2017] ultimately propose that a continuous process of user consent akin to that proposed by Rodden et al. [2013] – rather than an all-in-one consent – may aid in alleviating some of the identified privacy and data stewardship concerns.

Crucially for the premise of our research, the study identified that this lack of trust is fundamentally rooted in opacity. Participants explicitly linked their hesitation to their ignorance of the system's internal state and data handling practices. Cowan et al. [2017] highlight a pervasive anxiety stemming from a concern about what the companies that made the IPAs do with the user data, thereby quoting participant [FG5, P1] who stated that the access to

confidential data is not the core issue, but the uncertainty on how the company uses it: "they don't talk about it, so we don't know and we have doubt, whether they will store it or not, or keep it or not". This uncertainty was compounded by the black-box nature of voice data processing and storage. Additionally, users expressed frustration with the lack of personalisation and customisation, noting in particular that the IPA does not account for their preferred apps and instead uses apps from the same service provider.

### 3.1.2 Users' Envisioned Dialogues with Perfect Voice Assistant

Völkel et al. [2021] present results of their dialogue elicitation study on how users envision interacting with a future "perfect" voice assistant, focusing on dialogue structure, assistant initiative, and the social texture of interaction. Their paper was published at the Conference on Human Factors in Computing Systems (CHI), and the dataset had been made publicly available by the authors[1].

The data was generated by 205 Prolific users who stated their native language was British English. The participants were gender-balanced (49.3% male, 50.2% female, 0.5% non-binary), with a mean age of 36.2 years and a range of 18–80 years. Völkel et al. [2021] asked the participants to "imagine living in a smart home with a voice assistant", which is "perfect" and has no technical limitations to its capabilities or natural-language understanding. The participants were then elicited to write dialogues for both the user and assistants in given scenarios that reflect 8 use cases: 1) search (finding film times for a local cinema); 2) music (selecting background music for solo cooking); 3) IoT (managing lights when falling asleep while reading); 4) volume (adjusting levels to accommodate noise-sensitive neighbours); 5) weather (planning attire for a trip to Italy in two days); 6) joke (entertaining a group of friends); 7) conversational (seeking assistance when unable to fall asleep); and 8) alarm (preventing oversleeping for an important morning meeting). This framing is explicitly designed to surface expectations about what the assistant should do "on its own initiative" and how it should coordinate with the user.

In total, the participants produced 835 dialogues, with each dialogue rendered in a two-column format (for the user and the assistant) and grounded in a scenario prompt, with an average dialogue length of 44.23 words and 5.03 lines (standard deviations of 28.71 and 2.73, respectively). Völkel et al. [2021] then analysed the resulting data by coding dialogue moves and dialogue properties (e.g., whether an utterance expresses a question, humour, or an opinion), quantified distributions across scenarios, and related these properties to individual-differences measures included in their protocol.

[1] At https://github.com/kartikn27/nlp-question-detection (last access 2026-03-10)

Across the elicited dialogues, Völkel et al. [2021] report a prominent initiative-bearing assistant role: participants frequently cast the assistant not merely as a command executor but as a cooperative partner that anticipates needs, asks clarifying questions, and proposes next actions within the ongoing context of the user's day. Still, dialogues were user-initiated in 91.9% of cases, with a significant cross-scenario variation (in the scenario about volume being too high, the assistant initiated the conversation in 27.3% of the dialogues). Similarly, endings were user-terminated in 67% of the cases). Nevertheless, 80.0% of participants handed over the lead of the dialogue to the assistant on at least one occasion. This interaction pattern indicates that even in "perfect assistant" imaginaries, initiative is often only expressed as a assistant-side initiative within a user-framed episode, rather than as fully autonomous, assistant-triggered interventions. Notably, only three participants in four dialogues have imagined the assistant refusing a request, implying strong user authority. Still, 16.1% of participants in their dialogue imagined trusting the assistant with complex tasks or demanding social interactions without providing detailed instructions. A notable example of this is P41, who envisioned delegation of communication: "I need you to write an email to my daughter's college. [...] The additional help provided for her because of her dsylexia [sic]. They promised reader pens and a dictaphone but she hasn't received them yet. Please ask why". Moreover, Völkel et al. [2021] report that participants had imagined the assistant choosing a suitable outfit for them, paying for expenses, editing business presentations, and planning trips.

Beyond task-completion moves, the envisioned dialogues also contain substantial "relational" material: the authors identify that a notable portion of assistant turns involve social talk, including chit-chat (reported as 40.0% of participants have at least once imagined a chit-chat with the assistant). In addition to chit-chat, a thematic analysis has identified two further types of elicited social talk: social protocol and interpersonal connection. Völkel et al. [2021] also note that humanisation was not uniformly desired, highlighting its variation across scenarios. The proposed humanisation utility included enriching the conversations through joking or steering opinion-sharing. The desired degree of humanness was reported as ranging from purely functional to that of a friend. In reference to work on the evolution of language by Dunbar [1996], Völkel et al. [2021] also suggest that social talk may serve to establish rapport and increase trust.

Crucially, the paper reports a strong desire for proactivity and the need for personalisation. The majority of participants envisioned an assistant that is "smart, proactive, and has knowledge about the user", with Völkel et al. [2021] concluding that this aligns with the desire for a conversation UI. Still, user preferences regarding the assistant's role, opinionatedness, and use of humour have varied, which may partially be explained by differences in personality. Namely, participants' envisioned dialogue properties were contrasted with their Big-Five (OCEAN) personality scores, which uncovered that the humour in the dialogue is negatively predicted

by Neuroticism (with the model implying a reduced likelihood of humour as Neuroticism increases) and that Conscientiousness positively predicts the presence of user questions in the dialogue.

### 3.1.3 Reanalysis Method

For the study presented by Cowan et al. [2017] on the infrequent users' experiences, the first author provided us the "flow-of-consciousness" transcripts of the semi-structured interviews within five focus groups. Rather than trying to recreate the findings by extracting common themes, we read the transcripts and highlighted statements indicative of user expectations and predictions regarding IPA proactivity, delegative capabilities, scrutability, and relevant use cases. These statements were then revisited and the most notable ones were copied to an Excel table and classified according to the four topics of interest: 1) envisioned features that overlap with our predictions (proactive/delegative/negotiating); 2) confirmation/validation; 3) explanations/studying of the model/recommendations/decisions; and 4) insightful remarks regarding current use cases.

The data for the study reported by Völkel et al. [2021], in which the users were asked to write dialogues with their envisioned perfect future assistants, is publicly available at the git repository linked in their publication (cf. Appendix A1 for all links referred to in this dissertation). Due to the dataset's much larger size (205 participants writing dialogues for nine scenarios with each dialogue having around 5 lines on average), a Jupyter notebook was created in Python that filters out dialogues not containing at least one of the terms relating to the themes of interest. The remaining dialogues for the prescriptive scenarios as well as all dialogues for the open scenario were read and an Excel table was filled, similarly as explained above.

The selected statements and dialogues snippets from the two Excel files were then compared and analysed again. Finally, the most representative statements were selected and used to create tables A2.1 and A2.2 in Appendix A2 on page 259[2]. The resulting findings are summarised in the subsection below.

### 3.1.4 Key Findings of the Reanalysis

One of the scenarios by Völkel et al. [2021] prompted the users to write down the dialogue with their assistant about visiting a local cinema. Participant *155* imagined the IPA to ask the user for a seat position preference, whereas for participant *105* the IPA just booked seats and noted that they are in the user's "preferred seating area". Therefore, the participants by

[2]Both Excel files and the Jupyter notebook can be seen at https://gitlab.com/Jovan_NS/envisioned-ipa-scrutability (last access 2026-03-10).

Völkel et al. [2021] envisioned that the need for explicit verification would depend on prior interactions. Likewise, users in both studies envisioned proactivity (both in terms of initiating conversations and proposing activities) and delegation (negotiating meetings, writing letters and text messages on the user's behalf). Some users by Völkel et al. [2021] also envisioned elements of scrutability such as studying why a certain suggestion had been given or a proactive action performed. The IPA's awareness of social relations is implied in many open scenario dialogues from the study by Völkel et al. [2021], in which the IPA is capable of proposing, discussing, and elaborating on gift ideas for the user's significant others.

When discussing their views on Siri's capabilities for the study by Cowan et al. [2017], participant *F1P3* noted that they do not trust the IPA and that they constantly feel that they have to check whether the appropriate action has indeed been performed. Providing explicit verification and having the IPA ask for additional details in case of ambiguous requests could potentially resolve this. In the imagined dialogues with the perfect assistant written by Völkel et al. [2021], such behaviour has been envisioned by many participants in numerous situations. For instance, when comparing *search-45*, *search-105*, and *search-155* (that is participants in rows 45, 105, and 155 of the so-called search scenario that focused on finding information concerning cinema screenings, cf. Appendix A2), we can see that the assistant verified that it had truly booked the tickets. In addition, in the dialogue *open-143*, the assistant proposed an event time, which the user then explicitly verifies: "Hello Sara, I have a busy schedule with work and shopping, what would be the best time for me to exercise *Hello, Looking at your schedule the best time would be in the evening at 6pm* Thank you Sara that time would be great". Therefore, the participants in the study by Völkel et al. [2021] expected personalisation and envisioned that the need for explicit verification would depend on prior interactions and the task. That such assistant-based verification might be a desired improvement was also suggested by an infrequent user in focus group three who had stated: "I would have liked a question-answer to be more of a dialogue. So Siri would ask me a question, what kind of, 'are you making this [recipe] now or would you like me to create a shopping list for the ingredients?' ".

Proactivity – both in terms of starting an interaction and providing innovative suggestions or additional actions – has also been suggested by multiple participants, such as in the case of dialogue in *volume-142*, where the IPA noted that the music might be too loud and suggested not only lowering the volume but also setting that volume for the new default value. This is in stark contrast with current non-proactive and explicitly initiated voice assistants, that often turn on in inopportune moments, as stated by *F2P2*. Additionally, in the dialogue in *volume-169*, the IPA also suggested preemptively sending an apology to the neighbours. After the user's confirmation, the IPA formulated and sent the message as well as reported back the neighbours' response. Other users also envisioned such cases of delegation by the IPA

in electronic communication, notably in dialogues in *search-5*, *volume-173*, and *open-107*. In these three cases, this delegation is for the purposes of arranging a meeting with other humans on the user's behalf. Being aware of the user's calendar and time constraints was also envisioned by other participants, for instance in dialogues *alarm-197* and *open-143*.

The participants in both studies also envisioned elements of scrutability in a broader sense, where the users can actively study their assistant's functioning. In the case of dialogues from data from the study by Völkel et al. [2021], *joke-135*, *joke-137*, and *open-26* are particularly interesting. In them, the users explicitly asked the IPA for explanations regarding certain suggestions or the source of the assistant's knowledge. Moreover, in the case of *joke-135*, the IPA even offered to alter its behaviour in a manner that would prevent it from collecting similar pieces of information in the future. Additionally, many dialogues contain lines where the IPA unpromptedly offered its "rationale", e.g. in *search-5*, *search-155*, and *volume-142*. With regards to the issue of *wasted user models*, a user in focus group 3 by Cowan et al. [2017] noted that they think Siri is likely capable of much more than they are aware, but that they are unwilling to invest time in learning how to utilise it fully. Contrastingly, in the data from the study by Völkel et al. [2021], in the dialogue in *conversation-5*, the assistant offered its user to create a note in addition to highlighting the text, which the user declined but thanked the assistant for reminding them of a feature they had forgotten existed.

In conclusion, our reanalysis found statements that would indicate that participants of these studies had anticipated elements of scrutability, proactivity (both in terms of initiating conversations and proposing activities), and delegative capabilities to be integrated in the design of future IPAs. Delegation was envisioned to involve negotiating meetings, writing letters and text messages on the user's behalf. Some users in the study by Völkel et al. [2021] also envisioned elements of scrutability such as studying why a certain suggestion had been given or a proactive action performed. The IPA's awareness of social relations is implied in many open scenario dialogues by Völkel et al. [2021], in which the IPA is capable of proposing, discussing, and elaborating on gift ideas for the user's significant other. In the next subsection, we extract these findings into a diagram of the envisioned scrutable interaction.

### 3.1.5 Envisioned Scrutable Interaction Diagram and User Modelling

Following this reanalysis and the state of the art at the time, we proposed our scrutable IPA interaction and modelling diagram, which can be seen in Figure 3.1. As elaborated in Section 2.1, the terms agent, assistant, and voice interface are sometimes used interchangeably in the literature on conversational assistance. In this diagram, however, we strictly differentiate between the three. Here, the *agent* is the central control and communication engine of the

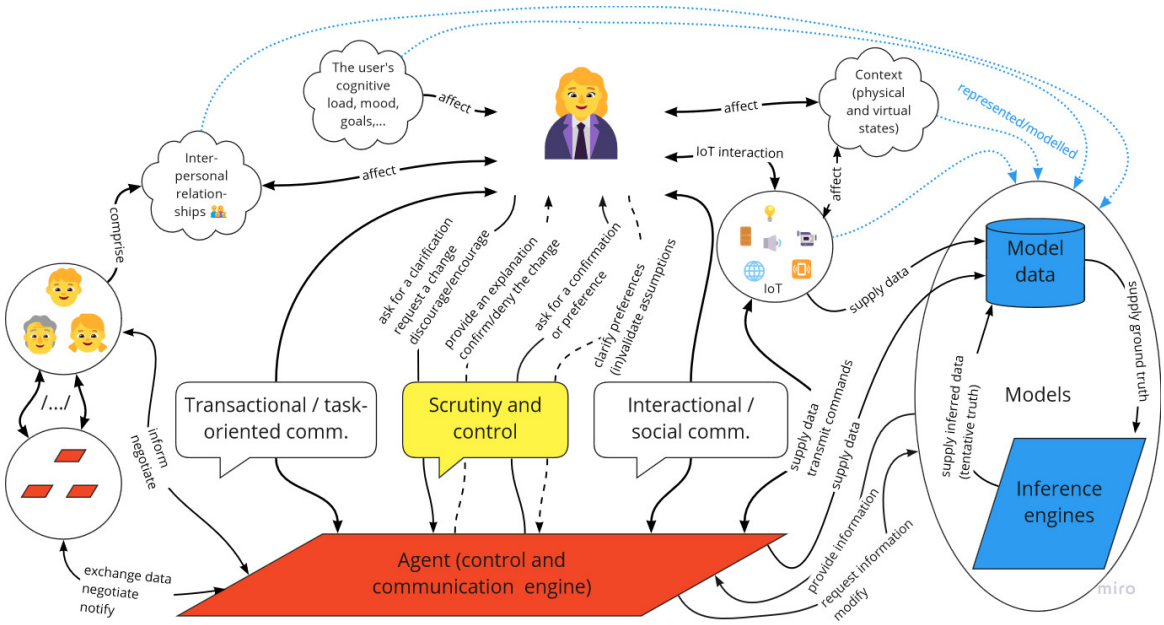


Figure 3.1: Diagram of the envisioned IPA interaction design, as published in [Jeromela, 2022]

intelligent personal *assistant*, where the data from various models is utilised to provide services and power the *voice interface*[3]. Following the classification from Clark et al. [2019c], the voice interface can facilitate transactional (centred around a specific goal) and interactional (ranging from chit-chat to longer dialogues) communication with the user. In addition, we see communication for the purposes of scrutiny and model control as a novel interaction type, and the main aim of our further research. As indicated by Clark et al. [2019c], even though the purposes of these classes differ, they can overlap in natural conversations. Many IPAs are capable of performing actions in the physical world by e.g. controlling the IoT devices [Chung et al., 2017, Campagna et al., 2017]. Many widely deployed and popular commercial IPAs such as the one surveyed by de Barcelos Silva et al. [2020] and Tulshan and Dhage [2018], as well as on other systems based on open-source technology [Campagna et al., 2017] or through purpose-built assistants such as by Vashistha et al. [2019] are also able to affect physical surrounding by interacting with the IoT.

The envisioned IPA would be highly personalised and thus underpinned by modelling the user (representing their preferences, goals, state), interpersonal relations, context, and connected IoT devices (both in terms of their sensors and possible actions). This view largely overlaps with the six areas of digital personal assistance recognised by Meurisch et al. [2017]. Employing the multi-model, metadata driven approach to providing personalisation of adaptive systems, as proposed and evaluated by Conlan [2005] within the context of e-learning, allows for a greater

[3]In the years following our reanalysis published as [Jeromela, 2022], the term *agent* became even more overladen [Botti, 2025], which is why the following iterations of our framework no longer use the term for the reasoning and inference engine of the assistant.

degree of module reusability and control flexibility. As illustrated in Figure 3.1, these models would save both the received evidence as well as inferences from evidence interpretation. This conceptualisation is akin to that of ALIGN system architecture presented by Peirce et al. [2008], which comprises a cyclical process of updating models through adaptive interaction with the user. Unlike ALIGN, and crucially for the further alignment of the framework, we envision that the user would also be able to explicitly scrutinise and modify the models.

## 3.2 Expert Discussion on Proactive Time Management Assistant

To further refine our findings from the reanalysis and the literature review presented in the previous section and to gain further insight into desired features of an intelligent time management assistant, we organised a multidisciplinary expert discussion. The participants were four university professors and three PhD candidates. We asked the experts to voice their ideas and opinions regarding potential use cases, data stewardship principles, usability, and information modelling of such IPAs for time management, as well as to sketch how they envision apt human-assistant interaction. This section includes findings originally presented as a work-in-progress paper at the 5th ACM Conversational User Interfaces (CUI 2023) conference [Jeromela and Conlan, 2023a] and as a full conference paper at the 11th ACM Human-Agent Interaction (HAI 2023) conference [Jeromela and Conlan, 2023b].

### 3.2.1 Motivation

Murad and Munteanu [2019] observed that, regardless of the ever-advancing underlying artificial intelligence, voice assistants offer "marginally more than an information kiosk or automation control" for many user groups due to the lack of specific interaction guidelines. In choosing the method for our expert discussion, we wanted to contribute to establishing such interaction guidelines and, more concretely, to outline a model of a voice-enabled and proactive IPA for time management that would guide our exploration of the role of scrutability further. To frame the exploration, we chose time management as a desired feature of IPAs [Völkel et al., 2021] that has the potential to save the user's time [Cranshaw et al., 2017] and facilitate collaboration [Toxtli et al., 2018]. With these goals and frame in mind, we chose the participants and devised the discussion structure.

### 3.2.2 Discussion Participants

We contacted researchers with expertise in areas identified as related to IPAs for time management. Due to the unavailability of some initial invitees, the final group comprised six male

and just one female researcher. Four university professors and three PhD candidates participated as experts. The experts described their fields of research as follows: 1) personalisation, adaptive systems, and proactivity; 2) data integration, databases, and the semantic web; 3) trustworthy AI, data governance, and standardisation; 4) conversational user interfaces, psychological and cognitive aspects of human-machine dialogue, and speech interface interaction design; 5) empathetic chatbots; 6) simulation-based development of proactive IPAs; and 7) explainability and counterfactuals in reinforcement learning.

### 3.2.3 Discussion Structure

The expert discussion took place in person in a seminar room of Trinity College Dublin. There were five distinct parts, as reproduced in full in Appendix A3. The participants were first asked to individually think of and write down use cases and scenarios in which proactive time management IPAs might be helpful for 10 minutes. A 15-minute-long presentation of initial ideas followed. The main discussion proceeded for just over an hour. The moderator delivered three prompts, roughly 20 minutes one after the other. The prompts asked the participants to discuss: 1) what data the IPA should collect and what constraints it should observe; 2) principles of communication, control, and data sharing; and 3) a graphical representation of the model of the envisioned IPA. The last prompt also asked the participants to sketch their ideas.

The discussion structure and the prompts were inspired by a prior study by Meurisch et al. [2017]. A notable adaptation from their approach – apart from a different assistance domain – is that we included a prompt on data sharing, communication, and control. These elements of human-assistant interaction were explicitly emphasised due to the inherently personal nature of time management. The dissertation author moderated the discussion, intervening only to provide prompts and ensure that all participants had an opportunity to voice their opinions. To moderate the discussion, a PowerPoint presentation was used as an aid, which is reproduced in full in Appendix A3.1. The event was organised in line with the institutional ethical guidelines, and all participants signed informed consent forms.

### 3.2.4 Data Processing and Analysis

Our analysis method is illustrated in Figure 3.2. First, the experts' sketches were photographed, and the recordings of the discussions were transcribed. Thereafter, we performed a *thematic analysis* of the outputs by following the five steps from the seminal paper by Braun and Clarke [2006]: 1) transcribing the recordings and familiarising ourselves with the data; 2) systematically generating initial codes; 3) searching for potential themes; 4) reviewing themes; 5) defining and naming the themes; and 6) producing the final analysis and the report. This

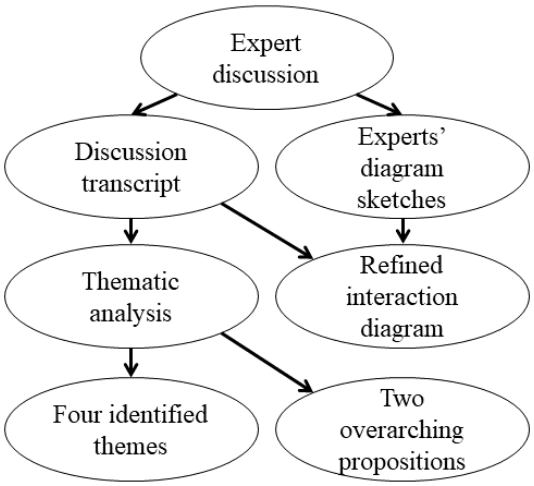


Figure 3.2: An illustration of the outputs of the expert discussion and the method taken

approach has been used extensively for qualitative analyses of human-assistant interaction, e.g. by Cowan et al. [2017], Olson and Xu [2021], Bartle et al. [2022], and Gonzalez-Jimenez et al. [2022]. One recent work that also used thematic analysis to understand perceptions towards IPAs (for home health aids) is by Bartle et al. [2022]. A notable difference is that we conducted the analysis with a realist/essentialist rather than a constructionist approach [Braun and Clarke, 2006, Potter and Wetherell, 1987, Widdicombe and Wooffitt, 1995], considering the fact that our participants are experts in areas related to the subject of the study and our own background allowed us to interpret their words and experiences in a more straightforward manner. The authors collaborated on data coding and refining the candidate themes over several weekly meetings, resulting in four final themes. Additionally, in performing the analysis, we recognised two overarching proposals spanning ideas across the four themes, concerning the role of explanations and data control in stages. Finally, the sketches the experts made during the discussion were combined and extended according to the accompanying statements in the transcript, resulting in the final interaction diagram. The data, including the anonymised transcript, the presentation shown by the moderator, the sketched diagram, and use cases from the post-it notes, is available in a shared drive[4].

### 3.2.5 Thematic Analysis

We performed a *thematic analysis* of our data following the five steps from the seminal paper by Braun and Clarke [2006]. Contriving and revising themes was done collaboratively, during weekly supervisor meetings, so to discuss the themes over four rounds of revisions. As can be seen in Figure 3.3, four main themes were identified. In addition, scrutability, i.e. the manner

[4]The data can be seen at: https://drive.google.com/drive/folders/1-crFEKo2jyr7d-_5OgBSLlSF7B9SASP3 (last access 2026-03-10)

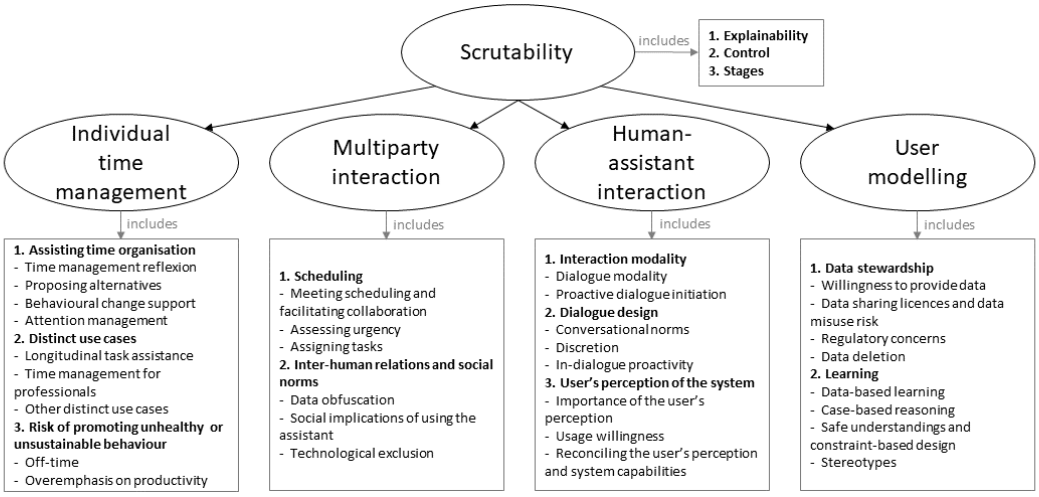


Figure 3.3: The four extracted themes and their subtopics; scrutability is seen as an overarching meta-theme of the discussion

in which the users can study, understand, and control the personalisation of the IPA was identified as a meta-theme that spans through all themes. Table 3.1 contains a summary of all the themes, while the subsections below describe them in more detail. *Stages* emerged in the discussion as an onboarding mechanism through which the user would be gradually introduced to the system's capabilities, and are discussed in more detail in Subsection 3.2.8.

### Individual Time Management

The theme of individual time management comprised ideas and comments on how the IPA could help an individual manage their time for tasks that do not directly affect other people, such as *behavioural change*. For instance, it was suggested that the IPA could nudge the user to meet their physical activity goals or stay within the desired screen time through proactive notifications asking the user to get back to the main task or take a break. The ideal mode of delivery of such disruptions was discussed extensively, especially when the user is potentially involved in safety-critical tasks such as driving or performing medical procedures. Furthermore, as *p*articipant *B* in *s*tatement *46* – shortened as *PB-S46* in the remainder of the text – explained, there are professions, such as lawyers and consultants, where the IPA could potentially propose a more goal-appropriate schedule. Moreover, PD-S136 suggested that longitudinal tasks – spanning a long time or comprising multiple steps – would be a specific use case where the assistant could propose concrete plans by, for instance, finding a suitable time in the calendar for practising a skill the user would like to master and then allocating the appropriate number of time slots.

Table 3.1: The names (and descriptive names, as in [Jeromela and Conlan, 2023b]) of the four identified themes and the main points contained within them

| Theme | Main points |
|---|---|
| Individual time management (Realising balanced assistance in personal time management) | 1. There is potential to assist people in organising their own time (to rethink their priorities and event allocations);<br>2. The IPA could propose an alternative structure for the week;<br>3. The IPA must balance its suggestions to prevent encouraging unhealthy or unsustainable behaviour. |
| Multiparty interaction (The role of the assistant in multiparty meeting negotiations and task allocations) | 1. The role of the IPA in facilitating human-to-human negotiation is delicate due to complex and dynamic social relations;<br>2. When sharing data about its primary user with a third party, the IPA must obfuscate/censor the data appropriately for the relation the user fosters with the party. |
| Human-assistant interaction (Enabling the user to gauge and utilise the assistant's capabilities) | 1. The IPA should help the user understand its actual capabilities (the user's model of the assistant);<br>2. The user's perception of the system capabilities is tied with modality choice, the assistant's conversational capabilities, and the appropriateness of the proactivity. |
| User modelling (Data stewardship and learning mechanisms in user modelling) | 1. Time management assistance relies on personal data and thus requires appropriate data use and sharing licences;<br>2. Case-based reasoning, user-set rules, and proactive explanations are potential ways for the user to control and adapt the IPA's (proactive) behaviour. |

### Multiparty Interaction

The theme of multiparty interaction included discussed scenarios where the IPA would interact not only with its main user but with other humans and their own IPAs as well, in order to schedule meetings or facilitate collaboration. Thinking again in terms of stages, it was envisioned that a simpler stage could be to offer an obfuscated view of the calendar with approved people. A more data-intrusive and technologically more challenging feature would be allowing the IPA to infer the social relationships of the user and to adjust both what information to share and with whom. The importance of model explainability for such potentially socially consequential features was repeatedly stressed. PB-S207 cautioned that users might refrain from having the IPA delegate arranging their meetings in order not to give the other party the impression that they were not worthy of a personal invitation. Contrastingly, should the system become widespread, those unable or unwilling to use the IPAs for time management might be disadvantaged through technological exclusion. Lastly, PD-S166 warned of accidental multiparty interaction. The given example was that the IPA might notify the user of a sensitive medical appointment via loudspeakers while the user has guests in the house. Proposed

remedies included privacy-sensitive or – more challengingly and intrusively – context-sensitive adaptive modality selection.

### Human-Assistant Interaction

A significant part of the human-assistant interaction discussion focused on dialogue design and modality selection. PD-S46 proposed that the assistant could look for contextual cues when choosing the output modality. The proposed cues for voice interaction were situations when the user's hands or eyes are focused on another task (given examples: driving, cleaning, assembling furniture). Moreover, PD-S200 underlined that personalisation should ideally also allow the user to specify when they would prefer the assistant to communicate with them using voice output or any other modality. Regarding dialogue design itself, PC-S199 believed the assistant "should also keep track of the social norms where the agent is deployed" and that it "shouldn't be lying". PD-S204 believed communication norms between humans and IPAs qualitatively differed from interhuman conversation.

PD-S200 proposed to differentiate between proactivity within and outside of ongoing dialogues. The former is reified in the IPA's responses within ongoing dialogues, whereas the latter comprises the actions the assistant performs from a "dormant state", such as rescheduling or initiating a new dialogue itself. The participant also noted that both forms of proactivity would ideally be personalised to the user and adapted for the relevant context. The importance of scrutability in this aspect was stressed by multiple participants (e.g. PB-S46, PE-S151). Furthermore, PD-S86 emphasised that the user should be able to adjust how proactive the IPA is in adjusting their calendar: "So imagine, say, for instance, if you're writing a paper [and] you're thinking about something you're doing, and then your agent is going 'Your calendar has been changed!' and it takes you out of your primary task and into that secondary task there".

The perception the user would have of the assistant and its capabilities was discussed. The participants (statements 233–251) explicitly associated the work of Roger K. Moore with the role of voice in human-assistant interaction. Specifically, they noted that the degree of common understanding depends both on the assistant's user model and the user's "internal model" of the assistant. The participants also noted that the user's perception of the IPA would change with usage and improve or deteriorate depending on how successful it was in understanding and fulfilling tasks. Furthermore, bad communication design (inappropriate voice, modality choice, dialogue design) could lead to misaligned expectations and abandonment of the technology, as "we assume that it will do very little [for us]" (PD-S249). Regarding the usage willingness itself, PG-S96 hypothesised that it would also depend on one's general openness towards new technologies and that it would span to encompass those who would likely eagerly adopt an

automated time management system for both private and personal tasks, over those who would prefer to keep their off-work calendars completely private, to those "who would really need to be convinced that this will help". The *stages* combined with scrutability were offered as a potential remedy for both the perceived lack of system utility and the general misalignment of perceived and actual system capabilities. As described by PA-S192, the user could "test-drive" more advanced system capabilities through hypothetical scenarios showcasing possible functionalities.

### User Modelling

The points the participants raised regarding user modelling can be split into two categories: data stewardship and learning mechanisms. Regarding the former, a recurring question (e.g. PG-S72, PB-S103, PF-S154) was the user's willingness to provide data, given the complex capabilities that were envisioned for the system (urgency and task complexity approximation, modality switching, inferring time constraints, understanding social and physical context, tracking attention and cognitive load) as well as the misuse risk. In that sense, PG-348 proposed all data collected about the user be tied to licences, which would also limit what the data may be used for. Participants agreed (e.g. PB-46, PD-S137, PG-S188) that the scrutability of data collection, sharing, and use becomes more pronounced when the assistant is making decisions on the user's behalf, e.g. by rescheduling events or filtering incoming or outgoing information. As PB-S46 put it: "If there's some sort of preference or there's some model being built – who and how is that managed and owned? Because that's potentially very intimate information. So, the actual act of managing or scrutability or transparency there or explainability is really important".

There have been several proposed mechanisms through which the assistant could learn about the user's time management preferences, ever-changing social structures, and the broader context. PE-S202 and PA-S203 contrasted data-driven learning and modelled understanding, noting that a degree of prescribed instructions would likely be needed and debating the appropriateness of constraint-based learning for scheduling. Specifically, it was suggested that the assistant may keep a list of "hard" (i.e. never to be done) and "soft" (preferred) constraints. PB-S191 believed sharing sets of constraints to be a potent control mechanism, as it would allow control of both data usage and time assistance preferences. However, PB-S185 also accentuated that constraints alone are limited by humans' incapability to imagine all potential situations (not) bound by the constraints. An alternative learning mechanism dubbed "proactive explainability" was proposed by PF-S320, whereby the assistant would proactively propose an action to the user and offer an explanation for its suggestion. Through case-based reasoning and after enough similar situations, the IPA would potentially act without asking for prior validation. However, PG-322 deemed such an approach too risky, given how complex

the context of interhuman interactions may be.

### 3.2.6 Proposed Interaction Diagram

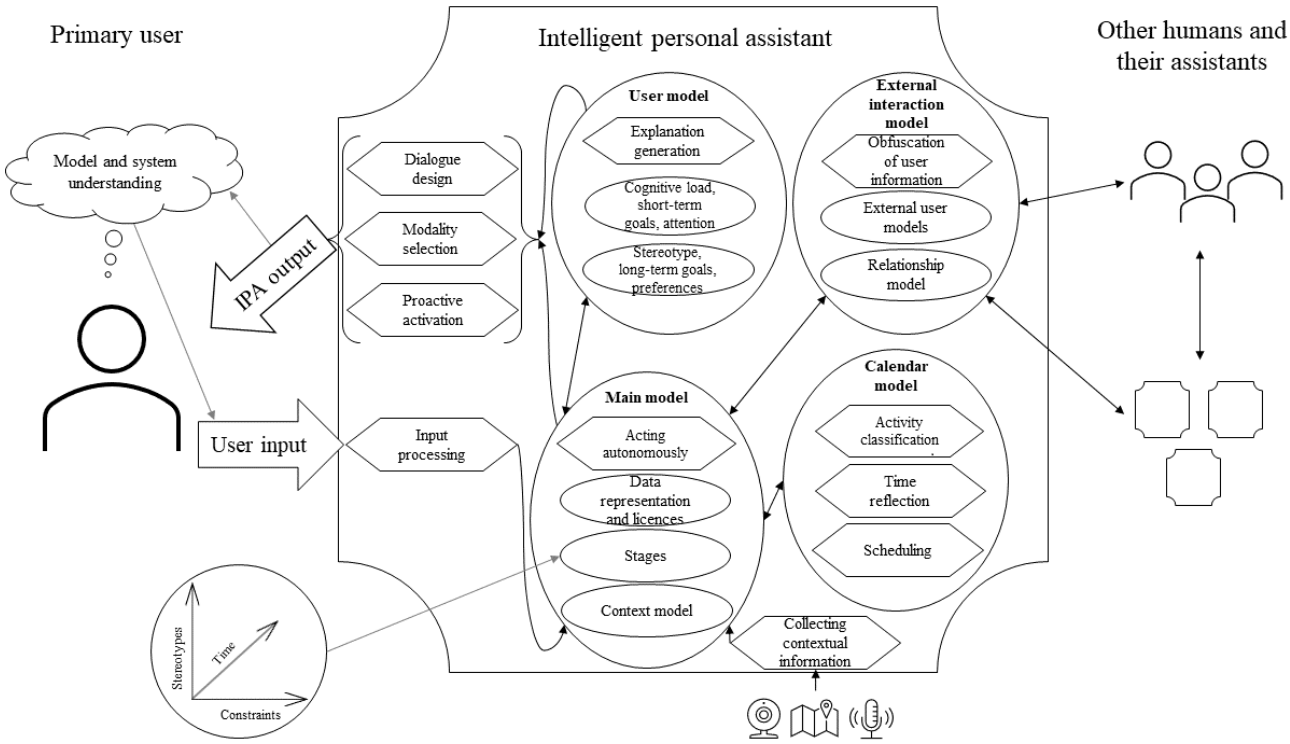


Figure 3.4: Diagram of the envisioned proactive time management assistant's internal design and its interaction with the user, arising from the expert discussion. The ellipses showcase data modelled and diamond actions performed by the assistant.

The last prompt asked the experts to summarise their discussion graphically and to sketch the assistant's interaction with the user and its internal structure. Based on the photographed drafts and the concurring discussion, the diagram in Figure 3.4 was made. On the left-hand side, the diagram emphasises the importance of the user's perception of the assistant, or the user's "internal model" of the assistant, that both structures the human-assistant interaction and is structured by it (PB-S241). The right-hand side showcases the assistant's communications with other humans (who are not its primary user) and their assistants. Thus, the IPA delegates scheduling and notifying. The bottom of the diagram shows that the IPA collects contextual information that would help it situate the user's request and choose its output modality.

Regarding the IPA's internal structure, PF-S299 claimed that if the assistant delegated the primary user's communication with other humans, it would require both its primary inward-facing user model and an outward-facing interaction model. The former would contain both shorter-term information about the user (e.g. their estimated attention level and cognitive

load as well as the short-term goals and intentions) and longer-term information (such as priorities and goals shaping the desired behavioural change, the user's preferences, and an assigned stereotype for system adaptation). The outward-facing interaction model would be used when the assistant communicates with others. Depending on the modelled social relationship and the task, the IPA would obfuscate information about the primary user to protect their privacy. Conceptually, each interaction might involve sharing a simplified external version of the user model. PF-S299 believed that such external user models would have to be fixed during the interaction with other humans and their IPAs, while the internal user model would be amendable during the interaction with the primary user (as it is scrutable). Since the investigated use case for the proactive IPA is time management, a calendar model is prominent in the diagram. The information about the user's calendar, as well as the activity classification, would enable the IPA to facilitate scheduling and to support the user in reflection on their time use. Lastly, the "main model" of the assistant would represent all data the IPA has access to, along with the licences for how that data might be used. As PG-S348 assessed, this is necessary for liability purposes and to enable the user's control of the assistant (through scrutiny of the IPA's behaviour). This is true in the case of data sharing as well (PG-S307). The stage would also bound the extent of the services. The stages, in turn, are based on constraints and stereotypes that also determine the assistant's autonomy levels (PA-S294, PD-S295).

The assistant must select the appropriate modality and dialogue design when generating the input. Moreover, the IPA could proactively initiate a conversation, propose an action, or notify the user about a change in their calendar. These actions would be based on the information from the main model, the calendar model, and the user model, which would facilitate personalisation. Through the information about the user, the IPA would try to generate an explanation personalised to the user (PF-S260). The depth of detail of explanations would ultimately be determined by the user scrutinising the assistant. Furthermore, as stressed by PD-S249, the system has to allow the user to scrutinise the limits of its capabilities. Otherwise, the user might underestimate its utility to them and stop using it.

### 3.2.7 Scrutability and Explainability

As elaborated in Section 2.3, we join Kay and Kummerfeld [2013] in their view of scrutability as an active user-centred process of investigating how the system is personalised to them that incorporates the whole user-system interaction. Accepting the reasoning from Khosravi et al. [2022], we see explainability as a more "algorithm-centric" concept, whereas scrutability is more focused on the user's understanding of how a system works for, and in conjunction with, them. Consider the question by PF-S195: "going back to explainability, if a certain

decision relies on certain other decisions that have happened in the past, how far back does that explainability-train go?". We believe that the "algorithm-centric" response to this question would be that the system should determine the sufficient level of detail so that the estimated user's understanding is maximised according to an internal measure. In contrast, the user-centric scrutable IPAs would allow the user to be the ultimate arbiter of the appropriate level of detail in the explanations through an active, ever-more-targeted process of scrutinising the system.

Both explainability and scrutability were deemed essential elements of human-assistant interaction by the experts. Regarding the former, PE-S63 and PA-S153 stressed that the IPA should be able to provide reasons for any (proposed) rescheduling or urgency estimate. Returning to the example from PA-S99 in which the IPA would decide which spouse is to drive a child to training practice, PE-S151 pointed out that the chosen spouse might justifiably want to know the explanation for the decision. Similarly, should the IPA recommend some background music (as proposed by F-S64 and already studied by Newbold et al. [2017]), the user may want to know if the song is recommended because the assistant holds a particular belief about them or because of sponsorship (PE-S151). Relatedly, the importance of transparency of potential goals of commercial entities in the IPA's actions was repeatedly emphasised, e.g. by PB-S103 and PG-S188. Yet, PB-S183 noted that explaining decision-making and the implications of privacy policies might be challenging for some types of algorithms.

Scrutinising the system's capabilities would allow for a more accurate user perception, as PB-S244 and PD-S249 noted. For instance, the user might want to scrutinise the extent of the IPA's proactivity in rescheduling events in their calendar (PF-S79). Privacy in assistant-facilitated multiparty interaction was seen both as a reason and a challenge for scrutability. On the one hand, scrutiny would allow the primary user to study and control what information is shared and with whom (PB-S46, PD-S178, PG-S186). On the other hand, to facilitate scrutability, the IPA must reveal private information about other users without their explicit permission (PA-S153, PD-S180).

### 3.2.8 Stages

Towards the end of the discussion, PA-S252, PA-S290, and PF-S320 proposed for the IPA to facilitate onboarding in *stages* so that the user can gradually adopt more system functionalities. This can be seen as an extension of scrutability, as the stages would frame the human-assistant interaction and allow it to change as the user's understanding of the system becomes more complete, and the IPA's modelling of the user becomes more accurate. The stages comprise services and functionalities the IPA offers and the corresponding necessary user permissions. A graphical representation of stages for time management assistant can be seen in Figure 3.5.

Stages would enable coping with vastly different user time management needs (PG-S98, PB-S103), different openness to technology and adoption reluctance (PG-S68, PA-89, PA-S292), the misalignment of user expectation and system capabilities (PD-S200, PB-S235), and issues with data control and with onboarding of new users (PA-S192, PB-S318).

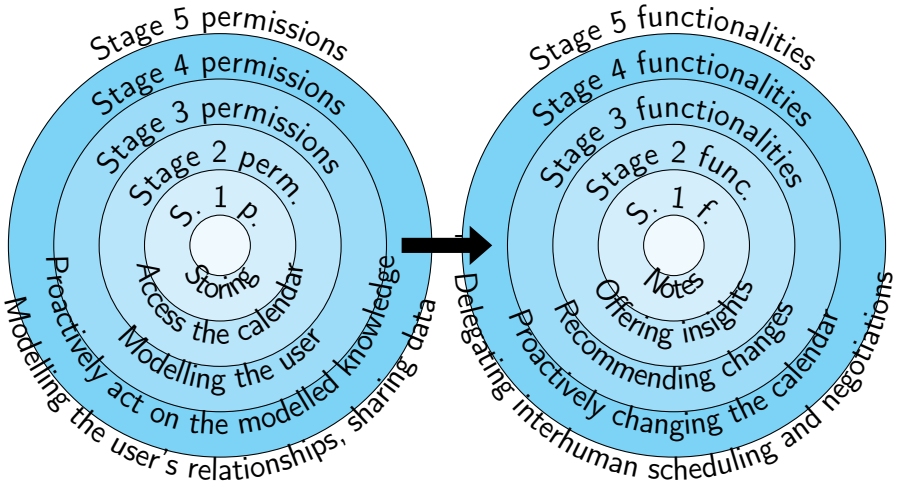


Figure 3.5: The proposed onboarding and control in stages: the user may "unlock" additional functionalities by permitting more data access and IPA autonomy. The experts envisioned stages as concentric and linear, yet they might conceivably "branch out", and the user might choose not to utilise one of the "inner" stage's functionalities.

Thus the IPA might allow the user to "test-drive" higher stages by showcasing what it would do in hypothetical situations personalised using the user's data currently available to the IPA. Specifically, the lowest stage might only allow the assistant to store the user's notes and priorities. As PG-S68 noted, "helping somebody to think through what their priorities are /.../ and actually to enunciate them to the inanimate object in some sense, that then helps them manage their time". Notably, this stage would not even necessitate the assistant accessing the user's calendar. As PA-S252 put it, the next stage would mean that the assistant "will help you reflect on what's going on in your calendar. Give you maybe some data-driven analysis of it, but you're still fully in control of any action to do with that calendar. It may not even make direct suggestions. All it's doing is helping the reflective capability". Stage three, as PA-S292 continued, activates when the user would "like the agent to recommend to [them] certain things [they] might change in [their] calendar directly rather than simply reflect on it". Further stages would involve proactive autonomous acting by the IPA and delegation of communication (PF-S320). As specific functionalities would be tied to stages (which in turn are tied with specific permissions and constraints), this human-assistant interaction mechanism would make the assistant more scrutable, as observed by PA-S319. Moreover, as suggested by PA-S192, the IPA might use stereotypes and data already made accessible to allow the user to "test-drive" the system, i.e. understand better the functionalities which would become (or remain) available should they change into a higher (or lower) stage. This might be useful as

an onboarding procedure for new users as well.

Significantly for our research question, control in stages (where functionalities are linked to permissions granted) may be the initial additional scrutability principle, whose importance is raised by the envisioned IPA's advanced capabilities. Similarly, a new role of scrutable models would be sustainable onboarding, which would allow gradual familiarisation of the user to the user model and the IPA capabilities that rely on it. In addition to the problems of privacy, invisibility, and lack of control, this preliminary new role and principle would also address the three novel problems that arise when considering use cases for advanced IPAs. These three problems relate to *proactivity* (When/why to initiate interaction? How do the user model and the context relate to the interruption design?), *delegation* (When/how to delegate for the user?), and the *broader effect* (How do the user model and the assistant's actions affect others?). In the next chapter, we will thus implement stages within a calendaring assistant, while Chapter 5 also evaluates the aptness and utility of stages, as a new scrutability principle.

### 3.2.9 Summary of the Findings

As illustrated in Figure 3.6, our analysis has resulted in three principal sets of findings. The discussion reaffirmed that designing proactive IPAs in general poses challenges relating to interruption design, the realisation of multimodality, data stewardship, and user control. Even assistance focusing exclusively on time management was seen as onerous, as organising time is a deeply personal and context-specific endeavour. Individual time management – reflecting on how the user spends time, optimising plans, etc. – was seen as currently more graspable than advanced multiparty time management, which necessitates modelling of complex social relations to facilitate appropriate data obfuscation and appropriate scheduling. In both individual and multi-party time management assistance, there are unresolved challenges concerning data stewardship, interruption timing, and modality selection, which we will discuss in the next section.

In addition, scrutability and staging were assessed to be important aspects of envisioned human-assistant interaction, and these design elements are evaluated in a user study with non-experts in Chapter 5. The results of the expert discussion also yielded several potential use cases to be implemented in our calendaring assistant. Finally, our analysis provided a step towards guidelines on proactive general IPAs, continuing the work by Meurisch et al. [2017] and Zargham et al. [2022], and contributed towards a set of general principles for conversational interfaces, sought by Murad and Munteanu [2019]. In the next section, we elaborate further on the remaining open challenges.

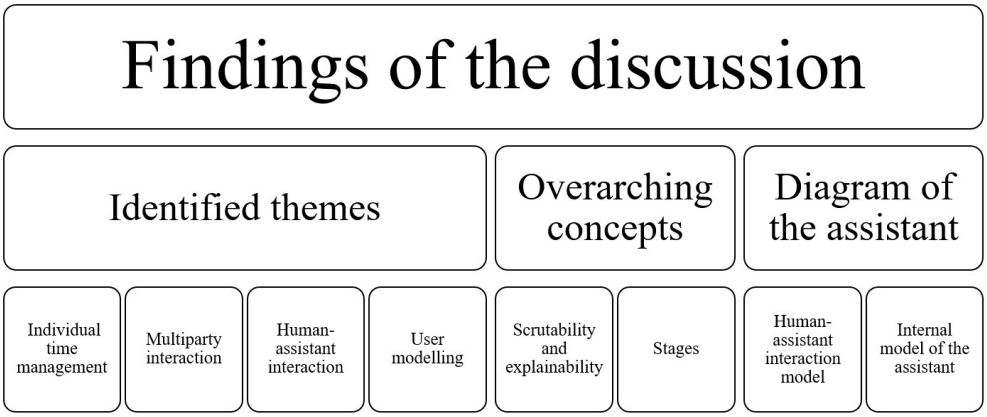


Figure 3.6: An overview of all findings from the expert discussion

## 3.3 Devising the Desired Design

As we have explored in this chapter so far, both the prior user studies and the experts envision the desired IPAs to be designed as proactive delegative systems capable of complex multimodal and personalised interaction made possible though comprehensive, yet steerable and scrutable user and context models. In this section, we organise these expectations as a set of interaction and modelling challenges, that are to serve as checkpoints for our IPA implementation in the following chapter. This section includes findings originally published in the ACM User Modelling and Personalisation (UMAP 2024) adjunct proceedings and presented at the adjacent Explainable User Modelling (exUM) workshop [Jeromela and Conlan, 2024].

### 3.3.1 Interaction Challenges

Prior user studies and our expert discussion outline IPAs capable of initiating and leading conversations with the user multimodally. The IPAs also envisioned that they might ultimately substitute the user in tasks such as meeting scheduling or task allocation negotiations. Moreover, the desired proactive interactions were seen as not taxing on the user's time and effort while also cognisant of the user's level of awareness of the system capabilities. Such use cases are novel from the scrutability perspective, and we should consider how they reflect on willingness to use. We categorise all interaction design implications of such complex use cases into six challenges detailed below.

#### Proactivity

Based on the use cases the experts in our study and the findings of Völkel et al. [2021], the envisioned proactive behaviour of time management assistants can be triggered by the following four cues:

1. The current time or scheduled events, such as alarms and appointment reminders;
2. the activities the user is engaged in, for instance if the user is busy, so the IPA chooses to postpone an engagement or because they are free so an opportune time for raising a non-urgent matter is identified;
3. the environment the user is in, i.e. whether an engagement is appropriate for the surroundings of the user; and
4. other people's needs, e.g. an attendance confirmation request.

The experts in our study suggested that the IPA may proactively nudge the user to scrutinise some of its decisions and models at opportune moments. Several prior studies, including by Zargham et al. [2022] and Kraus et al. [2020], had already looked into the acceptability of proactive agent behaviour, but further research is needed to establish how scrutability would affect the user attitudes towards proactive IPAs and the receptiveness towards "scrutiny nudges". As all these activity engagements are enabled by user and context models (to be discussed below), they should all be scrutable.

### Delegation

Having the IPA substitute the user in certain actions is another envisioned use case in both the reanalysis and the expert discussion. For instance, an expert proposed the IPA might delegate scheduling negotiations and event information dissemination, a use case similar to Calendar.Help [Cranshaw et al., 2017]. More challengingly, the experts also proposed for the IPA to take over task allocation communications (specifically, on agreeing whether they or their spouse would be the one to take their child to an extracurricular activity). Delegation presents a challenge for the facilitation of scrutability because, in providing the information, the IPA has to communicate how the decision was made and why it was judged to be most appropriate or fair while not revealing information that the other parties in the negotiation would not want to have shared. This suggests further work might continue from approaches that personalise explanations – as previously suggested by Chambers et al. [2022] – and adapt them for scrutiny of multi-user collaborative scenarios in a privacy-preserving manner.

### Modality Selection

Previously investigated scrutable applications relied primarily on a GUI, as indicated by the overview in Kay and Kummerfeld [2013] and the tools developed by Czarkowski [2006]. Contrastingly, the interviewed experts envisioned time management assistants as systems that would be able to interact with the user using both a GUI and a CUI, while simultaneously enabling a change of the response modality, depending on the content and context of the

communication. This adaptive multimodality poses a two-fold challenge for the realisation of scrutability. On the one hand, it should be determined how the IPA is to facilitate the scrutiny if the required information is more straightforward to display visually (e.g. how the assistant sees the user's availability over the upcoming period), but the prior communication was vocal. On the other hand, as experts further indicated, the user must be able to scrutinise and control what classes of information the IPA might proactively share vocally and under which circumstances. Otherwise, the user might be hesitant to share any sensitive or potentially embarrassing information with the assistant due to privacy concerns.

### Humanness and Expediency

Prior studies, including by Cowan et al. [2017], Doyle et al. [2019], and Luger and Sellen [2016], have looked into the perception and desirability of the humanness of IPAs (including the perceived personality, humour, and empathy). However, the IPAs considered in these studies typically have non-adaptive responses and are incapable of providing meaningful explanations as to why a particular trait was displayed. Allowing the user to scrutinise and control the "humanness" of the assistant may thus improve interaction satisfaction. Prospective (ante-hoc) methods, championed by Shneiderman [2022, Ch. 19], could be integrated here as well, allowing the user to set the desired level of humour and empathy of the IPA. More tightly related to the use case of time management is the expediency of the assistant's feedback. Dennis et al. [2012] reported that adapting feedback to the personality of learners might benefit their performance. Likewise, allowing the user to set their desired feedback (say, on a spectrum ranging from just giving the performance statistics to providing human-like praise) might support the expert-envisioned goal of facilitating and supporting reflection on time use.

### Disclosure of Capabilities

As was described in Subsection 3.2.8, the interviewed experts underlined the necessity for the assistant to disclose the extent of its capabilities accurately. Mishra and Rzeszotarski [2023] describe factors influencing user expectations of interactive interfaces, highlighting the role of user experience with the task at hand and program type, as well as the aesthetics of the interface. Similarly, Moore [2017] argues that the human-like aspects or the dialogue capabilities of the voice assistant should match the extent of its actual functionalities to give the user a fitting intuition. Contrastingly, state-of-the-art IPAs as well as chatbots based on large language models, such as ChatGPT, may lead the participants to overestimate their capabilities, as humanness sets unrealistic expectations [Cowan et al., 2017, Mishra and Rzeszotarski, 2023, Bender et al., 2021], amplifying a call for design interaction guidelines by Murad and Munteanu [2019]. Our analysis of the expert discussion also indicated that an accurate view

of the system's capabilities (or the user's model of the assistant) is tightly related to usage willingness and the user's openness to "forgive" the IPA's mistakes. This finding was reported by Luger and Sellen [2016] and Cowan et al. [2017]. Thus, further studies might look into how scrutability can aid the user in understanding the IPA's capabilities.

#### Effort-Efficiency

In prior studies on scrutable applications – e.g. on text editors by Kay and Kummerfeld [2013] or recommender systems by Musto et al. [2020] and Alkan et al. [2021] – the only practical way to complete the task was to use the provided application. In contrast, with time management assistants or IPAs in general, the user can always do the task "manually", without any additional assistance. Given that scrutability is a prerequisite to user control but requires that the user invests time in understanding the system first, it remains to be investigated how scrutability affects usage willingness. Thus, as underlined in Subsection 3.3.2 in reference to findings by Völkel et al. [2021], it might be worthwhile investigating how scrutable IPAs can be effort-efficient, that is, how to achieve that their use is principally less time-consuming manner for the user than just doing all tasks without them. In that respect, the benefits scrutability bears should be leveraged by the fact that it is a time-consuming, user-led process.

### 3.3.2 User, Context, and Risk Modelling

In defining the modelling challenges, we were led by the vision originally championed by Kay [1999], of scrutable user models that allow the user to investigate and control the manner in which the data is captured, stored, and used to facilitate personalised services, which fits with the envisioned proactive IPAs for time management. We do not mould the conceptual framing of these modelling challenges towards specific technologies or existing approaches. Instead, we defined the three challenges detailed below that need to be addressed for the IPA and its models to fit our emerging vision of scrutable-by-design assistance.

#### Meaningful Control

Kay and Kummerfeld [2013] illustrated the importance of user control over the models and correcting found errors in systems such as text editors and eLearning platforms. However, the experts in our discussion and the participants in the studies we reanalysed, envision much more potent systems capable of proactively modifying events in the user's calendar, partaking in negotiations, or spending money to book tickets. This potency and the extent of possible consequences reinforce the need for apt user control of IPAs. The IPA frameworks we described in Subsections 3.1.5 and 3.2.6, as well as the framework put forward by Meurisch et al. [2017] also envisioned that the IPA would model not only the user, but also the surrounding context

and the potential risk of its actions, further research should investigate how scrutiny and control of these models can be facilitated. A metric to consider in this pursuit is user trust, which was previously shown to be affected by factors including predictability and understandability [Kraus et al., 2020, Glass et al., 2008, Tolmeijer et al., 2021, Olson and Xu, 2021].

Should the IPA be able to negotiate with others, it would have to model risk for the user should it come up with an arrangement that is suboptimal for them. According to this risk model, as well as through the inferred or explicitly stated user preferences in the IPA's user model, the assistant is then to decide if it can proceed with the arrangement without consulting the user beforehand. Furthermore, as indicated in the outline of the proposed IPA's internal design, this control ought to extend to all models that affect the IPA's services.

### Underlying Technology

The ability of the user to scrutinise the IPA's actions is a prerequisite to enabling user control [Kay and Kummerfeld, 2013]. However – as we had reviewed in Subsection 2.3 and will discuss in Chapter 7 – some forms of machine learning algorithms are inherently opaque. How, or even if, such machine learning systems can be made scrutable is a matter of active research (cf. also Section 6.2 for final deliberations on Scrutable-by-Design principles in an era of opaque models). In that pursuit, we see value in both the algorithm-first ("opening the black box" explanations) as well as the human-centric approach, which aims to understand user perception and expectations. To this end, a meaningful aim may be to define a "minimal" set of requirements for interpretability that an IPA must facilitate for it and its models to be considered scrutable.

### Data Stewardship

Both from the sources in this chapter and from further studies, including by Tabassum et al. [2019] and Bolton et al. [2021], we can identify concerns regarding the manner in which IPAs collect, store, and share their data. Tabassum et al. [2019] found users to be more willing to share a conversation with an assistant if the content of the conversation was less sensitive and the service provided in return more tangible. As detailed in Subsection 3.2.4 and expressed through a user study quote by Tabassum et al. [2019], time management data is inherently sensitive because it deals directly with one's daily schedule and social contacts. To attempt to remedy this concern, the idea of *stages* is introduced in Subsection 3.2.8. Each stage refers to a set of IPA functionalities that is explicitly connected to a set of specific data licences, which allow the assistant to collect and use a class of data to facilitate these functionalities. We hypothesise that facilitating stages might comprise a novel scrutability principle, mitigating data stewardship concerns while facilitating user control and revealing

the extent of IPA capabilities. We thus ought to investigate whether stages affect the user's willingness to allow the assistant an insight into their personal or professional schedules.

## 3.4 Chapter Summary

We conducted a purpose-oriented reanalysis of the user study data originally collected and presented by Cowan et al. [2017] and Völkel et al. [2021]. We found numerous instances of study participants expressing a desire for personalised and proactive assistance, as well as for the ability to scrutinise the inner workings of the assistant.

Based on the literature reviewed to date and the reanalysis, we presented our first diagram of the envisioned interactions and inner modelling of the anticipated proactive IPAs (Figure 3.1). In the envisioned interaction, the IPA would facilitate communication for the purposes of user scrutiny and control, in addition to task-based and social communications.

To better understand the anticipated opportunities and concerns of the anticipated proactive IPAs within the domain of time management (a use case chosen as complex enough to uncover paradigmatic user perspectives, while less encompassing than general-purpose IPAs), we conducted and analysed a workshop-style discussion with seven experts in fields adjacent to human-computer interaction (HCI).

Among the main contributions of the study is an experts-devised summary of anticipated use cases for proactive IPAs for time management and a revised diagram (Figure 3.4) of the anticipated human-assistant interaction with an outline of the inner modelling structure. The discussion also raised an idea for coupling data use licences and specific features in discrete stages as a potential new principle of scrutability.

Based on the research to date and the outcome of the expert discussion, we formulated a set of six interaction challenges for advanced IPAs (proactivity, delegation, modality selection, humanness and expediency, disclosure of capabilities, and effort-efficiency) as well as three challenges regarding the realisation of the IPA's models (meaningful control, underlying technology, and data stewardship).

# 4 SIPA4TM: Scrutable Intelligent Personal Assistant for Time Management

This chapter details our implementation of SIPA4TM – our scrutable calendaring assistant – and our user study protocol. We then detail how the implemented assistant and the protocol together investigate the challenges and design requirements from the previous chapter. Specifically, Section 4.1 broadly defines the scope of our implementation, necessitated by the technical and temporal constraints of a PhD research. Section 4.2 then provides key technical details regarding the implementation of SIPA4TM, which was to serve as the focal point of the user study. Section 4.3 details the user study protocol we devised, which enabled us to investigate the role of scrutability in both the implemented assistant and the more advanced AI systems that may offer services beyond the scope of our implementation. Finally, in Section 4.4, we elaborate on ethical considerations regarding the SIPA4TM implementation and the process of obtaining ethical approval from the Trinity College Dublin Ethics Application System.

## 4.1 Scope of the Assistant

As shown in Figure 4.1, the assistant we implemented – dubbed Scrutable Intelligent Personal Assistant for Time Management, abbreviated as SIPA4TM – offers assistance for individual time management only, in order to maintain the feasibility of the technical implementation. This decision aligns with the expert discussion findings (summarised in Subsection 3.2.9) that individual time management may be considered separately from assisting multi-party interactions, given mostly distinct sets of challenges and use cases. Crucially, multipersonal time management is much more culturally and ethically complex than individual assistance to be implemented meaningfully. Thus, that aspect of assistance and its implications for scrutability are assessed through the discussion queries, where the users are asked to extrapolate their

experience with SIPA4TM, rather than being implemented as actual automated features.

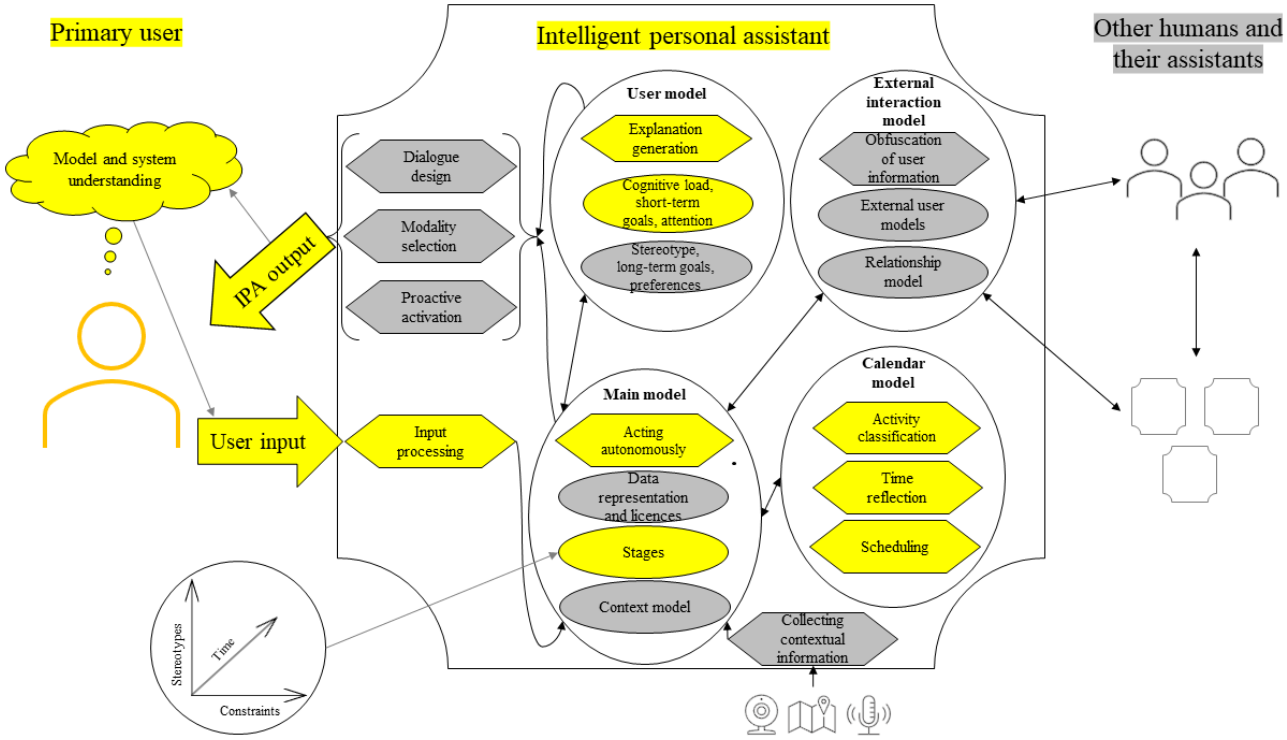


Figure 4.1: A coloured diagram from the expert discussion described in Subsection 3.4; yellow elements are researched through corresponding features of SIPA4TM; the grey elements are out of scope of implementation and are addressed through questions in the participant discussion protocol

With regard to user input, we bound our approach to a click-and-type approach, in which the user controls a desktop calendaring program with the integrated assistant using a mouse to select appropriate actions. The actions are selected from context-sensitive menus and clickable calendar elements. Where appropriate, the user can also type in relevant information, such as event details or user model values. This scoping of user input also made input processing implementation much more practical, by reducing it to appropriately incorporating the user's selections and data entries into the assistant's models.

Conversely, IPA output is limited to textual suggestions. The suggestions themselves are proactively shown (without the user explicitly requesting them), depending on the user model, the selected day, the selected stage, and context. As we detail below, suggestion creation is based on rules that take into account the application context, calendared events, and the user model. In this manner, we kept the creation logic transparent and reproducible. The rules take into account different activity classes, the schedule of the day and the week, and (in the highest stage) support time reflection.

As an alternative to such scoping of the input and output, we also considered implementing a

chat-based control, potentially utilising large language models to parse user intent or provide personalised suggestions. However, as we had discussed in Subsection 2.4, LLM-based interaction design is a subject of ongoing research, especially when it comes to achieving appropriate user reliance on the system. As our focus was to investigate user perspectives on scrutability and its for the user's model and system understanding, we aimed to have a user interface that is appropriate for the tasks at hand without being unnecessarily complex, potentially detracting users. The click-and-type input and text-based output made this goal much more traceable by reducing the need to decide on the most appropriate menu hierarchy and when to provide context-relevant additional information.

As elaborated in Subsection 3.2.8, the experts also proposed onboarding and control in stages alongside scrutability might help remedy both the misalignment of user perception of the system capabilities as well as the concern about data use. Therefore, we also aimed to have a meaningful implementation of stages, enabling us to investigate the non-expert perspective regarding the utility and appropriateness of stages as a potential scrutability principle.

Acting autonomously is limited to proposing event changes and alternative event times. Notably, SIPA4TM does not consider data representation and data-specific licences (beyond stages) and long-term goals; however, the questions in the discussion protocol do address them.

SIPA4TM does not support natural-language or multimodal conversations and is intended for use during a single interaction session. For that reason, it does not directly address challenges of dialogue design, modality selection, and proactive activation. As with the delegative features, these aspects of the assistance are addressed through questions in the focus group instead.

By contrast, IoT-related assistance and sensor-based detection of the user's environment were deemed outside the scope of the dissertation, as they were tangential to time management. Table 4.1 provides an overview of the scope of our assistant implementation and the extent of consideration addressed through protocol design, while the sections below provide more details on their actualisation.

## 4.2 SIPA4TM Implementation Process

The calendaring assistant we implemented was named after its principal research goal and task domain – Scrutable Intelligent Personal Assistant for Time Management, SIPA4TM for short. SIPA4TM is a locally-run desktop calendaring assistant, with persistent data storage for events and suggestions, logging for debugging and data analysis and a Google Calendar API integration for optional event data syncing or overwriting, enabling easy data restart between

Table 4.1: Design challenges mapped to implementation and study protocol

| Challenge | SIPA4TM implementation | Questions in the study protocol |
|---|---|---|
| Proactivity | Context and action-based triggering of time-management suggestions | Comfort with proactive calendar changes and proactive delegation; expectations around notification before any changes |
| Delegation | Not implemented directly | Trust in allowing the assistant to contact others on the user's behalf; willingness to invest effort in controlling such delegation |
| Modality selection | Primarily text-based interaction, with visual explanations at the second level of scrutiny | Preferences regarding voice vs. text interaction; opinions on voice assistants; attitudes towards avatars |
| Humanness and expediency | One-size-fits-all textual suggestions, with personalisation driven by user-model values | Comparison of tone and style between SIPA4TM suggestions, voice assistants, and LLM-based chatbots |
| Disclosure of capabilities | Scrutability via three levels of explanations and an explicit, transparent user model | Behaviours that would prompt users to question how a system works; approaches to learning about system capabilities |
| Effort-efficiency | Users may ignore suggestions they do not find helpful; exploration of explanations and the user model is optional and is logged | Perceived task facilitation; liked, disliked, and desired features; willingness to integrate SIPA4TM into everyday calendaring tools |
| Meaningful control | All user information is explicitly controllable via the user model, supported by explanatory guidance | Opinions on the explicit user model; perceived control in voice assistants, LLM-based chatbots, and other AI-based products |
| Underlying technology | Interpretable, rules-based system | Interest in how AI systems work; perceived importance of personalisation; attitudes towards recommender systems in daily life |
| Data stewardship | Stage-based permission control and scrutable data usage for each suggestion type | Prior experiences with data retrieval; opinions on memories in LLM-based chats; attitudes towards data control in stages and online privacy concerns |

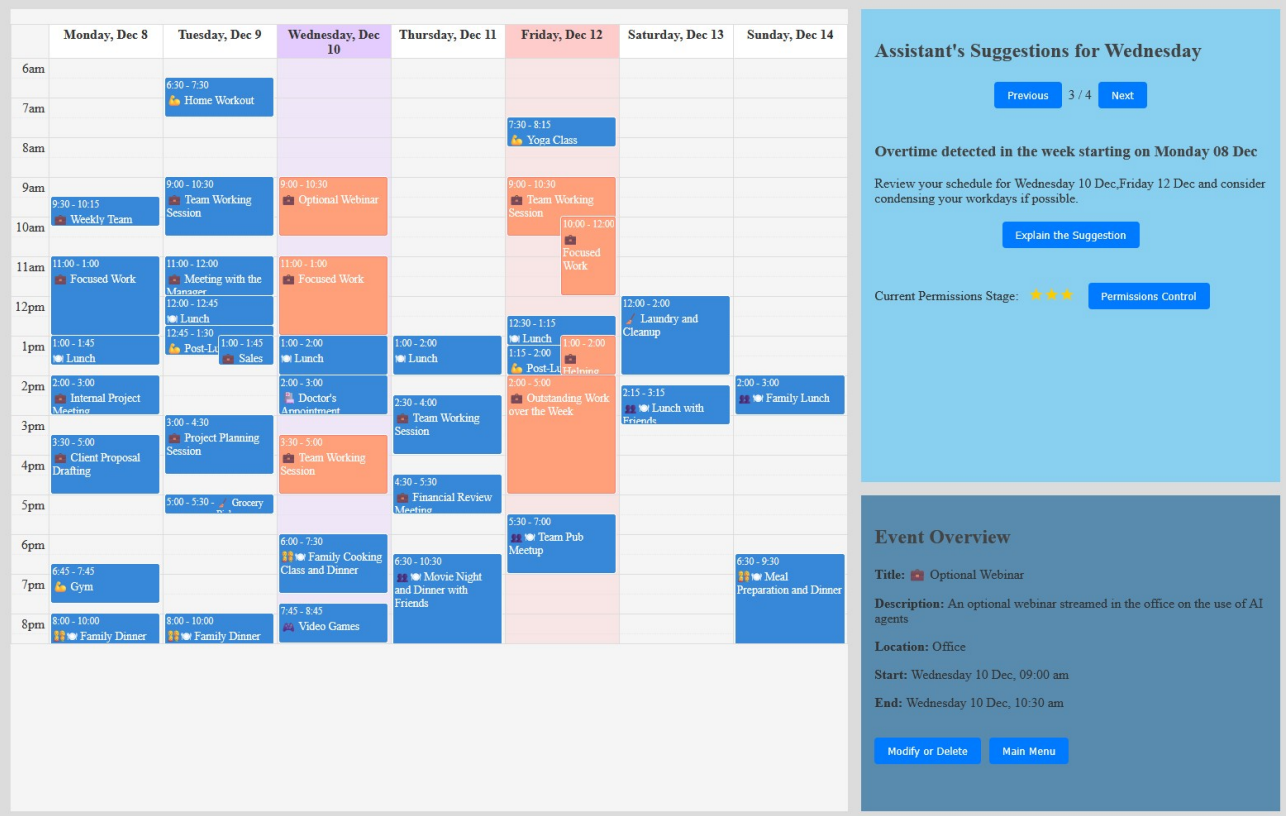


Figure 4.2: A screenshot of SIPA4TM with its three main components visible: the calendar view on the left-hand side; the assistant component in the upper-right corner; and the control and context component in the lower-right corner

users.

As discussed in Section 4.1, SIPA4TM is a rule-based calendaring assistant that offers suggestions for adapting events within the upcoming week to better align with the user's preferences stored in an underlying explicit user model. The main interface and the three components it comprises can be seen in Figure 4.2. The most prominent of the three components is a calendar view with the user's events for the upcoming week. On the right-hand side are the two components whose content changes dynamically, with the upper one displaying the assistant's suggestion, suggested explanations, and stage control, and assistant-related options, whereas the lower component is where additional contextual information about the selected event appears and which also hosts most of the buttons and fields through which the user controls the program.

Notably, the calendar view is dynamic, reacting to both the user's clicks and the assistant's suggestions. The day the user clicked is highlighted blue if there are no suggestions for the day and purple if there are. If the suggestion also affects events on other (unlicked) days that week, those days are highlighted in light red. Lastly, the events affected by the assistant's

suggestions are coloured orange (rather than dark blue). Additionally, the event the user has clicked and for which additional details are provided (or which is being edited) has a red frame around it.

In line with our research question and the guideline championed by Kay and Kummerfeld [2013], we designed SIPA4TM to be scrutable-by-design. For every suggestion, the user may click the "Explain the suggestion" button to scrutinise why it was displayed, as shown in 4.3. Specifically, on the first click, a high-level textual explanation was generated, with the option to go back or to explore further. The short explanation is rather general and conveys the general idea of each suggestion, as can be seen in Table 4.2. Should the latter option be chosen, a visual explanation would be provided, along with options to either return to the textual explanation, exit scrutiny mode, or explore further. The final level of scrutability then showcases which specific argument in the assistant's user model can be changed to alter future suggestions of the same type. At every level of scrutability, the user can also open the full user model to view what the assistant knows about their preferences and, optionally, change a preference. The fields the user model contains depend on the permission control stage, but since the suggestions are also tied to stages, the fields affecting the scrutinised suggestion are always visible. Notably, the design of levels of explanation and the broader architecture were inspired by the design guidelines by Kay and Kummerfeld [2019], as detailed in Subsection 4.2.1.

Table 4.2: Scrutiny Levels and Explanation Types for Each Suggestion Category in SIPA4TM

| Suggestion Type | Level 1 (Base Explanation) | Level 2 (Graphical Visualisation) | Level 3 (User Model Field) |
|---|---|---|---|
| **New Event Placement** | Indicates that the assistant highlights the largest free block of time (outside sleep hours) suitable for new events. | Plots free blocks and highlights the longest viable one, alongside sleep boundaries. | Returns sleep fields and `default_commute_duration`. |
| **Modify Event Placement** | Explains the available wiggle room around an existing event, respecting sleep boundaries. | Plots the modified event, nearby free blocks, and sleep boundaries. | Returns the same user model fields as new-event placement (sleep fields and commute duration). |

Continued on next page

Table 4.2 – Continued from previous page

| Suggestion Type | Level 1 (Base Explanation) | Level 2 (Graphical Visualisation) | Level 3 (User Model Field) |
|---|---|---|---|
| **New Event Creation Warning** | Indicates missing or invalid data (title, location, timing) when creating an event. | Plots proposed event placement and any overlaps with existing events. | Stage 0; no user model data used. |
| **Event Modification Warning** | Indicates invalid data or inconsistent modification of an existing event. | Plots modified and original versions of the event alongside nearby events. | Stage 0; no user model data used. |
| **Clash / Chain Clash** | Describes that two or more events overlap in time, based solely on start and end times; may be intentional or accidental. | Shows a horizontal bar chart of overlapping events. No user model data used. | Indicates a Stage 0 suggestion; no user model fields referenced. |
| **Transit Short** | Explains insufficient travel time between two events, based on event times and commute durations. | Plots both events, with a shaded region representing the minimum commute time (from user model or stored distances). | Lists `default_commute_duration` and, when available, stored location-to-location travel times. |
| **Sleep Time** | Explains that an event conflicts with the user's preferred sleep schedule (for weekdays and weekends). | Plots the event relative to bed and wake times using vertical lines for sleep boundaries. | Returns sleep preferences: `workday_bed_time`, `workday_awake_time`, `weekend_bed_time`, `weekend_awake_time`. |
| **Duration Short** | Explains that an event is shorter than its minimum modelled duration (hobby or chores). | Plots actual duration vs. required minimum duration. | Returns either `hobby_min_duration` or `chores_min_duration`. |

Continued on next page

Table 4.2 – Continued from previous page

| Suggestion Type | Level 1 (Base Explanation) | Level 2 (Graphical Visualisation) | Level 3 (User Model Field) |
|---|---|---|---|
| **Busy Day of the Week** | Explains that a day in the upcoming week exceeds the user's tolerance for busyness. | Plots event counts per day with a dashed threshold (median + tolerance). | Returns `busy_day_tolerance`. |
| **Overtime** | Explains that work-related events exceed the user's preferred maximum daily or weekly work duration. | Plots all work events and a vertical threshold representing the user's maximum work duration. | Returns `work_duration`. |

As shown in Figure 4.5, the user model comprises three possible sets of preferences, each corresponding to a permission control stage, as elaborated below. Should the user want to change any fields relevant to the current stage, they may simply click the current value in the assistant component. This then refreshes the contents of the context component to provide a high-level description of what suggestions editing the user model field affects, acceptable value and the current value (which was also visible alongside other fields and values of the user model in the assistant component). The user may then change the field value, either by selecting days for fields such as "Office weekdays", selecting a time for time-based fields such as "Weekend bedtime", or entering an integer for "Chores Min Duration" and other numeric fields in the user model.

Permission control stages are explicit pairings of suggestion types and the data classes they rely on. The concept was originally proposed by the experts (cf. Section 3.2). The current stage is represented by one to three stars (seen under each suggestion the assistant offers). Should there be no suggestion for the current stage, the assistant would also advise the user to increase the stage by clicking the stars. Should the user click the stars, they would see a window, as in Figure 4.4, with a description of the current stage and guidance on how to change it.

SIPA4TM was implemented with two main goals in mind. First, it was intended to present users with a clear understanding of the type of calendaring advice envisioned by some prior studies (including by Cowan et al. [2017] and Völkel et al. [2021] that we reanalysed in Section 3.1). The assistant was also implemented to embody scrutable-by-design user modelling and thus serve as an inspiration for focus group discussions. Second, the assistant was to serve as a quantitative measure of the actual user interest in investing their effort to scrutinise and understand why advice was given and on the basis of what data. For that reason, logs were

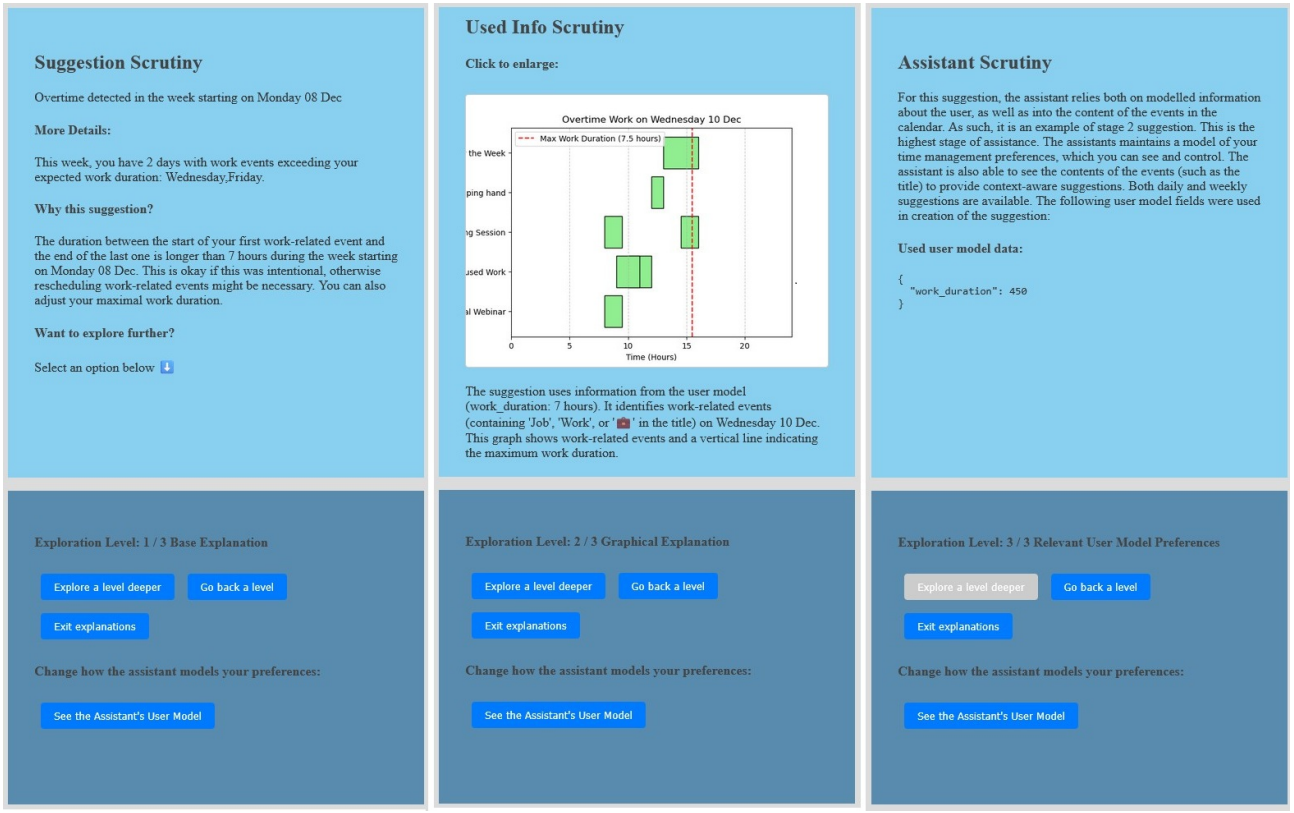


Figure 4.3: Three screenshots assembled to showcase the three levels of explanations: the first level is a short text-based explanation, the second is a visualisation with more details, and the final level exposes user model fields relevant for the scrutinised suggestion. The explanations are shown in the assistant component, and the user controls the scrutiny and may enter the user model to change the underlying values through the context component underneath the assistant component.

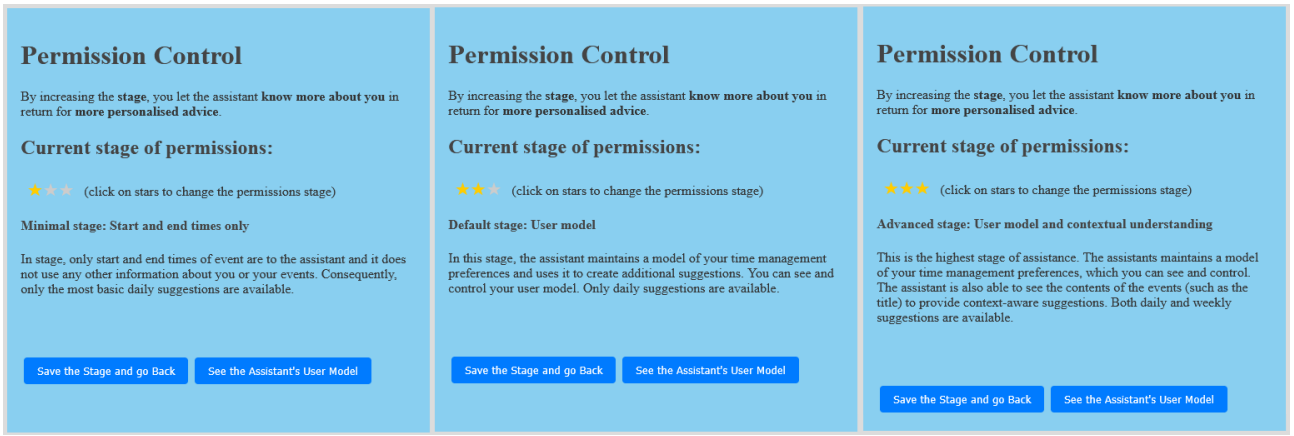


Figure 4.4: The view of the permissions control, displayed in the assistant component for the three stages, visualised by stars and described through a high-level description

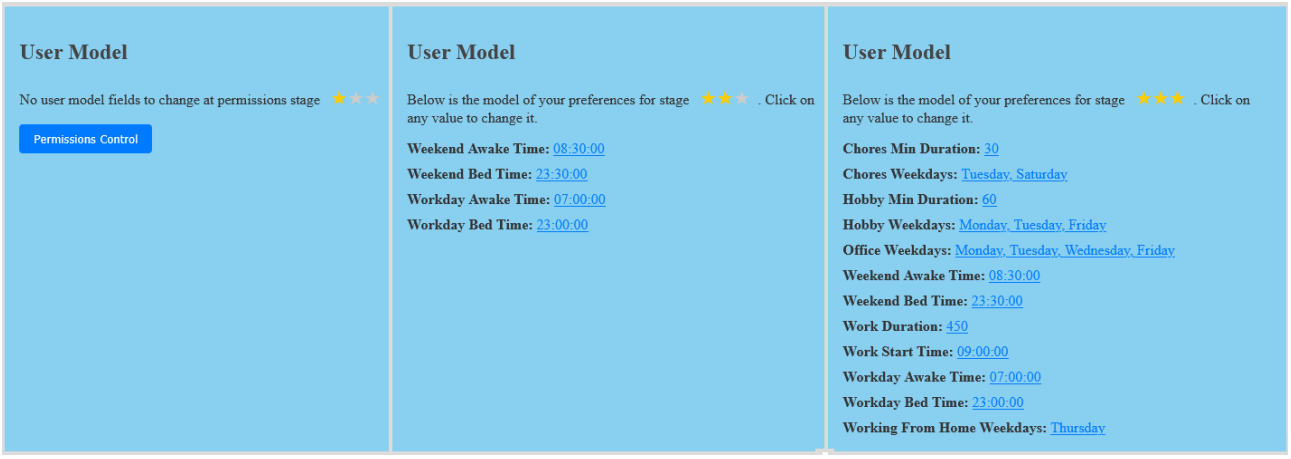


Figure 4.5: The assistant's contextual user model grows with the selected permissions control stage: for the lowest stage (★☆☆), user preferences are not saved or considered, at the middle stage (★★☆) the assistant only tracks the user preferred sleep pattern and commute times (the latter is not displayed), and at the highest stage (★★★) the assistant also tracks durations for different event categories

created during the interaction to record when users clicked the explain button, the level of scrutiny they selected, the stage they used at different times, and whether they changed user model values at all. The subsections below elaborate on additional technical and architectural details of the implementation process and provide a brief summary of the software engineering design patterns employed, which make the system extendible for potential future user studies. We also share practical lessons learnt from the implementation process that may benefit future related research endeavours.

### 4.2.1 Implementation Architecture

At a high level, SIPA4TM follows a client–server architecture, illustrated in Figure 4.6. The frontend is a single-page web application built in Vue and a calendaring library exposing a week-view of the user's schedule. The backend is a Flask application with a relational database accessed via SQLAlchemy models. All of the assistant's internal state – including user models, suggestions, and interaction logs – is persisted on the backend. The frontend holds only a lightweight representation of the current view (including selected date, which suggestion is currently shown), synchronised through a small number of API endpoints (such as "get suggestions for this date and stage", "update user model", "commence scrutiny process of the suggestion"). This separation was an important methodological step: it ensured that the experimental logic and logging could be inspected and reproduced independently of the client UI.

The suggestion engine itself is entirely rule-based in line with the scrutability principles by

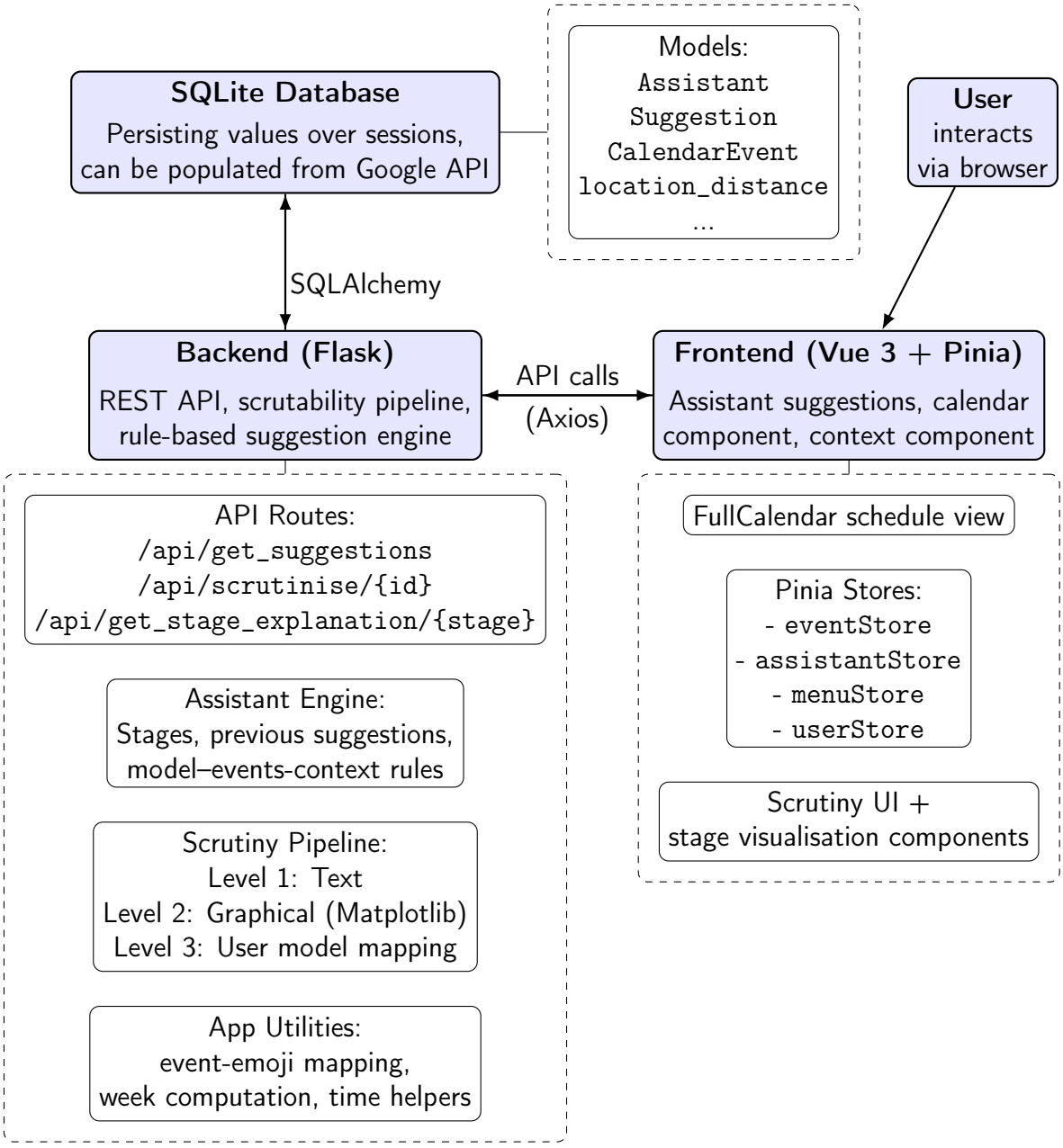


Figure 4.6: System architecture of SIPA4TM, utilising Vue 3 with Pinia frontend, Flask backend with scrutability pipeline and rule-based assistant modules, and an SQLite database storing user model, suggestions, and events.

Kay and Kummerfeld [2013] (cf. the related discussion in Section 6.2). All suggestion types are contained in Table 4.3. It operates on calendar events and a structured user model, and produces suggestions of different types. Suggestions include clash detection, short transit between events, overtime, sleep-time violations, and event-manipulation checks (as previously detailed in Table 4.3). Internally, each suggestion is represented as a database record with a type, a title, a free-text description, a recommended action, and denormalised links to the affected events and days. When the user selects a day or modifies/creates an event, the

backend computes a set of "candidate suggestions" for that context. These candidates are then passed through a deduplication and refresh layer that ensures suggestions are relevant to the context and prevent the recreation of semantically identical suggestions (such as when a user clicks the same day twice). If a matching suggestion already exists but has expired, it may be "refreshed", whereby its status is returned to active and, if needed, its text and affected events are updated, rather than creating a new entry. Only genuinely new suggestions are added. This mechanism gives SIPA4TM a form of memory: it can "remember" that a particular clash or odd pattern has already been pointed out, and queues such suggestions after the genuinely new ones.

Table 4.3: Descriptions and proposed actions for all SIPA4TM suggestion types in the final implementation

| Suggestion Type | Description | Proposed Action |
|---|---|---|
| **Clash / Chain Clash** | Clash found: <titles of events that clash> with intervals <start-end per event>. | Click to modify <any clashing event (id <id>)> and the assistant will show suggested new times at higher scrutiny levels. |
| **Transit Short** | Insufficient travel time between <event A> (<end time>) and <event B> (<start time>): available <gap minutes> vs. required <commute time>. | Consider rescheduling <event A (id)> or <event B (id)> to allow at least <commute time> minutes for travel. |
| **Sleep Time** | <event> interferes with the user's <workday/weekend> sleep schedule: event runs <event start-end> vs. bedtime <bed time> and awake time <awake time>. | Consider rescheduling <event (id)> to avoid conflict with the sleep window. |
| **New Event Placement** | Available time blocks on <date>: <list of free time blocks>. The assistant identifies all valid free periods between wake and sleep times. | Schedule a new event on <date> during the largest available block <start-end> for up to <maximum hours> hours. |

Continued on next page

Table 4.3 – continued from previous page

| Suggestion Type | Description | Proposed Action |
|---|---|---|
| **Modify Event Placement** | Shows the earliest and latest times <event> may start-/end without interfering with <neighbouring events> or the sleep window. | Reschedule <event> to start no earlier than <earliest allowable> and end no later than <latest allowable> on <date>. |
| **Duration Short** | <event> lasts <actual duration> minutes but the minimum duration for <Hobby/Chores> is <required minutes>. | Extend <event> by at least <required - actual duration> minutes. |
| **Busy Day of the Week** | <Day names> have significantly more events than the weekly median <median>, exceeding the user's tolerance <busy_day_tolerance>. | Reschedule selected events from <most busy day> to <least busy day> to balance the workload. |
| **Overtime** | Work-related events on <dates> exceed the user's preferred maximum daily or weekly work duration of <work_duration> minutes. | Review the schedule for <affected dates> and consider condensing or redistributing work events. |
| **Event Manipulation: Missing Data** | Required fields are missing: <title/location-/description/id>. [Description may be missing as a minor issue; other fields block completion.] | Add missing <element> or reset to <previous value>. For minor issues, click Save again to proceed. |
| **Event Manipulation: Impossible Times** | Start time is at or after end time: impossible duration <start-end>. | Reset times to <old start-old end> or enter a valid time range. |
| **Event Manipulation: Unusual Duration** | The event is either too short (<15 minutes) or too long (>24 hours): duration <actual duration>. | Correct the duration or click Save again to accept the unusual value intentionally. |

Continued on next page

Table 4.3 – continued from previous page

| Suggestion Type | Description | Proposed Action |
|---|---|---|
| **Event Manipulation: Outside Calendar Hours** | Event occurs partially or fully outside allowed calendar hours: <start-end> vs. allowed <earliest-latest>. | Reset to <old start-old end>, enter new times within allowed hours, or click Save again if partially outside. |
| **Event Manipulation: Date Changed** | The new event date <new date> differs from the old date or the selected date <selected date>. [Minor but requires confirmation.] | Revert to <old date> or intentionally confirm by clicking Save again. |
| **Event Manipulation: Duration Changed** | The duration changed from <old duration> to <new duration>. [Minor but requires confirmation.] | Reset duration to <old duration> or confirm by saving again. |

A distinct set of rules governs event manipulation. When the user is in a "New Event" or "Modify Event" context (visible in the lower right-hand side window in Figure 4.2), the assistant checks for proactive suggestions for data entry and also verifies the proposed event data before allowing it to be saved. These checks cover, for example, missing required fields (such as title or location), impossible times (end before start), unusual durations (very short or extremely long events), dates that were moved off the selected day, and events placed outside the user's typical waking hours. Each of these conditions can trigger an "override" suggestion that temporarily blocks the action until the issue is either corrected or (if minor) explicitly acknowledged by the user.

The permission control is integrated into this suggestion logic and represents a reification of the idea of feature-data coupling in stages (cf. Subsection 3.2.8). In the final implementation, each stage corresponds to a specific configuration of which data the assistant is allowed to use and which classes of suggestions are active, as shown in Figure 4.4. Stage 0 (graphically represented as ★☆☆) restricts the assistant to non-personalised suggestions, i.e. clash-checking and input-completion verification for event creation and modification. Stage 1 (★★☆) offers more advanced daily suggestions that account for commute times between locations and the user's bed and wake-up times. Stage 2 (★★★) offers all implemented daily and weekly suggestions. In stage 2, the assistant also considers the event type to determine whether the duration constraints in the user model apply (e.g., typical working hours, minimum durations for hobbies or chores, work time).

The decision to represent stages using the star-based visualisation was made after a preliminary round with friends and colleagues, where the previous numerical system with (a plus and a

minus button) was reported to be unintuitive to understand and change. The star-based visualisation is interactive wherever it appears – clicking on the stars opens the permission control view and, within that view, the user can click to raise or lower the stage. When no suggestions are available at a given stage, SIPA4TM explicitly informs the user and suggests increasing the stage, but only as an option, not as a requirement. The idea the design comes from Shneiderman [2022, Ch. 19], who suggested using slides as ante-hoc explanations of what users can expect from an AI system. Specifically, we were inspired by Figure 19.7 in the book, showcasing sliders for music control recommendations originally developed by Millecamp et al. [2018]. A similar design has also been developed by Bemmann et al. [2024], who found that slider-based permission controls improve the perceived levels of system transparency and user control.

In parallel, SIPA4TM includes a scrutability pipeline that is tightly coupled to both suggestions and the user model. For each suggestion displayed in the interaction pane, the user can click an "Explain the suggestion" button. This enters a three-level scrutiny mode. The first level presents a concise textual explanation of why that specific suggestion was triggered, in natural language closely tied to the rule that fired (for example, pointing out that an event would overlap with sleep, or that the commute time between two locations is too short given the user's default commute duration). From there, the user may drill down into a visual explanation at the second level. The presence of a graphical explanation is inspired by the expert discussion's suggestion that some explanations might require different modalities. Moreover, we wanted to see whether the users would invest effort in exploring the presented graphs. At the deepest third level, the assistant exposes the link between the suggestion and the underlying user model parameters: it makes explicit which particular preference (for example, "minimum duration for chores" or "wake-up time on workdays") drives this type of suggestion and offers links to change that parameter for future recommendations. Throughout scrutiny mode, the user can always exit, step back up a level, or jump into the full user model view.

The permission control interface and the scrutability pipeline design were implemented to follow the design guidelines for interfaces that allow learners to control and scrutinise their models by Kay and Kummerfeld [2019]. Specifically, the multi-layered, complexity-increasing approach to explanations represents a scaffold for the user to scrutinise and, where appropriate, correct the model ontology, evidence, and reasoning (corresponding to meta-cognitive aspects M5 and M6 in the paper). Similarly, the stages as implemented are also meant to scaffold the user into adding and taking access from the system (corresponding to the control-aspects C in the paper). Lastly, the always-visible star-based stages indicator can be seen as a minimalist dashboard indicating what the user can change to get more advanced assistance and improve their calendar further (meta-cognitive aspects M4 and M5). From a practical standpoint, distinct levels of explanation also simplify the interpretation of interaction logs, as we have a

discrete depth the users went to while exploring the suggestion, thereby also indicating their level of interest and the level of effort they invested in it.

The user model in the control view is implemented to be transparent, listing all fields and values the assistant maintains about the user, as shown in Figure 4.5. The interface lists individually named fields (e.g., "Workday bedtime", "Default commute minutes", "Days you go into the office") with short descriptions and expected values. Clicking on any field name would allow the user to change the value. We were aware that it is not the standard nowadays, but rather that the recommendations are provided in an inscrutable manner that the user typically cannot explicitly change, as elaborated on by Tintarev and Masthoff [2022]. So in an attempt to give users an intuition of the affordance of control, we made times editable with time pickers, durations via numeric fields in minutes, and weekly patterns via checkboxes for days of the week, with a description of expected values. When the user attempts to save changes, the frontend sends an updated version of the current user model to the backend, where a schema and additional parsing logic validate the values (for example, ensuring that weekday names are valid, durations are non-negative and within a logically possible bound, and times are well-formed). If a value fails validation or cannot be parsed, the backend responds with an error; the frontend catches this, keeps the user's input visible, and displays a clear error notification rather than silently discarding or partially applying the change. In this manner, we aimed to implement user control, as the final element of scrutability, whereby users can experiment with their preferences and receive immediate, intelligible feedback when something is inconsistent and then see the change in suggestions accordingly.

To support the evaluation, SIPA4TM also incorporate logging of all user clicks and the interaction state – such as displayed events, new event information, displayed suggestions – as well as whenever the user would engage in scrutiny, such as when clicking on the "Explain the suggestion" button, when changing the scrutiny level, when interacting with the user model, or the stages. In the backend, the assistant tracks, for each suggestion, how many times it has been shown to the user and prioritises the less seen ones, and the frontend advances a suggestion iterator as the user presses "Previous" and "Next" to browse through available suggestions for a given day. Because these behaviours are implemented explicitly in the backend rather than left to an opaque learning model, they are amenable to precise instrumentation: every step between "calendar state" and "assistant output" can be reconstructed from logs and thus serve as the quantitative data in our user study.

Overall, the implementation of SIPA4TM operationalises our research question in three concrete ways. First, it demonstrates a bounded yet realistic style of proactive time-management assistance, where suggestions are dynamically created based on realistic calendar data and preferences stored in a user model. SIPA4TM thereby aims to provide an intuitive insight

into the types of time management assistance IPAs might provide. Second, in line with the scrutable-by-design maxim, SIPA4TM embeds scrutability directly into the interaction flow. Users can not only see that advice is being given, but also why, on the basis of which internal variables, and with which data-functionality alternatives available. In combination with the logging mechanism, this implementation serves to quantify how much effort users are willing to invest in understanding and shaping their assistant. Third, SIPA4TM reifies the idea of permission control in stages, integrating data stewardship control into interaction design. The stages structure the user model, determine which data may be used, filter the suggestion repertoire, and serve as a reification of the privacy-utility trade-off.

### 4.2.2 Maintainability and Extendibility

The implementation of SIPA4TM followed well-established software design patterns to support modularity, extensibility, and long-term maintainability, as originated by Gamma et al. [1994]. An overview of software design patterns used is provided in Table 4.4.

Table 4.4: Summary of software design patterns used in the implementation of SIPA4TM

| Design Pattern | Components | Purpose |
|---|---|---|
| Factory | Backend component creation, such as assistant construction | Centralised logic for creating or retrieving (assistant) instances; simplifies future additions to the SIPA4TM's architecture. |
| Singleton | Database access, database constraints, logging | Ensures that there is at most one assistant per user per user–stage pair and enables consistent logging across multiple uses. |
| Strategy | Independent suggestion-generation functions | Each rule is encapsulated as an interchangeable strategy, making it easy to add, remove, or modify suggestion types. |
| Chain of Responsibility | Suggestion-generation pipeline | Sequential processing of events and user context through multiple rule handlers; supports modular extensibility. |
| Observer | Vue.js with Pinia reactivity | UI components automatically update when store state changes, enabling seamless synchronisation of user interactions and suggestions. |
| Adapter | JSON serialisation of SQLAlchemy models | Translates internal model structures into frontend-consumable formats using the Axios client. |
| Model–View–Controller (architectural) | Flask backend (controllers), SQLAlchemy models, Vue interface | Improves maintainability by enforcing separation of concerns between data, logic, and presentation. |

At the architectural level, the system adopts a Model-View-Controller separation, enabling a clean division between data management (SQLAlchemy models), interaction logic (Flask controllers), and the frontend interface (Vue.js components). Several behavioural design patterns further structure the assistant's internal logic. Assistant instantiation follows a Factory pattern: the system dynamically constructs or retrieves the appropriate assistant model for each user and permission control stage, maintaining consistency through a database-enforced Singleton constraint that guarantees exactly one assistant instance per user–stage pair. Suggestion generation is implemented via a Strategy pattern, with each rule (e.g., clash detection, sleep-time conflicts, short-transit detection) encapsulated as an independent strategy that can be added or removed without altering the surrounding logic. These strategies are chained together into a sequential suggestion-generation pipeline, illustrating a Chain of Responsibility structure that processes input events and user context through layered transformations.

On the frontend, Vue's reactive Pinia stores act as an Observer system. Specifically, UI components automatically update when the stored state changes (e.g., via user clicks), ensuring synchronisation between interactions, user model updates, and suggestion display. The conversion layer between SQLAlchemy models and JSON responses functions as an Adapter, translating internal data representations into UI-compatible formats. Collectively, these patterns result in an implementation that is both extensible for incorporating new suggestion types or interaction modes (such as chat-based user interaction) and maintainable, supporting future research iterations without requiring substantial restructuring of the codebase.

### 4.2.3 Takeaways from the Implementation

In this subsection, we summarise some of the key takeaways from the development of our scrutable-by-design IPA for time management.

#### High-Level Implementation Approach

Prior research, such as by Osorio et al. [2011], in software engineering has found advantages of iterative development over the more linear and planning-based waterfall paradigm. While SIPA4TM was developed by a single researcher under supervision, rather than a team, the interactive development was still beneficial for two key reasons. First, while extensive and systematic testing of features was performed by the dissertation author, it was only through testing SIPA4TM iterations by the supervisor and volunteer colleagues that unexpected error states and previously undiscovered bugs could be found. Second, iterative consultations and external feedback helped in containing the IKEA effect and the curse of knowledge. The IKEA effect, as introduced by Norton et al. [2012], is a psychological phenomenon in which people overvalue creations they themselves complete. In SIPA4TM development, we let iterative

feedback guide the implementation process, even when it meant abandoning features that had been under development for long periods, such as the monthly calendar view and suggestions. Similarly, the curse of knowledge is a term originally defined in economics by Camerer et al. [1989] to describe the inability of humans to ignore the knowledge they have that others don't when making judgment calls. It is primarily through rotating feedback from colleagues that we recognised some features or UI elements that are non-intuitive to those uninvolved in the implementation. For instance, the idea of stages was not clear to several initial testers, whereas some explanations were found to be inadequate. Akin to what was reported about developers in a company investigated by Osorio et al. [2011] – even though an iterative approach was planned from the beginning – SIPA4TM development reverted to a waterfall-like approach at several points. This may have happened because of overfocus on building a system that matches the envisioned design in full, rather than building runnable iterations and steering further development through rounds of feedback.

Another finding we discovered during the implementation process is that suggestions given in natural language amplify the visibility of logical errors. Moore and Nicolao [2017] noted that the communication capabilities of voice interfaces affect the user's assumptions of system capabilities. While SIPA4TM offers just text-based suggestions, feedback from the colleagues during the implementation phase uncovered that a similar judgment process also occurred for them, where they inferred that unintuitive formatting or unnatural construction of suggestions means that the underlying assistance is also very limited. So while we decided not to use LLMs in our implementation, we took this finding as a motivation to smooth out suggestion phrasing before commencing our full user study.

Similarly, the question of appropriate user onboarding was repeatedly raised during the colleagues' testing of the pre-final SIPA4TM iterations. For instance, during some runs, users would perform all of the tasks unaware that the user model field values are changeable. This is a well-known issue and was also recognised by Kay and Kummerfeld [2013] as one of the domains where scrutable models can help by recognising and then advising the user about the presence of system features they were likely unaware of. The problem seems to have been exacerbated by the fact that opaque models have become more prevalent in recent years, as reported by Löfström et al. [2022], so users do not necessarily expect to scrutinise the system directly, but rather to rely on strategic action-based probing. For us, the solution determined to be most appropriate was to provide each user with a two-minute tutorial of the SIPA4TM's functionalities, the permission control stages, and the scrutiny approach, to showcase that they are there and that the user may explore them more, if they want to at any point.

Lastly, during the implementation, we recognised a trade-off between a more generic but less tuned initial state of the user model and the initial set of calendar events. Initially, our

idea was to have the users in our study fill in a questionnaire on their routines and time management habits, based on which the initial set of calendar events and the user model fields would be generated. However, as the implementation and the initial iterative rounds of testing took place, we realised that such an approach raises both technical issues and methodological challenges. Regarding the former, it proved difficult to ensure that SIPA4TM would be able to provide meaningful and non-trivial suggestions that would realistically warrant scrutiny by the user. Systemic testing of the overall behaviour beyond individual suggestion generation also became much more complex. On the methodology side, the more generic approach raised questions of why the user would want to investigate or adapt the user model if it is already fitted to their preferences from the get-go. For this reason, we decided to implement only one set of initial calendar events and user model values and present them to the user during onboarding, as we elaborate in Section 4.3 below.

### Technical Challenges

Following our reanalyses of studies by Cowan et al. [2017] and Völkel et al. [2021], we chose to bound our domain of scrutable proactive assistance to time management, as our findings indicated it represents a desirable and sufficiently complex use case for IPAs. Following our expert discussion, the domain was further narrowed to individual calendaring assistance. However, in the course of implementation and initial testing, it was revealed that even calendaring assistance remains a complex challenge. In other words, calendar data is deceptively complex — recurrence exceptions, time-zone and summer time changes, and transparent events create edge cases that generic interval libraries do not handle correctly. Relatedly, the (unspoken) assumptions of the original testers who, much like our users in the main study, stem from different generational and cultural cohorts, differed significantly in the expectations of what the expected resolution of such ambiguous behaviours would be. Some examples include expected input formatting (12-hour or 24-hour clocks) and the representation of commutes between events (as modelled implicit constants, as included in the main event duration, or as separate calendar events).

Similarly, providing context-based suggestions, such as alerting the user that their new event duration differs from the previous one, provided both an experience design challenge and an implementation endeavour. Regarding the former, the timing and the implementation of flow-interrupting reminders (as pop-up windows that pause event modification and prompt the user to explicitly confirm or modify their action) were chosen as the best practical solution, although some initial testers disliked them, echoing the findings Hertzum and Hornbæk [2023] who found control flow interruptions to be frustrating "goal thwarting". On the other hand, context-based suggestions provided a data consistency challenge, as they necessitate a separate branch in the suggestion refreshing and creation pipeline. Unlike regular suggestions, they necessitated

additional consistency checking and prioritisation, as they were not only dependent on the calendared events but on the currently-unsaved "draft events" as well.

Another technical challenge was that calendaring suggestions could not be considered in a fully atomic manner, as the existence of one type of affected others. The most straightforward example is that a commute buffer suggestion is not meaningful if the events overlap. Similarly, suggestions that the event duration is too short for the event class should also take into account the user's bedtime. Lastly, based on insights from our iterative design approach, we changed the calendar scope from monthly planning to weekly planning, which presented further challenges, as existing calendaring tools and libraries assume the user would switch between different views and time frames, which was deemed beyond the scope for our study and thus necessitated monkey patching. Similar issues arose because we wanted to restrict our calendar view to display only the events between five in the morning and midnight, to prevent most of the main calendar view from being dominated by empty space and necessitating excessive scrolling. However, later rounds of user testing then revealed unexpected behaviour when new events are placed outside of this displayed window or when the user model event class durations were set to very large values.

### User Interface and Interaction Design Insights

The content of the suggestions themselves was a non-trivial design decision, as initial rounds of testing showed that both too much and too little text led to users disregarding both the suggestion and attempts to scrutinise it. The final design choice was thus to provide just the suggestion title and the proposed remedy and to instead only show the description of the suggestion with more details for when the user clicks the "Explain the suggestion" button. Similarly, to make the suggestions less text-dense, we made calendar events affected by the suggestion a different colour, thus making the assistant's suggestions more prominent in the overall UI (cf. Figure 4.2). However, this then made colouring different event classes (job, hobby, and others) impossible. An initial remedy was to explicitly specify the event class, but this has pushed the superfluous text from the assistant component into the calendar component. The final fix for this was to use emojis to provide intuitive meaning, inspired by a similar approach in scrutable notification management implemented by Fraser [2021].

As we have indicated previously, during our implementation, we also discovered that interaction design elements that the users do not necessarily expect in more popular tools in the domain (for SIPA4TM, this was primarily Google Calendar and Microsoft Outlook), have to be additionally emphasised for user to consider them in their use. In addition to explicitly introducing them in the onboarding tutorial, we also consciously made the "Explain the suggestion" button prominently displayed in the assistant component. Similarly, we made the permission

control stages visible beneath every suggestion (which had been made possible after making the suggestion text shorter), rather than only once the scrutiny was initiated, as had initially been envisioned. Furthermore, we made the current and permissible levels of stages more visually intuitive through the stars-based visualisation, as discussed in Subsection 4.2.1.

A leitmotiv in SIPA4TM implementation has been to make a program whose user interface is not necessarily as polished as commercial e-calendars, but still does not hinder users in their task-solving and scrutiny. On a practical level, this meant that the interface had to be reactive to users' actions without needing a page refresh. We achieved this by using the Vue.js framework on the frontend with Pinia store, as elaborated in Subsection 4.4. While this additional work on frontend implementation has resulted in a significantly prolonged development time, we deemed it as necessary, as we anticipated refresh delays would cause too much frustration, as indicated by – among others – Hertzum and Hornbæk [2023], who describe how lag, slowness, and blocking interactions are classic sources of frustration. In our frontend design, we kept in mind the guidelines and principles outlined in the book on designing user interfaces by Shneiderman et al. [2018]. For instance, we aimed to have standardised sequences in different contexts (e.g. in choice and placement of options in the context component when scrutinising a suggestion and when editing an event), to use the most intuitive input elements for the data type (for the user model as well as event modifications), and to use thumbnail images to preview larger images (in level two explanations).

### Abandoned Features

Our initial scope for SIPA4TM included a more advanced implementation of proactive features, where the highest stage would also enable the assistant to change the events and just have the user accept or reject them, rather than just suggesting how to do so. Unfortunately, due to a limited time for implementation and issues with preliminary versions of these features, we had to ultimately abandon this endeavour and to instead investigate it through questions in the semi-structured focus group discussions.

Similarly, as previously discussed in this subsection, we agree with Kay and Kummerfeld [2013] that a scrutable personalised system ought to track features that the user has not used, even though they would likely be helpful, and to inform the user of the existence of these system features. However, the implementation of such a system was also seen as rather technically challenging and its use applicability impeded by the short time the user has to interact with SIPA4TM. The onboarding tutorial given to the user study participants was thus devised as a substitute for this interaction element as well.

## 4.3 User Study Protocol

To evaluate SIPA4TM, described in the previous section, and to investigate user attitudes towards scrutability, proactivity, and delegation in IPAs, we devised a three-phase user study protocol. The study was designed to be conducted in person at Trinity College Dublin, lasting approximately 90 minutes per session. The protocol was structured to query participants on their current experience with time management tools and AI systems, to facilitate task-based interaction with SIPA4TM, and finally to engage them in a reflective discussion on the broader implications of time management assistance. The full user study interaction protocol, alongside user study documentation and resources, is provided in Appendix A4.

### 4.3.1 Phases of the User Study

The first two phases of the user study were conducted individually, with a collaborating researcher guiding each participant. The third phase was the focus group discussion, attended by two or three participants and moderated by the principal researcher (i.e., the dissertation author).

#### Phase 1: Survey

All invited participants were informed of the study time and place via email, including a digital copy of the informed consent form. On the day, the participants were welcomed with a standardised greeting and asked to join a room with a collaborating researcher. There, the collaborating researcher gave the participant a printout of the informed consent form to sign. The form emphasised the voluntary nature of the study and data protection measures (including GDPR compliance). The participant also had a chance to ask questions before signing the form and proceeding with the study. Each participant then selected a pseudonym for the study. The pseudonym was checked against an existing list to ensure its uniqueness. Participants were informed that pseudonyms would be used for anonymisation and self-introduction during the group discussion. The pseudonym and date were recorded on the consent form to facilitate the deletion of data upon request.

Participants were then asked to complete the survey. This survey included both typical demographic information and questions about users' experience with IPAs, time management tools, and large language models. The latter part of the survey aimed to enable contextualisation of the remaining study data. The questionnaire was inspired by reported questions in similar qualitative IPA studies, including Luger and Sellen [2016]. The survey questions and the results are presented in Section 5.2.

### Phase 2: SIPA4TM Interaction

At the beginning of the second phase of the user study, each participant was informed by a collaborating researcher that they would be asked to make calendar changes for the upcoming week for an imaginary user. The collaborating researcher further explained that they would be doing so using a desktop program akin to Google Calendar that includes a text-based time management assistant. The researcher also reminded the participant that the remaining phases of the study would be audio-recorded before proceeding to start recording.

The participant was then given a printout of the high-level agenda for the upcoming week with noted priorities and habits of the user they are to role-play as. The participants were informed that they did not have to memorise the schedule, but just to gain a general insight thereof. This printed document is reproduced in Appendix A4.2.1. The researcher then asked the participant a few open-ended questions about their typical schedule and how it compared to the one they had just read. This was done to ensure the participant had understood the assignment and was content to proceed with the interactive part of the phase.

The participants were then introduced to the SIPA4TM on a researcher-owned computer. The introduction began with a brief onboarding tutorial to familiarise the participant with the three-pane interface described in Section 4.2. The rationale for this onboarding approach is detailed in Subsection 4.2.3. Next, the user was told they could explore the assistant and begin with the tasks when ready. The interface was already loaded with a standard template calendar, as shown in Figure 4.2. All participants were provided with the same initial calendar, assistant, and user model settings. The participants were then given a printout of the study tasks. All tasks, as printed, are reproduced in Appendix A4.2.2 and we elaborate on the task choice in Subsection 4.3.2.

To capture the users' immediate cognitive processing and decision-making strategies, we employed verbal protocol analysis, i.e. the think-aloud protocol [Ericsson and Simon, 1993, Gilhooly and Green, 1996, Fox et al., 2011]. The participant was instructed to verbalise their thoughts, uncertainties, and planning processes as they interacted with the calendaring system. In addition, interaction logs were automatically generated to capture the sequence of user actions. The collaborating researcher remained in the room to provide cursory help with potential technical difficulties without interfering with the experiments. Participants were not timed during their interaction with SIPA4TM and were told to complete the tasks at their own pace. Most participants took between 30 and 40 minutes to complete tasks in the second phase. The collaborating researcher emphasised that the user can proceed to the next task whenever they consider the previous task resolved.

Once the participants confirmed that they had completed all the tasks and interacted with

SIPA4TM to their satisfaction, the screen and audio recordings were saved for later analysis, named by the participant's pseudonym.

### Phase 3: Semi-Structured Focus Group Discussions

The last phase involved group discussions about the participants' experiences during their interaction with SIPA4TM. Groups of two to three participants were queried by the lead researcher in semi-structured group discussions. The discussion themes included participants' general approach to personal assistants and digital scheduling tools, their experience with the system, and feedback on components such as the assistant, user model, and calendar interface. The moderator's questions were divided into four main groups:

- *SIPA4TM-Related Questions:* Assessing the perceived utility of SIPA4TM, its suggestions, interaction design, and the user model; inviting the participant to suggest improvements of the assistant.
- *Scrutability:* Investigating the user's desire to inspect the user model and the logic behind personalisation of the provided suggestions, asking about their willingness to invest time and effort in understanding and improving personalised assistance.
- *Proactivity, Delegation, and Stages:* Enquiring about trust thresholds for proactive and delegative assistance; exploring whether such more substantive assistance causes new challenges for scrutability to emerge and whether the permission control in stages is a desired data stewardship paradigm.
- *Usage Willingness:* determining the factors that would influence the adoption decision for proactive IPAs for time management.

The study protocol with the final question recommendations is reproduced in Appendix A4.3.3. The formulation of questions was inspired by the transcripts from Cowan et al. [2017] (which we had received from a study author, alongside advice for our study), as well as by the findings from other qualitative human-assistant interaction studies, including Luger and Sellen [2016], Bartle et al. [2022], Zargham et al. [2022], and our practical experience from the expert discussion moderation. Specifically, based on questions and the findings from Cowan et al. [2017], we wanted to ask the participants what their impressions of SIPA4TM's calendaring assistants are (in general and in comparison to the commercial IPAs), what features they found particularly useful (if implemented in SIPA4TM) or missing (if not), whether they were able to estimate the assistant's capabilities easily, whether they would trust an IPA to set sensitive appointments for them (what the privacy considerations are and if the scrutability and/or control or data use in stages affects it), the future of IPAs and features they would like to see a time management assistant to have as well as whether they personally would

use an IPA for time management (what the primary factors influenced their decision are and whether scrutability affects that decision). Inspired by the findings of Luger and Sellen [2016], we asked the participants whether they were satisfied with the SIPA4TM's action feedback mechanisms, to what extent they were satisfied with the assistant revealing its capabilities, and if they believed IPAs would be more useful than they are now (whether they could become less task-centric and more akin to human assistants). The questions about proactivity are informed by our expert discussion as well as by the reported findings of Völkel et al. [2021] and Zargham et al. [2022]. Specifically, these studies lead us to ask questions on the opportune moments for proactivity, on the appropriate level of control they would want to have over the assistant's proactive behaviour, and on the type of suggestions and recommendations they would welcome.

The theoretical basis for the protocol of our focus group was the publications on semi-structured interviews and focus groups by Longhurst [2003] and Harrell and Bradley [2009]. As planned, our focus group lasted 30 to 50 minutes, depending in part on group size. This duration is in line with both the recommendation by Longhurst [2003] (who suggests focus groups of six to twelve participants to last 60–120 minutes) as well as the experience from the transcripts from Cowan et al. [2017] (where groups of three to five participants lasted from 20 minutes to just under 50 minutes). In line with Longhurst [2003], we see focus groups as appropriate for researching novel topics, such as the confluence of the research areas of scrutability and proactive IPAs. Focus groups also allow for discussion among participants, with the moderator playing a less significant role than in individual semi-structured interviews. Following the guidelines of Harrell and Bradley [2009] and keeping the exploratory nature of the user study in mind, not all protocol questions were asked in all focus groups, and their order and phrasing varied. Such choices were made to facilitate more natural conversation, enabling participants to express their opinions more freely while staying within the study's time constraints.

### 4.3.2 Chosen Tasks

In formulating the tasks, we aimed to simulate desired use cases suggested by the experts in our discussion, as well as the user studies we had reanalysed. We also wanted the tasks to grow in complexity, so we could observe whether users would engage in scrutinising SIPA4TM's suggestions and model, and, if so, at what level of task complexity and to what depth of exploration. To explore such "organic" scrutiny and the actualisation of the effort-efficiency principle hypothesised in Chapter 3, we set the first four of the seven tasks to be achievable without recourse to the assistant's advice at all. The user study protocol also stipulated that users be asked not to look at the following tasks until they had finished the ones before

them. Tasks five to seven were printed on the back side of the task sheet, so as to be hidden from view. All tasks are reproduced in Table 4.5 and their underlying research intents can be categorised into four groups, discussed below.

Table 4.5: Tasks for the second part of the user study (interaction with SIPA4TM)

| # | Task |
|---|---|
| 1 | Add a grocery pickup on Sunday before the meal preparation time. |
| 2 | Move the Monday Weekly Team Meeting 15 minutes early, so that it starts at 9:15. |
| 3 | Ensure there are no clashes over the week. |
| 4 | Can you find any other improvements to the calendar for the upcoming week? |
| 5 | What are the different stages of the assistance?<br>• Can you see what information the assistant requires for each of them?<br>• If you haven't already, increase the stage to the maximum.<br>• Does this give you more ideas for potential improvements to next week's schedule? |
| 6 | If you have not done so already, make sure that you are on time for the doctor's appointment on Wednesday. |
| 7 | Set up an event to pick up a friend from the airport at 12:30 on Saturday.<br>• Can you discover how the assistant sees what matters to the user?<br>• Explore the suggestion explanations until you feel you understand how it works. |

### Basic Event Manipulation

The initial tasks (#1 and #2) were designed to acclimatise the user to the click-and-type interface and the constraints of the simulated schedule. They also served as indicators to see how often the suggestions or the user models would be scrutinised for such relatively straightforward tasks and assistance.

### Aggregate Constraint Verification

Users were explicitly asked to "Ensure there are no clashes over the week" in task #3, prompting them to rely on the calendar's visual feedback or the assistant's suggestions. An underlying hypothesis was that the assistance would be more valuable to participants in more aggregate tasks. Furthermore, we wanted to see whether higher task and suggestion complexity, associated with tasks #3 and #4, would motivate users to invest more effort into understanding the assistant.

### Explicit Prompting to Engage in Scrutiny and Permissions Control

A pivotal moment in the interaction is task #5, whose subtasks required the user to interact with the *permissions control stages* mechanism. Users were asked to observe the different stages of assistance and identify what information was required for each. They were then

instructed to increase the stage to the maximum to observe how the system's proactivity and suggestion granularity shifted in response to increased data access. The last subtasks, then, for the first time, explicitly invite the user to engage with the suggestions. Still, the users do not have to engage with the explanations for the suggestions, unless they choose to. Moreover, it is up to the users to decide whether they want to explore each stage in more detail or whether they would be content with the one-sentence explanation, shown in Figure 3.5. Task #5 thus offers an insight into effort-efficiency in data stewardship. Relatedly, task #6 was added, so that the changes in the suggestions that SIPA4TM can generate for each stage can be noticed, which would not have been guaranteed otherwise due to the open-ended task #4.

#### Explicit Scrutability

The final task guided the user directly to scrutinise SIPA4TM's personalisation. The first subtask of task #7 first asks for an event with a location not previously in the system to be created. This served as an indirect invitation into a non-trivial exploration of the suggestions and the knowledge that the assistant has (previously entered commute distances) and does not have (commute time to and from the airport). Similarly, it was up to the participant to decide which suggestions and which model fields to explore, as well as the depth. As with all other tasks, for task #7 too, the collaborating researcher did not decide what constituted sufficient exploration, and was only there to help with potential technical difficulties and to remind the participant to verbalise their thoughts, should they forget to do so.

### 4.3.3 Initial Events and User Model Values

During the initial study ideation, we planned to have the events at the beginning of SIPA4TM interaction in phase 2, generated to align with participants' survey responses on what they imagine a typical week's schedule looks like. Nevertheless, such an approach was abandoned in favour of a single set of initial events to standardise the tracking of logged events and to make the assistant's behaviour (across the three stages) more predictable and easier to analyse. In this subsection, we briefly explain how this initial set of events was selected.

Following our expert discussion and the evaluation of interaction and modelling challenges, we drafted an initial set of high-level user tasks for SIPA4TM. Following the creation of the initial draft of the tasks and the expected SIPA4TM suggestions, we outlined the preliminary set of calendar events. In essence, we imagined a calendar of a highly organised, quite busy knowledge worker, whose agenda includes meetings, scheduled focus time, doctor visits, as well as domestic obligations and social events. In line with the iterative approach of the Design-Based Research, we refined both the tasks and initial events over multiple rounds of consultations and supervisor meetings.

Similarly, following the initial set of calendar events and in line with the high-level idea of expected SIPA4TM suggestions, we drafted the preliminary default values for the user model. As with the events and SIPA4TM, the user model values were also iteratively refined. The final selection of both the initial calendar events as well as the user model fields and values is guided by three key considerations, discussed below.

- **Professional and private constraints:** The schedule was anchored by office hours (roughly $09:00 - 17:00$) and two major projects. Meetings related to these projects were deemed important and difficult to reschedule. Preferences regarding further event types (such as family obligations or hobbies) were not specified in the printout given to participants, allowing the participants to make assumptions in line with their own preferences and to increase the associated feelings of ownership over task solutions, hopefully leading to a more natural system interaction.
- **Sleep schedule:** The constructed person using SIPA4TM, whom the participants were to role-play as, prioritises healthy sleeping habits, with a bedtime at $23:00$ on workdays and $23:30$ on weekends. Such a sleep schedule was modelled as a soft constraint in the assistant's model, and participants were informed of it in the printout about the person's time management preferences. Still, two initial events were intentionally scheduled at this time to trigger the relevant assistant's suggestions and to give participants a chance to either modify the events or the values in the user model.
- **Commute times:** To show assistant's travel-time suggestions, a temporal topology of locations was defined. This included the constructed person's office of work (45 minutes from their home), the person's gym, store (both near their home), social venues (further away), etc. A commute timetable was created and encoded in the user model, defining both exact travel times between known locations and a default commute buffer time with an unknown location, both of which the users could investigate in the user model. This timetable is shown in Table A4.1 in Appendix A4.2.3.

The exact printout with the constructed person's time management preferences is given in Appendix A4.2.1. The commute times in the assistant's user model were not included in the printout, and the participants could only uncover them through interaction with SIPA4TM. The full commute timetable can be seen in Table A4.1 in Appendix A4.2.3.

## 4.4 Ethical Considerations and Approval

The Trinity College Dublin ethics committee approved the study before it took place. To apply for and obtain the approval, we ensured that our method, study surveys, tasks, and protocol all minimise harm risk for both the collaborating researchers and the participants. Practically,

this meant sending the consent form (reproduced in full in Appendix A4.1) in advance to participants, notifying them that the study includes an in-person discussion element, but that they will be introduced using a pseudonym and that they would be free to leave at any point should they wish to do so. All collaborating researchers had phone numbers of security services and were informed about the user study consent form, user study protocol, and other study details.

We also ensured compliance with best practices for data storage security and data anonymisation, in line with the GDPR. All non-anonymised user study data was stored either on the devices owned by the main researcher or on a university-affiliated cloud drive. The signed consent forms were kept locked in the university premises until the anonymisation was complete and they could be safely destroyed. Audio files created during phases two and three were automatically transcribed and manually verified within seven days of the study's completion. The automatic transcription was done locally, using WhisperX [Bain et al., 2023]. During the manual check, the transcripts were anonymised by stripping all potentially identifying data (e.g. "my sister Sarah, who is 24, and I might use the assistant to remind us of shared tasks", would have been changed to "[my sibling, who is in their twenties], and I might use..."). After this was done, and seven days had elapsed since the study began, the audio recordings were deleted. The participants were informed that they could request deletion of all data associated with them within seven days after the study. To facilitate this, a map linking the pseudonyms to participants' identities was kept for seven days after their participation, after which it was destroyed. We received neither data deletion requests nor complaints on the conduct of the study from the study participants after the study was completed.

## 4.5 Chapter Summary

We implemented a Scrutable Intelligent Assistant for Time Management (SIPA4TM) as a desktop application to serve as the anchoring point in our main user study, evaluated in Chapter 5.

SIPA4TM was implemented in line with the principles of scrutable user models by Kay and Kummerfeld [2013] and the design guidelines by Kay and Kummerfeld [2019]. Each suggestion has three levels of explanation. The user could control the model ontology, evidence, and reasoning.

SIPA4TM was also designed to showcase the idea of permission control in stages – derived from the expert discussion in Chapter 3 – which enables the user to add and remove access to the system.

Seven interaction tasks were created to guide the user's logged interactions with SIPA4TM;

the first four do not explicitly ask the user to engage with its scrutability features, thus serving as a proxy for actual user interest.

In line with the Design-Based Research approach, SIPA4TM was developed iteratively, incorporating feedback from the volunteering testers, while the user study protocol for post-interaction focus-group discussions was created, so as to prompt views-sharing about aspects of IPAs and LLM-based chatbots that SIPA4TM does implement (such as a conversational interface, proactive calendar changes, and delegation of communication with others).

# 5 User Study and Evaluation

In this chapter, we provide more details on a SIPA4TM-anchored user study with 22 participants. We begin by describing the chosen user study conduct and evaluation methods, recapitulating task design, detailing our participant selection process, and elaborating on data preprocessing. We continue by reporting on the results from our verbal protocol analysis, the thematic analysis of the post-interaction focus groups, and the quantitative results of the study. We also present a central theme of our qualitative evaluation, the idea of *prospects*, that connects the user's interaction with the system and the user's trust in the system to act proactively and delegate. We close by evaluating the results from the lens of proactive time management IPAs.

## 5.1 Task-Based User Study

Our task-based user study investigated user preferences, trust, and willingness to use regarding the personalisation and proactivity of Intelligent Personal Assistants (IPAs). For this study, we define personalisation as the assistant adapting its services, responses, and suggestions to the user's preferences. Proactivity means the IPA offers suggestions or other services without being explicitly prompted beforehand. To date, most commercial IPAs, e.g. Google Assistant, Amazon Alexa, and Apple Siri, remain purely reactive (only responding to explicit requests by the user), simple (mainly used for tasks such as setting reminders or alarms), and non-personalised (meaning that they do not learn habits or preferences of the user to perform their services better). In this study, we aimed to understand design principles that would enable future, more advanced IPAs to be more helpful, more enticing, and better aligned with potential users' priorities and values.

In particular, we are focusing on how the user's data ought to be modelled and how the scrutiny of the personalisation is to be facilitated. Inspired by the study by Cowan et al. [2017], this research has three phases: 1) a survey; 2) an interaction with our assistant; and 3) a group discussion on related impressions and ideas. We focus on time management assistance (calendaring, event management, reminders, etc.) as a specific use case for the envisioned

IPAs. Further details about each of the three phases, as well as the implemented calendar assistant, are presented in Chapter 4. In short, the study began with participants completing a questionnaire on their demographic information, prior experiences with AI-based systems, and their interests in explainability, privacy, and AI. In the second phase, the participants interacted with SIPA4TM, our purpose-built Scrutable Intelligent Personal Assistant for Time Management. The assistant was locally hosted in a laptop provided to each participant. Participants were asked to think aloud and were audio-recorded while solving the provided time-management tasks. The final phase was a semi-structured group discussion. The group discussions consisted of two to three participants, with each group comprising individuals who had concurrently completed the individual interaction phases immediately prior. This session was facilitated by a moderator using a structured discussion guide designed to elicit feedback on the participants' experiences and perceptions of interacting with the system. The last stage was also audio-recorded, and the resulting transcript served as the basis for the thematic analysis.

To evaluate our results, we relied on survey data, interaction logs, thinking-aloud audio recordings, and group discussion audio recordings. The main inspiration for our user study method and the number of participants was the work of Cowan et al. [2017], which investigated perspectives on Apple Siri among infrequent IPA users. Our user study was further informed by Luger and Sellen [2016], who presented an earlier study on IPAs and employed audio-recorded individual interviews, which were then thematically analysed. In formulating the protocol and conducting the focus groups, we also kept in mind the exploratory qualitative work by Bartle et al. [2022], who used several hypothetical usage scenarios to elicit desired use cases that go beyond currently implemented IPA capabilities. Lastly, Werner et al. [2023] also relied on focus groups and thematic analysis to investigate older users' perceptions of speech interfaces.

### 5.1.1 Research Collaborators

The study method and SIPA4TM implementation were done entirely by Jovan Jeromela (the dissertation author), under the supervision of Prof. Owen Conlan. The user study had additional collaborators: Matt Murtagh White, Jasmina Gajcin, Alok Debnath, and Dipto Barman. All four were PhD candidates or PhD graduates from Trinity College Dublin at the time of the study, and their collaboration has been approved by the institutional ethics committee. The work done by Murtagh White, Gajcin, Debnath, and Barman towards this user study was not included in their respective PhD dissertations. The contributions of each of the researchers are provided in Figure 5.4[1].

[1]Fully anonymised user study data, including a backup of SIPA4TM code, can be found at `https://drive.google.com/drive/folders/14fvkx9oScqboDqrDj2Tmtaw0zrHwVNe8?usp=sharing`, whereas additional materials relating to the thematic analysis of focus group discussions, such NVivo files corresponding to iter-

### 5.1.2 Participant Selection

This study was conducted with 22 participants divided into nine groups of two to three members for group discussion. The number of participants was consistent with prior qualitative user studies, such as by Herder and Staring [2024], Luger and Sellen [2016], and Cowan et al. [2017]. We recruited prospective participants by pinning study participation call posters on university campus notice boards; by posting information about the study on the university social media; and by distributing leaflets with an interest registration link in postboxes of houses up to two blocks away from the campus.

To register their interest in participating in the user study, prospective participants needed to indicate their availability, age, gender, and university status. From there, we selected the participants on a first-come, first-served basis whenever we had enough interested registrations and available collaborating researchers for the next three-person slot. In cases where multiple groups were available, we chose to first contact participants who would improve gender and age balance, and increase the number of non-university-affiliated participants. We did so, aiming to have a more diverse and thus more generalisable insight. Once a group was formed, the participants were informed by email of the suggested time and asked to confirm their attendance. Still, several groups had participants not show up despite prior confirmation, which contributed to a suboptimal balance across the three demographic categories.

## 5.2 Participant Demographics and Surveyed Attitudes

All participants were surveyed on their experience with human personal assistants and AI-based Intelligent Personal Assistants, their opinions on explanations for autonomous actions and data privacy, and their general use and interest in AI. Table 5.1 provides an aggregate overview of the responses to the demographic questions.

Overall, the participant group represents a population that is both mostly accustomed to structured time-management practices and rather familiar with contemporary AI tools, while remaining heterogeneous in attitudes toward privacy, explainability, and trust. Our participant sample was skewed towards younger adults (77.3% under 35), with a strong representation of university-affiliated participants (63.6% staff or research students), and a comparatively high level of formal education (54.5% holding a Master's degree or higher). This composition likely reflects the self-selecting nature of applying for the participation and the advertisement of the study on boards of the university campus.

While the relatively high levels of formal education and interest in AI likely diverge from

ation of the process can be found at `https://drive.google.com/drive/folders/1vxB0sGqi7-NbP3TX_MCfNhXZrMfMvHbW?usp=sharing` (last access 2026-03-12).

Table 5.1: Aggregate demographic profile of the study participants ($N = 22$)

| Variable | Category | Count | Percentage |
|---|---|---|---|
| **Gender** | Male | 13 | 59.1% |
| | Female | 9 | 40.9% |
| **Age** | 18–24 | 10 | 45.5% |
| | 25–34 | 7 | 31.8% |
| | 35–44 | 1 | 4.5% |
| | 45–54 | 3 | 13.6% |
| | 55–64 | 1 | 4.5% |
| **Highest degree** | Doctorate | 3 | 13.6% |
| | Master's | 9 | 40.9% |
| | Bachelor's | 8 | 36.4% |
| | Secondary | 2 | 9.1% |
| **University status** | University staff | 9 | 40.9% |
| | Research student | 5 | 22.7% |
| | Taught student | 4 | 18.2% |
| | Non-university | 4 | 18.2% |
| **English knowledge** | Mother tongue | 15 | 68.2% |
| | Taught speaker | 7 | 31.8% |
| **Fluent languages** | One (English only) | 9 | 40.9% |
| | Two | 8 | 36.4% |
| | Three or more | 5 | 22.7% |

the general population, it also indicates a sample that is both prone to analytical thinking and practically experienced in scheduling, making them particularly relevant for investigating scrutability and trust in time-management IPAs. At the same time, the group was diverse in terms of linguistic or cultural background. Nearly one-third of participants (31.8%) reported English as a taught rather than a native language, and over half reported fluency in two or more languages. This linguistic diversity is noteworthy in light of prior findings, including by Cowan et al. [2017], that language competence and accent can significantly shape experiences with IPAs and speech interfaces more broadly. While voice-based interaction was not facilitated by SIPA4TM, participants' reflections on questions about voice assistants are thus more likely to reflect varied user experiences. Table 5.2 provides an overview of participants' individual demographic information.

As shown in Figure 5.1, daily use of digital calendars was nearly universal across the sample, confirming that calendaring is an established, routinised practice for most participants. In contrast, prior use of commercial IPAs (e.g., Siri, Alexa, Google Assistant) was markedly lower and more fragmented, with many participants reporting only past or irregular use. This pattern aligns remarkably with the population studied by Cowan et al. [2017], who similarly focused on users for whom IPAs are available but not integrated into everyday workflows. As in that

Table 5.2: Demographic characteristics of individual participants

| Participant | Gen. | Age | Highest degree | Status | English | Lang. |
|---|---|---|---|---|---|---|
| [G1P1] | M | 18–24 | Bachelor's | Research student | Native | 1 |
| [G1P2] | M | 25–34 | Master's | Research student | Native | 1 |
| [G2P1] | M | 18–24 | Secondary | Taught student | Taught | 2 |
| [G2P2] | M | 18–24 | Bachelor's | Research student | Native | 2 |
| [G2P3] | M | 45–54 | Bachelor's | Non-university | Native | 1 |
| [G3P1] | M | 18–24 | Secondary | Taught student | Native | 1 |
| [G3P2] | M | 18–24 | Master's | Staff | Native | 5 |
| [G4P1] | M | 25–34 | Master's | Staff | Native | 1 |
| [G4P2] | F | 18–24 | Bachelor's | Taught student | Native | 2 |
| [G5P1] | F | 25–34 | Doctorate | Staff | Taught | 2 |
| [G5P2] | M | 45–54 | Doctorate | Staff | Native | 1 |
| [G6P1] | F | 18–24 | Bachelor's | Non-university | Native | 3 |
| [G6P2] | F | 45–54 | Master's | Staff | Taught | 2 |
| [G6P3] | M | 18–24 | Bachelor's | Non-university | Native | 2 |
| [G7P1] | F | 35–44 | Master's | Non-university | Taught | 3 |
| [G7P2] | M | 18–24 | Bachelor's | Staff | Taught | 2 |
| [G7P3] | F | 25–34 | Master's | Research student | Taught | 3 |
| [G8P1] | F | 25–34 | Master's | Staff | Native | 1 |
| [G8P2] | M | 18–24 | Master's | Taught student | Taught | 2 |
| [G8P3] | M | 25–34 | Master's | Research student | Native | 3 |
| [G9P1] | F | 25–34 | Doctorate | Staff | Native | 1 |
| [G9P2] | F | 55–64 | Bachelor's | Staff | Native | 1 |

work, participants in our study were generally aware of IPAs and their advertised capabilities, yet described limited sustained engagement – often attributing disengagement to reliability issues, misaligned expectations, or insufficient perceived value.

We further queried participants on their prior experience with relevant technologies to establish a baseline for their interaction with SIPA4TM. Figure 5.1 shows aggregated results of prior experience regarding time management tools and services, while detailed individual usage patterns are presented in Table 5.3.

As shown in Table 5.3, the vast majority of participants are daily users of electronic calendars (e.g., [G1P1], [G2P2], [G5P1]), indicating they are regular users in the domain of time management. Furthermore, a significant portion of the cohort interacts with LLMs on a daily or weekly basis. However, as previously stated, this engagement does not translate to traditional IPAs: the majority of participants reported either never using IPAs or only using them in the past.

A noteworthy finding from this data is the distinct divergence between the participants' use of electronic calendars (eCal) and Large Language Models (LLMs) versus their use of traditional

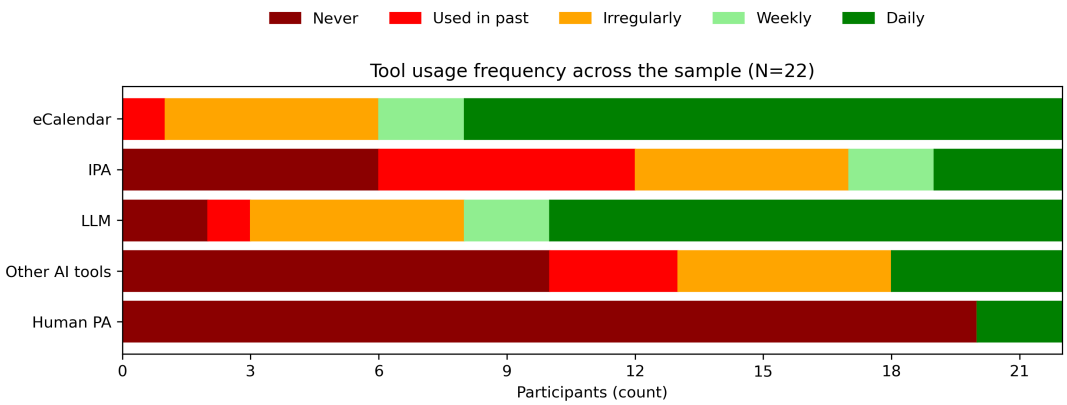


Figure 5.1: Aggregated results of prior experience with e-calendars and AI tools

IPAs (e.g. Siri or Alexa). This pattern echoes the findings of Luger and Sellen [2016], who – in reference to Norman [1988] – identified a "gulfs of execution and evaluation", where user expectations of conversational agents often outpaced system capabilities, leading to abandonment. The high prevalence of study participants who used IPAs in the past but do not use them now suggests that many participants may have tried and rejected earlier generations of IPAs due to functional limitations. However, unlike the infrequent users described by Cowan et al. [2017], who often hesitated due to social embarrassment or privacy opacity, our participants are active users of generative AI (LLMs). This suggests that while they may have rejected the "command-and-control" model of legacy IPAs, they are comfortable with the conversational and reasoning capabilities of modern LLMs. This positions our study at a unique intersection: our participants are sceptical of the old IPA utility but experienced with the current underlying technology (LLMs).

Figure 5.2 shows aggregated attitudes on the interest in AI, the importance of explanations, and the importance of privacy. Attitudinally, the participant pool exhibited substantial variation: interest in AI was generally high, but concern for privacy and the importance attributed to explanations varied widely, both across and within groups. Several participants who expressed high interest in AI nonetheless reported low trust in autonomous behaviour, while others with moderate AI interest emphasised strong requirements for explainability and control. Similarly, although many participants acknowledged privacy risks as an abstract concern, relatively few had actively exercised data retrieval rights in relation to the GDPR [2016].

At the individual participant level, the attitude data on the right-hand side of Table 5.3 reveals a heterogeneous landscape regarding privacy and explainability. While interest in AI is generally "High" or Medium" across the board, the importance placed on privacy varies significantly, ranging from High" (e.g., [G6P3], [G7P3]) to None" ([G2P1]). This variation is crucial to our thematic analysis, as it allows us to investigate whether scrutability tools are universally

Table 5.3: Individual results for AI tool usage (eCal – eCalendar, Oth. – Other AI tools), having a human personal assistant (H.PA), attitudes (AI – interest in AI, Priv. - Importance of Privacy, Expl. - Importance of Explanations), and whether they have ever requested their data (R.D.) for each participant (Part.)

| Part. | eCal | IPA | LLM | Oth. | H.PA | A.I. | Priv. | Expl. | R.D. |
|---|---|---|---|---|---|---|---|---|---|
| [G1P1] | Daily | Never | Daily | Never | Never | High | Low | Med | Once |
| [G1P2] | Daily | Past | Daily | Never | Never | High | Med | Med | Once |
| [G2P1] | Daily | Weekly | Daily | Never | Never | High | None | Med | Once |
| [G2P2] | Daily | Never | Irreg. | Never | Never | Low | Med | Med | Once |
| [G2P3] | Daily | Never | Never | Daily | Never | Med | Med | Med | Never |
| [G3P1] | Past | Past | Irreg. | Never | Never | Med | High | Med | Mult. |
| [G3P2] | Irreg. | Irreg. | Daily | Daily | Never | Med | Med | Med | Never |
| [G4P1] | Daily | Weekly | Daily | Irreg. | Never | Med | High | Low | Mult. |
| [G4P2] | Daily | Past | Weekly | Past | Never | High | Med | Med | Never |
| [G5P1] | Daily | Past | Daily | Irreg. | Never | High | Low | Med | Unsure |
| [G5P2] | Irreg. | Daily | Weekly | Daily | Never | High | High | Med | Mult. |
| [G6P1] | Irreg. | Never | Never | Never | Never | Low | Med | Low | Never |
| [G6P2] | Daily | Irreg. | Daily | Never | Currently | Med | Med | High | Never |
| [G6P3] | Weekly | Never | Irreg. | Past | Never | Low | High | Low | Never |
| [G7P1] | Weekly | Irreg. | Daily | Irreg. | Never | Med | Low | Low | Once |
| [G7P2] | Daily | Daily | Daily | Irreg. | Never | High | Med | High | Once |
| [G7P3] | Daily | Irreg. | Daily | Past | Never | High | High | Med | Never |
| [G8P1] | Daily | Past | Irreg. | Irreg. | Currently | Med | Med | Med | Once |
| [G8P2] | Irreg. | Irreg. | Daily | Daily | Never | Med | High | Med | Once |
| [G8P3] | Daily | Past | Daily | Never | Never | Med | Med | Med | Unsure |
| [G9P1] | Daily | Daily | Irreg. | Never | Never | Med | Low | None | Never |
| [G9P2] | Irreg. | Never | Past | Never | Never | Med | High | High | Never |

valued or appeal primarily to privacy-conscious users. Notably, several participants who rated privacy importance as "Low" still reported requesting their data from service providers at least once (e.g., [G1P1], [G7P1]), indicating that a lack of theoretical concern does not necessarily equate to a lack of data oversight behaviour. This aligns with Cowan et al. [2017], who noted that concerns about data permanence and transparency are significant for infrequent users. By including participants with diverse attitudes toward privacy and explanation, we ensure that the evaluation of SIPA4TM captures a broad spectrum of user needs, from those seeking efficiency to those demanding granular control.

## 5.3 Verbal Protocol Analysis

Verbal protocol analysis is a set of qualitative research methods for understanding participants' thought processes while performing a task [Ericsson and Simon, 1993]. Verbal protocols can be collected either while the participants are performing the task (concurrently) or retrospectively [Ericsson and Simon, 1993]. In our study, we opted for a type of concurrent verbal protocol analysis called "think-aloud", which captures participants' immediate thoughts as they perform

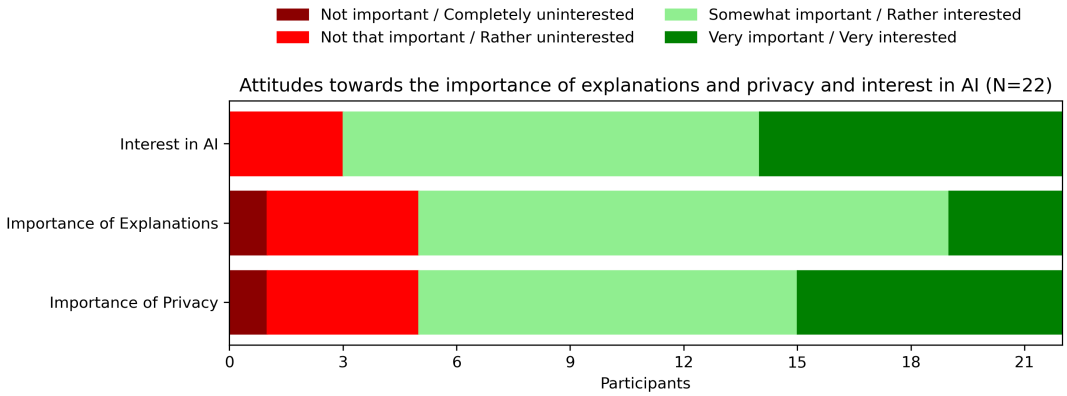


Figure 5.2: Aggregated attitudes on the interest in AI, the importance of explanations, and importance of privacy

the tasks at hand [Gilhooly and Green, 1996].

In choosing this approach, we consulted literature on verbal reporting procedures by Fox et al. [2011], who, in reference to Ericsson and Simon [1993], proposed that there are three levels of thought vocalisation, of which the "think-aloud" approach captures levels 1 and 2. Specifically, through "thinking-aloud", participants verbalise their inner speech ("thoughts already activated in attention as verbal articulations" [Fox et al., 2011]) as well as thoughts outside of the inner speech that participants find necessary to vocalise as reference points. Crucially, the participants in "think-aloud" studies do not interpret their activities or thinking processes [Ericsson and Simon, 1993, Fox et al., 2011]. While thinking aloud might feel unnatural to participants [Gilhooly and Green, 1996], Fox et al. [2011]'s meta-analysis has shown that this protocol does not significantly affect participants' task performance.

### 5.3.1 Interaction Tasks Summary

As detailed in Section 4.3, the study employed a set of seven interaction tasks, listed in Table 4.5 on page 79. Participants were asked not to preview subsequent tasks, and Tasks 5–7 were printed on the reverse side of the task sheet to prevent anticipatory behaviour. The facilitating researcher did not intervene in participants' decision-making, except to assist with technical difficulties, task-understanding issues, or to prompt continued verbalisation during the think-aloud protocol.

Tasks #1 and #2 focused on basic event manipulation, i.e. creating and editing calendar entries, primarily serving to familiarise participants with the SIPA4TM interface, its control style, and the constraints of the simulated schedule. These tasks also functioned as a baseline for observing whether participants would engage with SIPA4TM's suggestions or explanations

in situations where assistance was clearly optional and cognitive demands were low. Tasks #3 and #4 introduced aggregate constraint verification across multiple days, explicitly requiring participants to ensure that no clashes occurred within the schedule. While still solvable through direct visual inspection of the calendar, these tasks increased both temporal scope and reasoning complexity, thereby making consultation of the assistant's suggestions more utilitarian and the underlying reasoning less immediately interpretable.

In contrast, task #5 explicitly prompted participants to engage with the SIPA4TM permissions control and to increase the control stage to the maximum. Still, while participants were asked to compare stages, they retained full discretion over whether to explore the accompanying explanations in depth or to rely solely on the brief textual descriptions shown in the interface (cf. Figure 4.4 on page 61). Task #6 then invited participants to modify an event on a day for which SIPA4TM was expected to display commute-based advice in the highest stage, thereby illustrating the increased capability and personalisation afforded by higher data access and providing a potentially more enticing implicit cue for scrutiny.

Finally, Task #7 asked participants to create a new event in order to ensure that non-trivial suggestions would be generated, and then explicitly invited scrutiny of SIPA4TM's suggestion explanations. As with earlier tasks, participants were not instructed on how deeply to inspect explanations or model fields, enabling observation of how participants naturally approach explicit scrutability when prompted to do so.

### 5.3.2 Evaluation Approach

Once data collection was completed, all audio recordings were transcribed and anonymised by the main author, utilising the WhisperX Python library [Bain et al., 2023] and following the analogous procedure to that for discussion audio, detailed in Subsection 5.5. The collaborators then shared the transcripts, read them and manually noted statements of interest. A statement of interest was originally loosely defined by the main author to be any comment on at least one of the following:

- **Features:** An expressed sentiment towards a SIPA4TM feature or a comment regarding a feature the participants anticipated but SIPA4TM does not have;
- **Considerations:** Any explicitly stated time-management consideration or constraint participants voiced during their task performance;
- **Task design:** Indications that the user was confused about the user study tasks or what was expected of them to do;
- **Approach:** Explicitly voiced details regarding their approach to solving the tasks or

interacting with SIPA4TM; or

- **Other findings:** Other statements on time management, data control, explanations, scrutability, stages, artificial intelligence, or SIPA4TM that a collaborator finds insightful.

Before sharing the evaluation approach and the transcript with the collaborators, the main author evaluated four of the "thinking-aloud" transcripts and shared his first statements in a shared spreadsheet, with columns for participants [G1P1] to [G9P2] and rows for individual statements. Each non-empty cell in this spreadsheet corresponds to the lines in the transcript file populated by the statements of interest of that type. The collaborators were then invited to take the empty participant columns and populate them in accordance with their own best understanding, and to feel free to add further rows for comments on features, considerations, task design, task-solving approach, or any other findings they considered relevant. The preliminary analysis of thinking-aloud transcripts for [G1P1], [G1P2], [G7P1], [G7P2], [G8P1], and [G8P3] was done by Jeromela, [G2P3], [G3P1], [G3P2], [G6P2], [G8P2], and [G9P1] was done by Gajcin, [G2P1], [G4P1], [G6P3], and [G9P2] by Debnath, [G4P2], [G5P1], [G6P1], and [G7P3] by Murtagh White, and [G2P2] and [G5P2] by Barman.

Finally, once all the columns were preliminarily populated, the main authors reviewed several non-empty cells to verify alignment and merge similar rows into themes. Table A4.2 in the appendix on page 285 provides an aggregate overview of all thinking-aloud statements of interest and the themes and subthemes they were attributed to[2]. The following subsection provides an overview of each theme and subtheme and summarises corresponding conclusions.

### 5.3.3 Overview of Selected Statements

In this subsection, we summarise the interaction findings from the thinking-aloud phase, highlighting representative statements marked by the researchers during coding. In contrast to the focus-group discussions, where participants explicitly reasoned about trust, delegation, privacy, and desired future capabilities, the thinking-aloud recordings indicate what participants thought in the moment of solving the scheduling tasks: what they attended to, what they questioned, and what they treated as "obvious" or not warranting scrutiny. Table 5.4 contains a high-level overview of each theme identified in the thinking-aloud transcripts analysis, and we elaborate on each theme therein below.

**Scheduling Considerations** Across participants, the dominant interpretive lens for interacting with SIPA4TM was not the assistant's internal reasoning per se, but the practical

[2]A full final mapping of the statements of interest to categories is available in the shared Google Sheet: `https://docs.google.com/spreadsheets/d/1NuL6VJVpBSNjkRNj9OKicnLpf3-FNfcpm0C8jeNo1XI/edit?usp=sharing` (last access 2026-03-12)

Table 5.4: Overview of Thinking Aloud Findings: themes, statement counts (SC), and participant coverage

| Theme | Subthemes | SC | Participants | |
|---|---|---|---|---|
| Active Scrutiny | Scrutability | 53 | 19 | all but G2P2; G2P3; G9P2 |
| | User Control (Desired) | 6 | 4 | G2P1; G2P3; G8P1; G9P1 |
| | User Control (Implemented) | 7 | 6 | G2P3; G7P1; G7P3; G8P1; G8P2; G9P1 |
| Interface Friction | Control | 5 | 3 | G2P1; G6P2; G8P1 |
| | Interruption | 12 | 2 | G2P3; G8P1 |
| | Visualisation | 3 | 2 | G6P2; G9P2 |
| Scheduling Considerations | | 59 | 19 | all but G2P3; G3P1; G7P1 |
| Scrutability Barriers | Implicit Scrutiny | 86 | 19 | all but G2P1; G3P1; G9P2 |
| | Interpretability | 53 | 13 | G1P2; G2P2; G4P1; G4P2; G5P1; G5P2; G6P1; G6P3; G7P2; G7P3; G8P1; G8P2; G8P3 |
| | Interpretability (Stages) | 10 | 6 | G2P3; G6P2; G7P1; G7P3; G8P3; G9P1 |
| | Purpose-Driven Scrutiny | 48 | 14 | G1P2; G2P1; G4P1; G4P2; G5P1; G5P2; G6P3; G7P1; G7P2; G7P3; G8P1; G8P3; G9P1; G9P2 |
| Insufficient Autonomy | | 38 | 14 | G1P2; G2P1; G2P2; G2P3; G4P1; G4P2; G5P1; G5P2; G6P1; G6P3; G7P1; G7P2; G7P3; G8P1 |
| Study Artefacts | | 22 | 9 | G1P1; G2P2; G4P1; G5P2; G6P1; G6P2; G7P2; G7P3; G8P1 |
| Trust | | 7 | 3 | G6P2; G8P2; G9P2 |
| Utility of Assistance | Attention Steering | 37 | 11 | G1P1; G1P2; G2P3; G4P2; G5P1; G5P2; G6P1; G6P3; G7P1; G7P3; G9P2 |
| | Contextual Information | 15 | 9 | G1P2; G2P1; G3P1; G5P2; G6P1; G6P2; G6P3; G7P1; G8P2 |
| | Ideation | 15 | 8 | G1P2; G2P1; G4P2; G5P2; G6P1; G6P3; G7P1; G9P1 |

constraints of scheduling and the tacit rules that participants bring to calendar use. Participants approached the tasks with pre-existing, complex scheduling strategies that often dictated their interaction with the assistant. In particular, participants repeatedly reasoned about which events could not be moved, which could be adjusted, and what constitutes a "real" clash. A prominent code within this theme was the explication of social considerations (with 21 relevant

statements coded, expressed by nine different participants, as can be seen in Tables A4.2 and 5.4, shortened as $SC = 21$; $PC = 9$). The participants reasoned about the flexibility of events based on the people involved. For instance, [G1P2] noted, "So George can have a bit of my time, but George is going to have to wait until 1.15[pm], unfortunately", illustrating how users mentally negotiate constraints before or during their interaction with the system[3]. Similarly, in resolving a conflict, [G2P2] immediately grounded their reasoning in event immovability ("if it's a team session I can't really move that either") before attempting a modification.

Furthermore, participants highlighted that apparent conflicts might be intentional, with [G3P1] remarking that overlapping events "might be alright in cases when there's an intentional multi-tasking". Similarly, [G5P2] treated the assistant's clash detection as potentially correct but not decisive, noting that two events labelled as a clash "could happen at the same time", as one "can call during the walk". These interaction traces reflect a broader point later developed in the thematic analysis: time management is situated in the broader context, and the calendar itself is often an incomplete proxy for flexibility, priorities, and social relations (as we discuss later in Subsection 5.5.2).

**Utility of Assistance** The utility of the assistant's core features was frequently explicitly acknowledged. Based on the coded statements, the utility of the advice can be broadly categorised into three groups: attention steering, providing contextual information, and ideation aid.

*Attention steering* ($SC = 24$; $PC = 7$), such as SIPA4TM's clash-reporting, was often described as helpful. Participants often voiced relying on these prompts to navigate the schedule. For instance, [G1P2] stated: "it's telling me I should modify lunch or helping hand or post-lunch walk by the office, so I think, to be perfectly honest, I could probably do without my post-lunch walk by the office", indicating that the system successfully directed attention to problem areas. Furthermore, a few participants voiced relying on the assistant to see if they had missed something, such as [G1P2], who said: "I'm just going to go and check and see if there's any more assistant suggestions for any of the days where there would be clashes where I might have missed them before".

Clash detection in some cases served for task-solving *ideation*, as in the case of [G5P1], "Selecting the day there is a clash. Going for a walk and talking to a sales assistant [are highlighted as clashing]. Post-lunch walks is active if work permits [states the event description], in this case, it doesn't. If I take my post-lunch walk and do it after the sales meeting, because I can't move the meeting with my manager. That's one of the things that shouldn't be moved".

[3]George is the name of a fictional colleague with whom there is a meeting scheduled in the initial configuration of the calendar used for the participants' interactions with SIPA4TM.

This indicates that the assistant served as a decision-making scaffold, even if it did not implement the changes itself. In other cases, participants acknowledged that some system outputs were "correct" or "fair" while questioning whether they were necessary. For instance, [G5P2] agreed with the clash report ("Yes, that's correct") but followed with "it's good that it tells me there would be a clash, but yeah", and later questioned the utility of a day overview panel: "I don't know if this is necessary, though". This points to an important nuance in perceived utility: users may accept that an assistant is *right* while still evaluating it as insufficiently valuable, particularly when it surfaces information they already feel capable of noticing on their own.

Lastly, the analysis of transcripts also identified that participants acknowledged providing context-related information as helpful, specifically when it comes to suggested commute times or previously unnoticed duration changes. For instance, participant [G3P1] failed to notice that the new begin and end times they had entered had changed the total event duration time, which then led to a correction-eliciting pop-up advice: "So, set that from half ten to half twelve. Failed to modify? Um, okay. Reset to original duration, two hours. Was that not two hours? So, half ten, half twelve. Okay, modified successfully".

**Active Scrutiny** Participants demonstrated a desire to understand and control the underlying logic of the assistant, though their ability to do so varied. Below, we first provide an overview of insights regarding participants' verbalised thoughts during active engagement with the explanations and other scrutability mechanisms, as well as statements during their attempts to control the underlying user model and the assistant's behaviour. The latter group of statements is split between the control mechanisms that SIPA4TM already implemented and control mechanisms that participants voiced as desired or expected, but which had not been implemented in this version of SIPA4TM.

In most sessions ($SC = 23$; $PC = 11$), participants explicitly attempted to delve into deeper levels of explanations to scrutinise the assistant's personalisation and suggestion-creation logic. For example, [G2P2] described the interaction in terms of exploring options and then deliberately inspecting the explanation: "See options below. Explore a level deeper" until they were satisfied with understanding of the information used for the suggestion "only start and end times of an event are to be assisted, and it does not use any other information about you or your events". This finding indicates is engaged when users are diagnosing what the system is (and is not) taking into account. On the other hand, [G7P1], [G8P2], and [G9P1] used scrutability features early in their interactions to gauge the assistant's scope and capabilities before actively engaging with it. There was significant interest in the assistant's learning mechanism, with participants attempting to deduce how the system modelled their preferences, evident also from [G1P2]'s expressed desire for more specific transparency: "it would be

interesting if it specifically was like, well, usually we play video games for two hours, so you're going to need more time, basically". Additionally, transcripts indicate that participants were interested in SIPA4TM's capability to map event and travel times ($SC = 19$; $PC = 8$), with [G2P1] realising this feature after seeing a suggestion that there is not enough time for commute between two events and then exploring the corresponding explanations: "If we explore a little deeper, yet we see that there is insufficient time, explore a little deeper, there lies on... Okay, so that's a valid, so that's something we didn't actually know. So that's actually helped us there". Several participants, including [G1P1], [G3P1], [G3P2], and [G6P3] voiced their interpretation attempts regarding the visual explanation (second level). Regarding the clarity of the visualisation, [G2P3] stated "I mean this graph is pretty intuitive, it kind of shows where the clash is. And then this length explanation is the most intuitive", but this impression was not universal, with [G1P2], [G7P1], [G7P2], and [G8P1] only shortly acknowledging the graph before moving on, implying disinterest or a belief that it is not worth their time to engage with the graph, as we will discuss further below.

The *implemented user control* of stages and user model values was not fully understood by all participants. Among those who did, [G7P1] and [G8P1] wanted to change personalised suggestion, with the former interested in changing the sleep times and the later on commute time between two locations. The implementation of stages as a control mechanism was explicitly appreciated during interaction by [G2P3] and [G7P3]. [G7P3] even provided an explicit breakdown of their understanding: "By increasing the stage, we can let the assistant know more about us in return for a more personalised advice". Meanwhile, [G2P3], systematically explored all stages, "we'll go back to stage 1 for the time being and then we'll go up", concluding "that's interesting that it has different levels of kind of contextual clues". However, even they anticipated greater potency of stage control, especially with regard to interruption tuning, stating: "It still has that kind of slightly annoying thing where it stops you if the duration is different".

This was not the only example of participants *wanting or expecting more potent user control*. Frequently, participants anticipated that increasing stages would lead to more direct action, and commented when this did not occur or was not graspable. [G8P1] articulated this frustration clearly: "I feel like I'm going around in a circle here though, because I want to give it permissions control to make, to make the changes, but I can't". Conversely, others defaulted to a strategy of maximising permissions, hoping to minimise effort, with [G2P1] stating, "Maybe we'll keep it on maximum because why not". Similarly, [G9P1] used the highest stage, hoping that this would result in more proactive and personalised suggestions than what was implemeted in SIPA4TM: "To be honest, the more context it has, I'm happier with, so I was happy to use the three stars [i.e., the highest stage], because if I'm kind of, the less I'd have to think about it, the better, if it gets busier and busier. And that's where I like, kind of, if it thinks about,

oh, well, you're not a morning person, so maybe don't put the event there".

**Insufficient Autonomy** A major theme identified was the gap between the assistant's suggestions and its ability to execute them, with many participants explicitly stating during the interaction that they had expected the SIPA4TM to do more on its own ($SC = 19$; $PC = 10$). Among such expressed sentiments, the desire to have the assistant automatically make changes, such as resolving clashes, was the most frequently identified. [G1P2] voiced this disappointment clearly: "I was expecting perhaps some sort of functionality where it would change it automatically for me and solve the clash, or at least make a suggestion of like a different time that would solve the clash and that didn't happen". Similarly, [G2P1] remarked, "So unfortunately, it doesn't seem that the assistant can actually do the changing of the schedule for us, which would be quite nice". Other desired cases of more advanced SIPA4TM autonomy or proactivity included automatically adjusting end times to maintain event duration [G2P3,G8P1], prioritising events in its suggestions [G7P2], tracking numbers of specific event types scheduled over the week [G7P1], balancing the user's workload over the week [G2P3,G7P3], and even auditing logged times and managing multiple calendars [G4P1]. Notably, [G7P2] and [G8P1] stated during interaction that they wished there had been an "accept button", which would make the assistant act on its suggestions, a feature that had been intended for SIPA4TM but was ultimately not implemented, as elaborated in Section 4.2. Significantly, where participants did articulate gaps in the assistant's behaviour, these comments were typically framed as the system failing to take on enough of the burden, rather than failing to provide enough explanation. [G4P1] provided a characteristic example: after encountering multiple clashes, they described the situation as one in which the assistant is "pushing me on", yet still leaves the participant with the difficult reasoning work: "there's a little bit of cognitive load on this".

**Scrutiny Barriers** While scrutability was a design goal, several barriers prevented effective engagement with explanations. The thinking-aloud recordings also make clear that scrutiny is constrained by effort, interpretability, and the perceived relevance of inspecting deeper levels.

Participants' interest in scrutinising SIPA4TM was often *purpose-driven*, that is, tightly coupled with immediate task progress. We found 20 statements by 6 participants in which they opened explanations or inspected panels when they expected this would help them proceed to the next immediate step, not as a general practice. [G5P2] exemplified this reactive scrutiny when confronted with an unexpected prompt: "Why are you telling me this? I'll just click on it. Yeah, but I'm not quite sure, like, why". Conversely, when suggestions were not relevant to the user's immediate goal, they were ignored, as [G1P2] noted: "That's nothing to do with what I'm working on at the moment, so I'm not going to bother". Similarly, [G1P2], [G5P2], and

[G8P1] stated they were not interested in details unless they are task-relevant, while [G4P1] remarked they weren't interested in explanations but getting their tasks done.

*Interpretability issues* appeared both at the level of assistant output and at the level of UI cues. We identified 33 statements by 8 participants that indicate confusion regarding a feature of the assistant, and 22 statements by 7 participants that indicate labour to parse the meaning of that feature. For example, in attempting to follow the assistant prompts, [G4P1] noted: " 'Select an option below'. I don't see an option", indicating how UI issues can block users' interpretability or correction attempts. Participants sometimes struggled to parse the assistant's explanations, with [G7P2] expressing feelings of inadequacy when seeing the visual explanation: "Explore on a deeper level, similar graph. I'm not a huge statistics person".

We also identified *interpretability issues regarding stages*, that is, the manner in which SIPA4TM implements permissions control ($SC = 10$; $PC = 6$). For instance, [G7P3] admitted confusion, asking, "So this increases the stage to the maximum? I don't understand", thereafter answering the question on what data the assistant uses by guessing based on behaviour, rather than by engaging with stage and suggestion explanations. Similarly, [G9P1] failed to see the difference between the highest and the default stage, concluding "I can't really see a difference in the permission". [G2P3] struggled in seeing the difference between stages as well, while [G8P3] failed to understand what information is needed for each stage.

**Implicit Scrutiny** Our analysis of thinking-aloud transcripts has revealed a phenomenon that we call *implicit scrutiny*, to which 86 statements stemming from all participants except [G2P1], [G3P1], and [G9P2] were attributed. Namely, when participants encounter unexpected IPA behaviour or, in some cases, are explicitly prompted to do so by a task, they would scrutinise the IPA indirectly.

One such indirect strategy is attempting to "resolve" the SIPA4TM's suggestion through trial and error ($SC = 23$; $PC = 15$), which is seen, for instance, in the approach of [G8P3]: "Okay, we can just change your timing a little bit, so instead of 6:45 we just [put] 18:30 till 19:30 and then we have 30 minutes. Let's see... So, we moved it – now it's fine". A more direct approach would be looking at the suggestion description or explanation, which provides the alternative times and reasoning behind suggestions, such as dismissing the interruption-causing suggestions by clicking the save button again.

Similarly, several participants ($SC = 20$; $PC = 13$) engaged in mental calculation, such as that by [G4P2]: "Now looking at the family cooking class, which is at 6 to 7.30 on the 27th of August and then a video game session that is from 7.45 to 8.45. I'm just going to move the video game session by 15 minutes to get back home and change it to, and then still have ample time. So instead of 7.5 to 8.25, it'll be 8 to 9". Such calculations were in many

examples unnecessary, and the suggestion could be more directly examined through looking at the explanations or even modifying the user model to shorten the minimum amount of SIPA4TM expects the user to allocate for event types (in this case, for hobbies such as video games).

Furthermore, the analysis identified cases ($SC = 16$; $PC = 7$) where assumptions about the IPA behaviour and knowledge were constructed through suggestion-based extrapolations rather than scrutiny through explanations or user model, which would reveal a faithful insight. An instance of such a behaviour can be read in highlighted statements by [G6P1]: "Okay, so there's a clash at half-ten and half-ten, so I'll just make the focus work, make sure there's enough of a clash. I assume the AI is going to get mad at me if I reduce the time of the event. One thing I will say is that if it's like getting mad at you for changing the period of time, you're going to do something that seems a little bit controlling for a computer to tell you what to do. That's just me".

Lastly, when solving tasks four and seven, which aim to nudge participants to look at explanations and permission control stages, several participants instead extrapolated answers based on their own interaction-based inferences about the SIPA4TM's functioning. As we will discuss in Subsection 5.4.1, in some case this has resulted in participants like [G8P3] avoiding the pre-intended scrutiny channels entirely ("Okay, so what it seems like is that there are particular time blocks for every day and the assistant identifies that if it's just a work on a particular location or we have to physically move to a different location") or where the researchers intervened explicitly to direct a participant to look at explanations, as had happened in the case of the collaborating researcher observing [G4P1]'s interaction ("I will remind you, even though you may or may not need this, but there are multiple levels of the type of information that the system provides. So you can take a look at those if you get in trouble").

**Interface Frictions** Further statements reflected friction with interaction mechanics rather than assistant reasoning. These can be divided into expressed issues with SIPA4TM's control mechanism, interruptions, and visualisations.

Issues with basic *control mechanisms*, such as the inability to right-click to create events, were noted ($SC = 5$; $PC = 3$). In early use, a few participants occasionally struggled with the basic event-creation workflow and expressed uncertainty about what actions were available. For example, [G9P2] repeatedly attempted to infer the intended interaction pattern, such as: "Do I not just click there and do create new event? No. /.../ Maybe a right click", and expressed frustration: "I can't even do the first bit". Relatedly, [G8P1] was frustrated by the inability to make changes by dragging events over the grid: "Failed to modify. [sigh of frustration] It's... Okay. You can't click on the actual column it is. You have to create a new

event over to the right".

As evident from the timing of the sigh of frustration in [G8P1]'s response, one cause of UI friction was the control-flow *interruptions* caused by context-triggered suggestions ($SC = 12$; $PC = 2$). This had partially been anticipated during SIPA4TM implementation (cf. Subsection 4.2.3). [G8P1] reacted particularly negatively to the system's enforcement of constraints, stating, "Well, if it's my calendar, that's what I want it to do when it's not letting me do it", eventually even abandoning their modification attempt. Other participants, including [G2P3], voiced a similar sentiment. In their frustration with the pop-ups, participants often did not read the text of the context-based suggestion, which instructed them that they may also proceed without making changes by clicking the "Save" button again.

Two participants found the *visualisation* of the schedule lacking ($SC = 3$; $PC = 2$). [G6P2] commented events should be more clearly differentiated at a glance, exemplifying the case of different projects, saying: "I'm assuming that our project works are... They probably could be a little bit more better articulated in the calendar, so that it was, at a glance, you would have a feeling that you're dividing your time between the two projects. It doesn't seem to be too clear from the way the calendar is organised or the events were named at the moment". Similarly, [G9P2] voiced confusion regarding the overall layout of SIPA4TM to "Anyway, I'll probably need to talk to this as I go, but I found sometimes maybe the notification didn't work or it didn't happen and I missed something". Such remarks are minor in isolation, but they highlight a recurring observed phenomenon: when explanations or model views compete with the task-relevant representation (the calendar), the users' attention may be diverted toward interface management rather than toward verification or decision-making.

**Study Artefacts** Finally, some findings were attributed to the study design constraints ($SC = 22$; $PC = 9$). Researchers highlighted 10 statements that indicated that 6 participants, specifically [G1P1], [G5P2], [G6P1], [G6P2], and [G7P2], were struggling with the constraints of the task. In particular, participants were working with an artificial schedule belonging to a fictional persona, which occasionally triggered meta-level comments about the difficulty of reasoning without personal context. [G4P1] explicitly articulated this mismatch: "I would say there's a little bit of cognitive load on this. But it's not my diary, so why would I... I'm literally being someone else's personal assistant in this act". More generally, 10 highlighted statements indicate that 5 participants were confused about tasks themselves: [G2P2], [G4P1], [G5P2], [G6P1], and [G8P1]. These include assumptions such as that voiced by [G8P1] that they ought to change their behaviour to fit perceived system expectations: "Personally, if I had lunch before a doctor's appointment, I'd just leave halfway through my lunch and go to a doctor's appointment, but for the purpose of the experiment, we will change that and make lunch from 1 to 1.30". More generally, some of the participants did not engage with stages

due to a divergence from the study protocol by a collaborating researcher, as elaborated in Subsection 5.4.1. These statements are important to acknowledge because they suggest that some perceived cognitive load may be attributable not only to SIPA4TM's interaction design, but to the intentional removal of autobiographical knowledge that participants normally rely on when planning their own time.

**Trust** Explicit discussions of trust were less frequent in the thinking-aloud protocol ($SC = 7$; $PC = 3$) than in the focus groups, likely due to the concurrent nature of the task. However, instances of anthropomorphisation were observed, with [G6P2] noting, "it's nice to know that you're working really hard from an AI perspective". Conversely, a broad scepticism about the AI's capability was also expressed, as [G9P2] questioned a suggestion: "So I'm just wondering why this is wrong".

## 5.4 Logs-Based Quantitative Grounding

While the verbal protocol analysis in the previous section provides a qualitative overview of what participants attended to during task completion, we additionally rely on interaction logs as a quantitative grounding of how participants engaged with SIPA4TM's scrutability mechanisms in practice. So, while the think-aloud protocols provide insight into the *intent* behind scrutiny, the logs provide an objective record of the *extent* of that scrutiny, while the thematic analysis of the focus group discussions in Section 5.4 identifies underlining *causes and considerations* driving scrutiny. Table 5.5 aggregates log data on:

1. how often participants opened the *suggestion explanations* at each depth level (explanation levels 1–3; cf. Section 4.2 and the table on page 58 for an overview of all explanation levels and the reasoning behing them);
2. how often they opened the *stage explanations* (for each of the stages 0–2; cf. Figure 4.4 on page 61);
3. how often they inspected the *user model view* (for each of the stages 0–2; cf. Figure 4.5 on page 62); and
4. the distribution of participants' actions across the three permission control stages.

Across participants, the logs indicate that explanation access was common, yet unevenly distributed and strongly shaped by both task progress and stage selection. The data reveals a distinct "funnel of scrutiny" regarding the assistant's levels of explanations. There is a marked preference for surface-level information, with the number of visits for each sequential explanation depth level reducing roughly geometrically from the previous. Most participants opened

Table 5.5: Logs quantification: Opened suggestion explanation levels, stage explanations, user model views, and percentage of actions done under each stage

| Participant | Suggestions | | | | Stage Expl. | | | | User Model | | | | Totals Sum | Used Stage [%] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level 1 | Level 2 | Level 3 | Total | Stage 0 | Stage 1 | Stage 2 | Total | Stage 0 | Stage 1 | Stage 2 | Total | | Stage 0 | Stage 1 | Stage 2 |
| [G1P1] | 11 | 4 | 1 | 16 | 2 | 4 | 4 | 10 | 0 | 1 | 1 | 2 | 28 | 0.0 | 53.4 | 46.5 |
| [G1P2] | 14 | 2 | 1 | 17 | 7 | 8 | 3 | 18 | 1 | 2 | 2 | 5 | 40 | 0.0 | 76.0 | 24.0 |
| [G2P1] | 7 | 3 | 2 | 12 | 1 | 4 | 3 | 8 | 0 | 0 | 1 | 1 | 21 | 0.0 | 33.6 | 66.3 |
| [G2P2] | 11 | 10 | 5 | 26 | 6 | 14 | 8 | 28 | 1 | 3 | 3 | 7 | 61 | 5.8 | 66.5 | 27.6 |
| [G2P3] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | 100.0 | 0.0 |
| [G3P1] | 14 | 7 | 5 | 26 | 0 | 2 | 1 | 3 | 0 | 0 | 0 | 0 | 29 | 0.0 | 56.3 | 43.7 |
| [G3P2] | 2 | 1 | 1 | 4 | 0 | 1 | 2 | 3 | 0 | 0 | 1 | 1 | 8 | 0.0 | 40.6 | 59.4 |
| [G4P1] | 9 | 5 | 4 | 18 | 0 | 4 | 1 | 5 | 0 | 1 | 0 | 1 | 24 | 0.0 | 84.2 | 15.8 |
| [G4P2] | 3 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0.0 | 100.0 | 0.0 |
| [G5P1] | 0 | 0 | 0 | 0 | 1 | 3 | 3 | 7 | 0 | 1 | 1 | 2 | 9 | 0.0 | 72.8 | 27.2 |
| [G5P2] | 6 | 4 | 0 | 10 | 2 | 3 | 4 | 9 | 0 | 1 | 2 | 3 | 22 | 5.0 | 48.3 | 46.7 |
| [G6P1] | 8 | 1 | 0 | 9 | 1 | 1 | 1 | 3 | 0 | 1 | 0 | 1 | 13 | 0.0 | 90.2 | 9.8 |
| [G6P2] | 5 | 0 | 0 | 5 | 2 | 4 | 2 | 8 | 1 | 1 | 1 | 3 | 16 | 7.6 | 57.6 | 34.8 |
| [G6P3] | 9 | 3 | 2 | 14 | 0 | 2 | 2 | 4 | 0 | 1 | 1 | 2 | 20 | 0.0 | 58.6 | 41.4 |
| [G7P1] | 7 | 3 | 1 | 11 | 2 | 4 | 8 | 14 | 1 | 2 | 5 | 8 | 33 | 0.0 | 41.6 | 58.4 |
| [G7P2] | 11 | 5 | 1 | 17 | 2 | 6 | 6 | 14 | 1 | 3 | 4 | 8 | 39 | 0.0 | 62.8 | 37.2 |
| [G7P3] | 4 | 2 | 0 | 6 | 0 | 6 | 0 | 6 | 0 | 5 | 0 | 5 | 17 | 0.0 | 100.0 | 0.0 |
| [G8P1] | 12 | 4 | 0 | 16 | 0 | 7 | 1 | 8 | 0 | 5 | 2 | 7 | 31 | 0.0 | 93.2 | 6.8 |
| [G8P2] | 17 | 10 | 4 | 31 | 0 | 3 | 6 | 9 | 0 | 1 | 6 | 7 | 47 | 0.0 | 38.8 | 61.2 |
| [G8P3] | 8 | 3 | 1 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 0.0 | 100.0 | 0.0 |
| [G9P1] | 4 | 2 | 0 | 6 | 4 | 6 | 9 | 19 | 1 | 2 | 4 | 7 | 32 | 1.0 | 14.6 | 84.4 |
| [G9P2] | 7 | 3 | 0 | 10 | 0 | 2 | 2 | 4 | 0 | 2 | 2 | 4 | 18 | 0.0 | 66.7 | 33.3 |
| **1st Qt** | 4.25 | 1.25 | 0 | 6 | 0 | 2 | 1 | 3.25 | 0 | 0.25 | 0 | 1 | 13.75 | 0.0 | 49.6 | 11.3 |
| **Median** | 7.5 | 3 | 1 | 11.5 | 0.5 | 3.5 | 2 | 7.5 | 0 | 1 | 1 | 2.5 | 21.5 | 0.0 | 64.7 | 34.1 |
| **3rd Qt** | 11 | 4 | 1.75 | 16.75 | 2 | 5.5 | 4 | 9.75 | 0.75 | 2 | 2 | 6.5 | 31.75 | 0.0 | 88.7 | 46.6 |

suggestion explanations multiple times (median total = 11.5), typically concentrating their inspection at Level 1 (median = 7.5), with lower but still non-negligible engagement at deeper levels (median Level 2 = 3; median Level 3 = 1). This pattern suggests that participants frequently sought quick, surface-level justification or clarification, and is thus aligned with the qualitative finding that users are primarily interested in immediate error correction rather than deep architectural understanding – they scrutinise the immediate suggestion cause (Level 1) but rarely investigate the underlining logic and user control mechanisms (Level 3). We also interpret these results as consistent with the later thematic finding of maximising the prospect of time and effort, as will be further elaborated in Subsection 5.5.1: participants were willing

to "check" the assistant, but not to repeatedly invest effort into the most demanding forms of explanation unless needed.

Engagement with *stage explanations* was also widespread (median total = 7.5), indicating that many participants actively explored the meaning of the permissions control and what additional access would entail. However, the logs also show that *user model inspection* was comparatively rare (median total = 2.5), with a first quartile of only 1 user model view, and with Stage 0[4] user model views effectively absent for most participants (median Stage = 0). This provides an important complement to the qualitative accounts: even though participants frequently discussed the importance of understanding and verifying system behaviour and later articulated strong views on scrutability and control, the act of opening the explicit user-model representation was not a dominant behaviour during task completion. In other words, much of the observed verification likely occurred through direct inspection of calendar outcomes and suggestion text (i.e., implicit scrutability), rather than through explicit inspection of the differences in underlying model comprehensiveness.

When looking at the cumulative engagement with suggestion explanations, i.e. summing all suggestion explanation levels, stage explanations, and user model views across all stages, we observe a wide variance in behaviour, with a Median of 21.5 total interactions per participant. This aggregate metric may serve as a proxy for the user's overall interest in scrutability. Consequently, this column is used for statistical significance in relation to survey answers in Subsection 5.4.2.

Finally, Table 5.5 demonstrates that participants' actions were overwhelmingly performed under Stage 1 and Stage 2, with Stage 0 being nearly unused for most individuals (median Stage 0 proportion = 0.0%). The median proportion of actions conducted under Stage 1 was 64.7%, with Stage 2 comprising the remaining 34.1% (median). Apart from the outliers discussed in subsection 5.4.1, all participants operated with at least some form of modelling enabled (Stage 1). This is not surprising given that Stage 1 was the default and that participants were not explicitly instructed to change it until task #4. Similarly, the fact that many participants additionally engaged with the highest stage (Stage 2) for a substantial portion of their interaction is also consistent with the fact that leaving the highest stage as the option was the likely "low-effort" choice after task #4. At the same time, there was considerable variability between participants in how much they relied on Stage 2, ranging from near-zero to dominant use (e.g., [G9P1] conducting 84.4% of actions under Stage 2). This variability anticipates the later thematic discussion of stages as an expectation-control mechanism that different participants approached with different risk tolerances and delegation preferences and the fact

[4] As we have elaborated in Subsection 4.2.1 and shown in Figure 4.2 at page 57, the default permission control stage was 1, with each of the stages 0, 1, and 2 represented graphically in SIPA4TM's assistant component as ★☆☆, ★★☆, and ★★★ respectively.

that some participants explored the system before starting with the tasks (cf. Subsection 5.5.2).

### 5.4.1 Outliers

From Table 5.5 it can be noted that [G4P2], [G7P3], and [G8P3] form outliers in their stage distribution ratios, as they had not received any suggestion with permission control set to a stage other than the default stage 1 (i.e., ★★☆). Following this observation, the main author has re-examined the thinking-aloud transcripts, which also include the onboarding and moderation by a collaborating researcher tasked with observing the think-aloud interaction. This review revealed that all three participants had shared the same collaborating researcher, who had deviated in their understanding of the study protocol, reproduced in Appendix A4.3, relative to other researchers. Namely, as we discussed in Subsection 4.3.2, the fifth task explicitly asks the participants to explore the SIPA4TM's permissions control. However, from the transcripts, it can be inferred that the three participants had not fully understood the intent of the task, and the collaborating researcher had not steered them towards discovering that they can change permissions explicitly by clicking on the stars in the assistance component. Instead, the collaborating researcher moderating all three participants chose to maintain the general principle of non-interference, rather than to exceptionally clarify the rather specific intent of the fourth task, as other collaborating researchers and the main researcher had done where necessary. We thus conclude that the outlying used stage ratios for the three participants stem from a systemic fault with the realisation of a part of the study, rather than a semantically meaningful deviation from the typical interactional pattern.

Additionally, [G2P3] is an outlier with near-zero recorded interactions with explanations. However, this was due to a significant accessibility-related factor revealed from the thinking-aloud study transcripts. It should be noted that our study call had not excluded applications from participants with disability. This was a conscious choice, as capturing perspectives of participants with diverse experiences and requirements was a desired outcome. For this reason, we include the answers from the survey and the insights shared during the discussion by both [G2P3] as well as by [G4P2], [G7P3], and [G8P3] in our qualitative analyses, but we exclude the four participants from the statistical significance testing in Subsection 5.4.2 below.

### 5.4.2 Statistical Significance Testing

To complement the qualitative insights derived from the verbal protocol analysis, we performed statistical tests on the relation between the participants' self-reported attitudes (gathered via the pre-task survey) and their actual interaction with SIPA4TM (recorded via interaction logs). Specifically, we sought to determine whether participants who reported higher levels of interest

in AI, higher importance of privacy and explanations, or more experience with GDPR personal data retrieval have demonstrated a statistically significant increase in their engagement with the system's scrutability features.

We mapped the participants' survey responses into three ordered groups as follows: *Low*, *Med*, and *High* importance/interest. For the data requests questions, the grouping was done by mapping the answer "Yes, multiple times." into the group "High", the answer "Yes, once." into the group "Med", and the answers "Unsure, I don't remember." or "No, never." into "Low". For the remaining three questions, "High" corresponded to answers with a prefix "Very important" or "Very interested", "Med" comprised participants who chose answers beginning with "Somewhat important" or "Rather interested", while every other answer was categorised as "Low". The dependent variable was the "Total Sum", i.e. the aggregate metric summing all interactions with the Suggestions, Stage Explanations, and User Model components, which can be seen in Table 5.5. Figure 5.3 contains box plot visualisations for the four questions and the corresponding groups, with individual responses plotted as dots within each box plot.

For this analysis, we employed the Jonckheere–Terpstra (JT) test by Jonckheere [1954] and Terpstra [1952] as implemented in Python by Bereslavsky [2025]. The JT test is a non-parametric statistical test designed to detect ordered differences between classes of independent samples. Unlike the Kruskal-Wallis test, which tests only for a general difference between groups, the JT test is specifically sensitive to an ordered alternative hypothesis (e.g., *Low* $<$ *Med* $<$ *High*). This made it particularly suitable for our study, where the independent variables (survey responses regarding the importance of Explanations, Privacy, and AI Interest) are ordinal. JT operates by aggregating pairwise comparisons across groups: for each observation in a lower-ranked group, it counts how often it is smaller than (and, with tie-handling, equal to) observations in higher-ranked groups.

Formally, the JT null hypothesis states that the distributions of the outcome (explanation and user model views) are identical across the ordered groups (Low, Med, High), i.e. there is no monotonic trend between the groups. The JT alternative is that there is a monotonic increasing tendency in views from Low to Med to High. Intuitively, if the outcome systematically increases from Low to Med to High, the JT statistic becomes large; if there is no monotonic ordering, the statistic remains near what would be expected under the null hypothesis of exchangeability across groups.

Raw statistical significance results (shown as "JTP p" in Figure 5.3, found evidence of rejecting the null hypothesis in the case of Data Requests, at $p = 0.0231$, while other categories remained well above the significance threshold of $\alpha = 0.05$. However, after correcting for the four repeated hypothesis tests using the Holm procedure, the Data Request category also became non-significant, albeit remaining weekly significant relative to $\alpha = 0.1$ with the

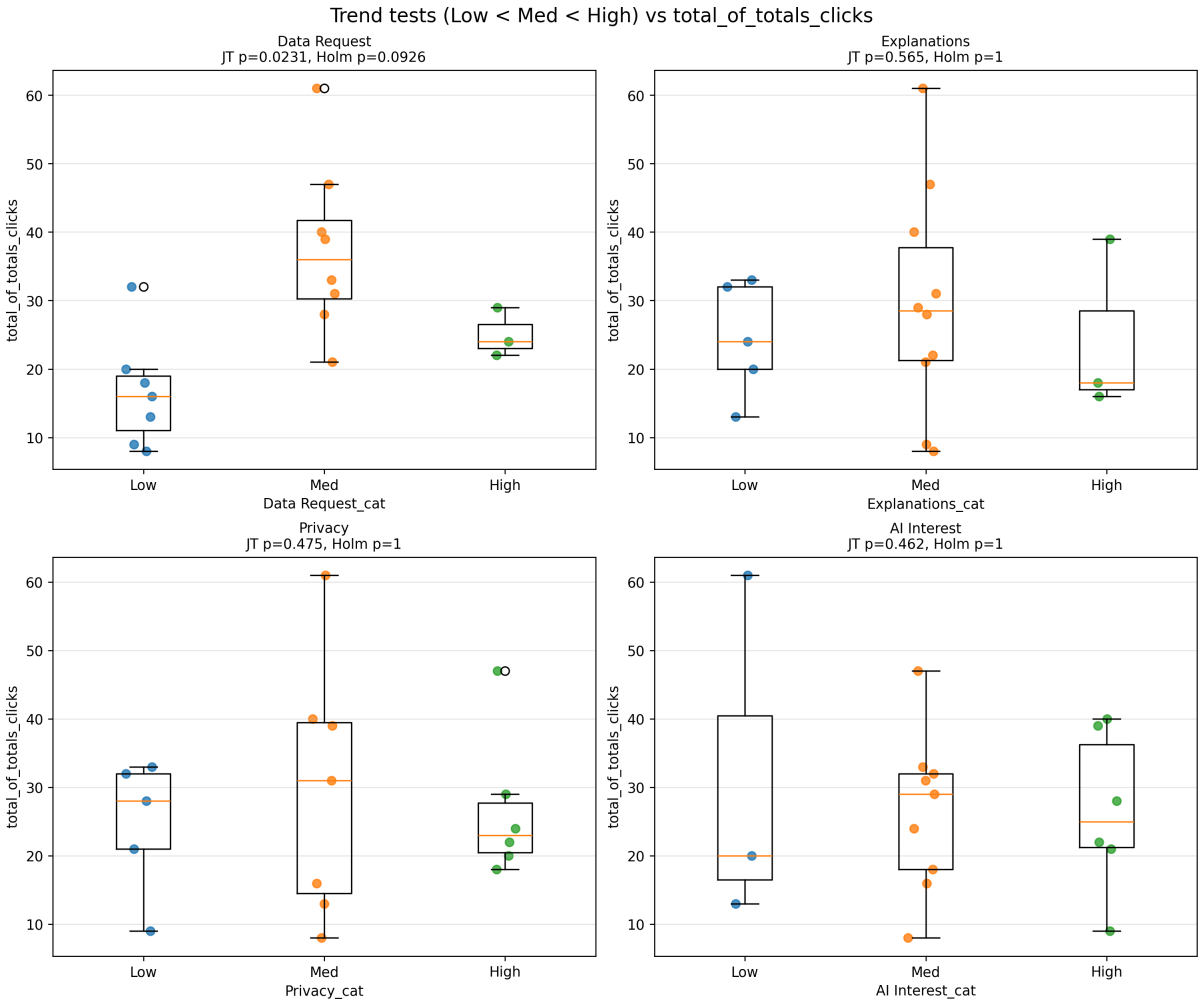


Figure 5.3: JT test p values before and after Holm-corrections above respective box plot visualisations of totals sums grouped by the four survey questions on the interest in AI, importance of privacy, importance of explanations, and the number of data access requests sent

Holm-adjusted $p = 0.0926$. In other words, within this dataset, participants who stated that explanations or privacy mattered more to them, or that they were more interested in AI, did not scrutinise SIPA4TM significantly more frequently than those who rated these aspects as less important.

Regarding the finding of the JT test that indicated a weakly statistically significant monotonic trend for the past of the GDPR data request category, an inspection of the distributions in Figure 5.3 reveals that the effect is primarily driven by a separation between the Low group and the remaining participants, rather than a strictly increasing pattern from Low to Med to High. In particular, the Med group exhibits higher view counts than the High group, suggesting a non-linear or peaked relationship. This highlights that JT detects ordered dominance in aggregate pairwise comparisons rather than requiring strictly increasing medians across all adjacent groups.

These remaining three null results align with the "transparency paradox" often observed in Human-Computer Interaction (HCI) research, where users frequently state a strong desire for transparency and control but rarely utilise these features when they are provided, as surveyed by Kokolakis [2017] and further reported by Colnago et al. [2023]. While the lack of significance suggests a discrepancy between reported attitude and actual behaviour, we must also consider the limitations of the study design.

The absence of strong significance must be interpreted with caution. First, the survey categories used here were coarse and may not have captured the specific contingencies that motivate explanation use in practice (e.g., uncertainty, perceived stakes, or moments of surprise). Second, the analysis is statistically underpowered due to the small number of participants ($N = 18$ after exclusions explained in Subsection 5.4.1), which results from a conscious focus on a study design that favours a qualitative approach with fewer participants and a long opinion-sharing phase. Third, the interactions with SIPA4TM occurred in a time-bounded study setting with scripted tasks, which may not reflect how explanation needs occur in longitudinal, personally consequential time-management practice.

## 5.5 Thematic Analysis

To analyse the output from focus groups – the main qualitative data source of our study – we performed thematic analysis following the steps by Braun and Clarke [2006]: 1) transcribing the recordings and familiarising ourselves with the data; 2) systematically generating initial codes; 3) searching for potential themes; 4) reviewing themes; 5) defining and naming the themes; and 6) producing the final analysis and the report. One of the foci of our analysis was understanding the participants' stance on the assistant proactively offering suggestions as well as on the offered scrutiny procedures. Figure 5.4 outlines our approach.

The thematic analysis presented in this section is grounded in semi-structured focus group discussions conducted as the third phase of the user study, following the protocol detailed in Section 4.3. In short, following individual interactions with SIPA4TM, participants took part in moderated group discussions (2–3 participants per group) that explored their experiences with the system and their broader views on intelligent personal assistants for time management. The discussion guide covered four interrelated areas:

1. perceived utility and shortcomings of SIPA4TM and comparable commercial IPAs;
2. attitudes towards scrutability, including willingness to inspect user models and reasoning behind personalised suggestions;
3. views on proactivity, delegation, and the staging of permissions as a data stewardship

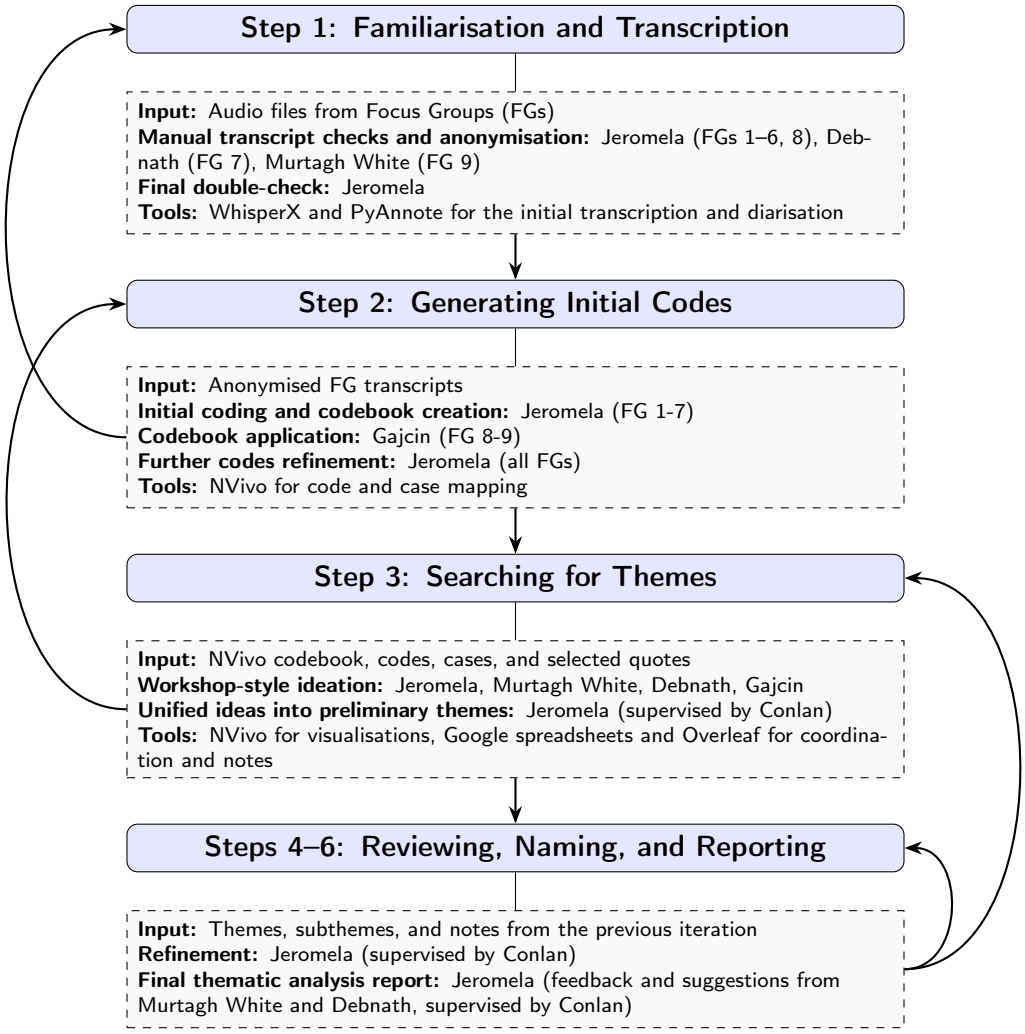


Figure 5.4: An outline of our thematic analysis approach for the focus group discussions; returning arrows indicate iterative improvement of previous steps

mechanism; and

4. factors influencing willingness to use, especially with regard to the proactive and delegative aspects of the assistance.

Question formulation was informed by prior qualitative studies of IPAs – primarily by Cowan et al. [2017] and Zargham et al. [2022] – and shaped by the challenges we envision in Section 3.3.1. In line with established guidance on semi-structured focus groups, including by Harrell and Bradley [2009], the main researcher, who moderated all focus groups, adapted question content, phrasing, and order to support natural discussion flow, while remaining within time constraints. All focus group discussions were audio-recorded.

Audio recordings of the focus group discussions were transcribed using the Python library

WhisperX by Bain et al. [2023] and diarised[5] using PyAnnote 3.1 by Bredin [2023]. Afterwards, groups 1–6 and 8 were manually audited by the principal researcher (i.e. the dissertation author, Jovan Jeromela), focus group 7 was audited by Alok Debnath and discussion 9 was audited by Matt Murtagh White. The audit involved reading the automatically diarised transcripts while listening to the audio recordings to find and correct transcription and diarisation errors.

While auditing the transcripts, we also removed all personally identifiable details and renamed participants in line with the naming scheme. The naming scheme used is to replace the participant's chosen acronym into GxPy, where x is the sequential number of the focus group (i.e. a digit 1–9), and y is the sequential number of the participant (which could be either 1 or 2 if the group had two participants, or 1, 2, or 3 if it had three). Finally, all transcripts were audited and double-checked one more time by Jovan Jeromela, by listening to the audio recordings and comparing them with the previously transferred transcripts in a format expected by NVivo. Transferring files in the format expected by NVivo involved removing timestamps and merging consecutive statements from the same participant into a single line of text to allow automatic attribution of statements to their respective participants. While this transformation indeed leads to loss of temporal information, this was deemed as appropriate, since the thematic analysis procedure as outlined by Braun and Clarke [2006, 2019] does not include a consideration of the timing of utterances. This concluded step 1) of the thematic analysis.

The next step was the generation of initial codes, which Jeromela completed during the preliminary coding of transcripts for focus groups 1–7. Jasmina Gajcin used this codebook to code transcripts of focus groups 8 and 9. The codes were further refined by Jeromela during the theme creation steps and during the supervisor discussions. Steps 2–3 of the thematic analysis were conducted in NVivo, a qualitative analysis software widely used in the field[6].

The third step, comprising the initial search for themes, was initiated by a workshop-style ideation meeting, at which Jeromela, Gajcin, Debnath, and Murtagh White received copies of the codes created in the previous step and then proposed their initial theme candidates. Over the course of multiple weeks and several revisions under the supervision of Conlan, as well as a few informal exchanges with the other collaborators, a set of potential themes was created by Jeromela. This set of themes can be seen in Figure 5.5.

The preliminary themes centre on trusting the assistant enough to use it for their time management, and in particular, trusting to proactively change the calendar or delegate the user's communication with others. We found that participants discussed the influence of factors such

[5]Speaker diarisation is the process of identifying speakers in an audio recording, allowing for an attribution of transcribed lines of text.

[6]The final NVivo file, alongside exports in spreadsheet forms are available at drive.google.com/drive/folders/1w8bkXw2WZDjMR0gKgZ6Bu8pkagID0OFg.

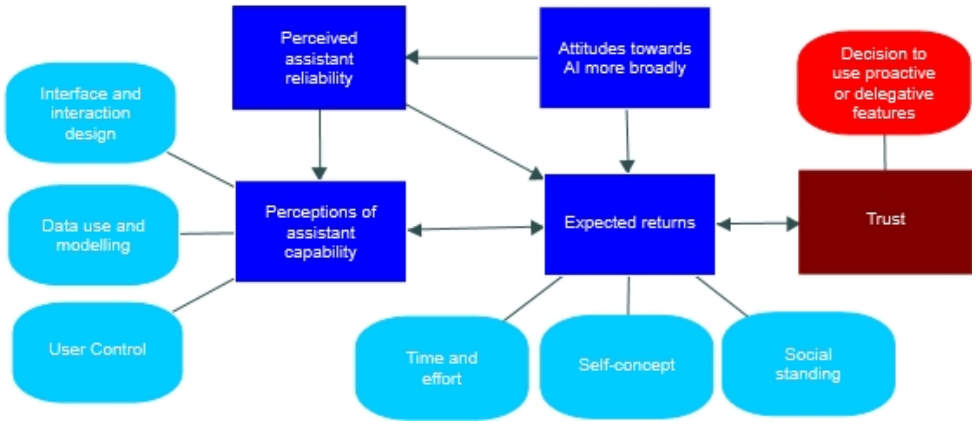


Figure 5.5: The preliminary themes, after the conclusion of step 3 of the thematic analysis

as system performance and explainability on their level of trust only indirectly, by focusing on expected returns. The preliminary themes included most of, but not all, the codes.

Steps four and five, that is, reviewing, changing, and naming the themes, were done by Jeromela under the periodic supervision of Conlan. This yielded a final set of three themes and a total of seven subthemes, comprising all codes from the transcripts, some of which were split and renamed, leading to a more internally consistent final codebook. The final analysis includes three main themes and seven subthemes, as shown in Figure 5.6. Our findings indicate that changes in feelings of trust towards the system are influenced by the user's perceived *prospects* of using or not using a proactive IPA or an AI system, rather than by their direct interactions with the system. We elaborate on this central theme of prospects, as well as on the themes of trust and interaction and their subthemes, in the subsections below.

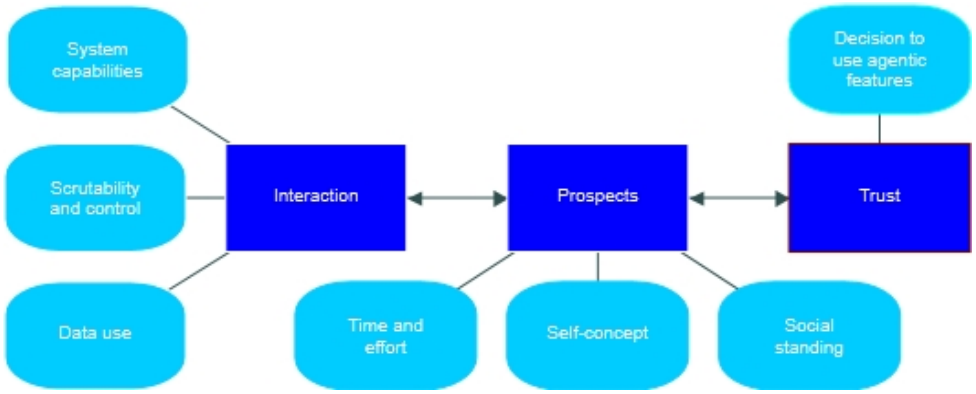


Figure 5.6: The final themes, after the conclusion of steps 4 and 5 of the thematic analysis

### 5.5.1 Prospects

The central theme is that of the expected returns, *prospects*. In discussing both the appeal and the concerns regarding AI-based assistance, the participants relied less on the features

themselves and more on the estimates of likely prospects, which can be divided into three broad categories: 1) time and effort; 2) self-concept; and 3) social standing, as discussed below. The word prospect was deliberately chosen, as our findings indicate that loss aversion is much more highly valued than the potential gains from using the system more freely. In that sense, thinking about prospects of returns stemming from system interaction is akin to thinking about monetary wealth, as made famous by a quote from behavioural economists Kahneman and Tversky [1979] that losses "loom larger than gains". We expand on this insight in Section 6.1, whereas in the subsections below, we report participants' points on each of the three identified prospects.

As decision-making is person and circumstance-specific, in Table 5.6 we summarise some of the statements regarding AI and eCalendar (e.g., Google Calendar or Microsoft Outlook) use and attitudes towards voice assistants that our participants shared. As expected, these sentiments align closely with the survey results (Figure 5.1).

Table 5.6: Expressed usage of LLMs, Voice Assistants, and eCalendars services

| Reported usage | Groups | Refs | Participants | |
|---|---|---|---|---|
| Uses LLMs | 5 | 5 | 5 | [G1P2], [G2P1], [G3P1], [G6P2], [G8P3] |
| Does not or rarely uses LLMs | 4 | 7 | 5 | [G2P1], [G2P2], [G3P1], [G6P1], [G7P3] |
| Uses eCalendars | 3 | 5 | 3 | [G6P3], [G8P3], [G9P1] |
| Does not use eCalendars | 3 | 8 | 4 | [G6P1], [G7P1], [G7P3], [G9P2] |
| Uses Voice Assistants | 7 | 36 | 11 | [G2P1], [G3P2], [G4P1], [G6P2], [G7P1], [G7P2], [G7P3], [G8P1], [G8P3], [G9P1], [G9P2] |
| Does not or rarely uses Voice Assistants | 9 | 29 | 13 | [G1P1], [G1P2], [G2P1], [G3P1], [G4P1], [G4P2], [G5P2], [G6P3], [G7P2], [G7P3], [G8P1], [G8P3], [G9P2] |

### Time and Effort

The first prospect relates to the anticipated effort, that is, whether using the IPA would indeed make calendaring easier, and, as [G2P3] put it, whether the assistant may be "useful for taking the drudgery out of scheduling things". Participants talked openly about the decisiveness of whether the assistant (and other AI systems) would ultimately increase or decrease the effort needed for them to achieve their tasks and goals. The ease of use and the time needed were also openly discussed. Below, we describe the main ideas identified within codes.

Participants generally anticipated that a calendaring assistant would be able to coordinate meeting timings, especially where several attendees are involved. Furthermore, [G5P2] pro-

posed that AI might suggest whether a meeting is necessary at all, thereby indicating the AI could serve as a voice of authority or objectivity, similar to how caregivers saw an ideal voice assistant as a "backup witness" in the study on IPAs for caregivers by Bartle et al. [2022]. Going beyond calendaring tasks themselves, participants considered how using the assistant would *affect their time* more broadly. For example, the automatic calculation of expected commute times between locations of consecutive events was often reported as useful by, for instance, participants [G4P1], [G5P1], and [G6P3]. The desired implementation was found to be subjective. For instance, [G6P3] noted that whether the commute time ought to be represented as an explicit calendar event or (as was the case with SIPA4TM) just as alerts, is person-dependent. Similarly, [G4P1] wanted more information control over commute notifications: "The thing that would make me feel more comfortable [is] if there was an API call to Google Maps instead of [just a commute time estimate of] 15 minutes or 35 minutes", whereas [G4P2] and [G8P3] liked the simplicity of a single time estimate. This divergence reinforces the finding of prior studies, including by Herder and Herden [2023], which indicate a need for a personalised approach to nudging in persuation. We also identified a similar sentiment for more advanced features that were discussed in focus groups, but had not been implemented in the current version of SIPA4TM, such as sending emails in the name of the main user: "By the time you've done all that, and then [the IPA] sent the email – just as quick as you [could]'ve done it yourself" [G5P2].

More generally, we found that the *ease of control* was a major contributor to the inner prospect calculus regarding the time and effort, with [G5P1] noting that they preferred putting events in a calendar with a longer duration than they actually are likely to take and that "if the system is telling me: 'Don't do this, don't do this, because you're not leaving enough transit time!', then I'll just do shorter blocks for the key events, per se. But, I don't know how long I can keep doing it. I could continue to learn from it and then get heavily chilled by doing that, or I could just give up and use another interface". Similarly, in all but one focus group, participants stated that the assistant would be more appealing if it had buttons to accept or refuse suggestions, thus reducing the effort and time needed for calendaring.

Similar mental calculations was reported for commercially-available voice assistant, with [G1P2] saying they "would find [voice assistants] to be a slower and *more effortful* user experience than just typing or using my phone" and with [G5P2] expanding this sentiment to overall frustration: "it became that, that level of frustration where it's now easier to take your phone out to look up for the information". The statement by [G5P2] is also very illustrative of the prospect-based mental calculus for deciding which technology to use, that is whether to use a LLM-based chatbot or to search manually: "why would I use ChatGPT instead of just googling it? Because I want more information, I want to think less about it, and I want to trust it as much as I can".

As we will discuss further in the theme of trust, the participants often assumed less time and effort would be needed to establish rapport with humans, with [G7P3] noting that people who have human assistants would be too busy to configure an AI-based assistant: "They won't have time to put in their own /.../ schedule in. You have to manage this and that". The participant then contrasted this with human assistants, who may be instructed while the primary task is being done, due to a higher degree of contextual understanding.

The assessment on whether using an AI system would lead to less time and effort needed overall seems to happen shortly after the start of the interaction and to be heavily influenced by the ease of control, as illustrated by [G2P1]: "I mean, in the first few minutes, like once I started doing the task, I found that the assistant wasn't actually really helpful so I just, like, inherently just avoided using it because I found that, okay, I know I can change it myself, so I'll just do that because it's easier because I'm trying to figure out the assistant". This indicates the role of the "halo" effect – as described by Kahneman [2011] in reference to Asch [1946] to mean that the initial impression significantly affects the final judgment – is present in human-AI interaction as well (as further reported in Subsection 6.1.1).

### Self-Concept

Self-concept is a term capturing the totality of an individual's thoughts, beliefs, and feelings towards themselves, as defined by Burns [1980]. The term was formalised in modern psychology in the theoretical system of Rogers et al. [1959]. The only motive in this psychological system is the actualising tendency, which is a tendency of an organism to develop capacities in ways that support it. According to Rogers et al. [1959], once the self-structure is developed within an organism, a *tendency towards self-actualisation* occurs, through which the discrepancy between the idea of the ideal self and the actual experiences in relation to the self is minimised. We adopt the terminology from the Roger's system to describe the subtheme of self-concept in relation to the AI-interaction. Our study indicates that users tend to evaluate a system by judging whether using it would aid or detract their self-concept actualisation.

**Supporting the envisioned best self.** On the positive side, the potential for the IPA to support (long-term) positive behavioural change was recognised, with [G1P1] expressing a desire to enter multi-year goals and have the assistant track and support their fulfilment. While no other participant expressed a desire for a such a longitudinal support, seven other participants from four further groups ([G4P1], [G4P2], [G5P1], [G6P1], [G7P1], [G7P3], and [G9P1]) expressed positive sentiment towards the role of the assistant in either, being more physically active (reminding them to take a break and stand up [G4P1], do a short exercise [G5P1], or a full training [G4P1,G7P3]), supporting habit forming (like a healthy sleep routine [G4P1] or doing chores [G7P1]), reminding them about deadlines or events [G9P1], or helping

make their calendar more aligned with their values [G6P1,G7P3,G9P1]. However, even in those situations, participants often expressed a desire to retain a level of autonomy over their time, only having the assistant be there as an objective supporter, rather then a benevolent dictator, as [G4P2]'s statement illustrates: "[The assistant's suggestion] gives me a bit more of security to be like, 'Okay, there is some research being done into me before making these decisions.' /.../ I could check [the suggestion] off or not because this is what I decided for myself. And then this is how I would start a habit to force myself into doing something and then checking on myself if I'm doing those things or not. So I like that it's asking me more questions than assuming more things". Or, as [G7P3] put it while answering a question on whether they would like there to be a button for the assistant to rearrange all events in the upcoming week more efficiently: "I'm unsure because that would be too much interference into my life. /.../ Like where I can just click and [have the IPA] manage my own time. Although it's very, you know, tough to enter everything on my own self, but if it gives me a solution, I should be able to click and move the stuff. But it should be my mind working towards my task, not somebody else changing everything on their own. So it would be too much for me".

**Essence of being human threatened.** We identified a concern about autonomy over one's time due to overreliance on an AI system. Several participants emphasised their dislike of scheduling advice regarding personal events. Among them was [G6P1], who described their experience interacting with SIPA4TM as "a funny balancing act because it feels like a computer is slightly telling you what to do and you're like, no free will", adding "I prefer to think about things myself. I just set a timer on my phone for each thing". Their preferred alternative was to memorise events or write them down in a notebook to help cope with feelings of overwhelm. Another participant who disliked the calendar in the study was [G9P2], who stated: "I hated seeing it all laid out like that. I would wake up on a Monday and look at that and think, I'm overwhelmed. /.../ social events, I wouldn't be scheduling them at all /.../ Honestly, I would be the sort of person who would be more likely to write down stuff on a piece of paper or have a traditional paper calendar with stuff written up on it that everyone in the house can look at". In response, [G9P1] suggested that it may be a "difference in personality" as she had the opposite experience: "I went from writing everything down, but then I'd get stressed having to revise because work can be fluid and depending on different events coming in. /.../ So then I end up going back and changing /.../. But I got quite stressed and messy having a, I would always have like a Filofax or notebook and it just got really ugly and I think stressful so then I switched to Google calendars and I actually felt less stressed being able to shift and move stuff down digitally and I'm the type of person I think if I see everything laid out I feel productive". Unsurprisingly, [G9P1]'s sentiment towards using the time management IPA was also more positive ("So I am a big user of Google Calendar normally, so I really enjoyed working

with that. And I liked as well, I think I would always love a personal assistant") than that of [G9P2] ("I would rarely use something like this").

**Values-driven AI usage.** Interestingly, even though [G9P1] suggested "I know maybe that's coming from someone of a different age group and less independent, like factors like kids and stuff that completely change over your timetable, but I liked /.../ it [when SIPA4TM was] coming back to me and going, yeah, you do have time for all of this today. I'm feeling better", thereby noting they were younger than [G9P2] and implying that the age might be the root cause of the difference in opinion. However, [G6P2] – a participant much younger than both of them – had a stance much closer to [G9P2] and offered a line-of-work explanation rather than the one based on age: "I don't think I'd be comfortable with it, to be honest. I live pretty manual, analogue, so to jump from that experience to having AI do that for me, I think it wouldn't sit with the way I operate. But I can imagine people with certain jobs might find it useful. But for me personally, in my lifestyle, it wouldn't". This is not surprising, as social phenomena are rarely one-dimensional, as illustrated in the complexity of the self-concept in psychology research, which we discuss in Section 6.1.

**Personalisation and Goal Modelling.** With this in mind, we argue that the decision on whether and to what extent to use AI-based time management assistants is made also with reference to the mental calculation on whether the assistance would help users actualise their self-concept. On the one hand, if the assistant is seen as helping them be the person they want to see themselves, some users, like [G2P1], [G2P2], and [G4P1], would even allow the IPA to restructure their week. On the other hand, if the assistant is deemed more likely to deny the user's self-concept, then the concerns about the loss of spontaneity become more pronounced, as is the case with [G3P1], [G3P2], and [G8P2]. Furthermore, [G4P2] and [G6P1] expressed concern about the environmental impact of AI-generated suggestions. Most participants were not as exclusive, however, finding their view somewhere in the middle, as [G7P2], summarised: "I also think that not a lot of people would be a huge fan of being told what to do, but rather having a suggestion of what could have been done to make things better". This mental calculation was summarised by [G2P2]: "[T]hat's kind of like using technology to free up all the time so you can be a person. Whereas the other way around is using AI to like do the people stuff, so that you can do work".

### Social Standing

The last major *prospect* we identified from our participants is social standing. This subtheme included considerations of how private and professional time management differ, and the view that it would ultimately be the user, not the system, to bear the social consequences of the

system's actions. More broadly, participants worried about how advanced AI technology would fit into or change social norms.

**Societal impact of e-calendars and AI.** Participants shared how e-calendars, widely used today and without AI integration, also affect one's professional relations. [G5P2] complained about letting other people enter events in their calendar "I don't have any time management and I'm just jumping to the tune of somebody else, right? So I found that the way around that is to actually block-book, a bit like this, we have focus booked". [G5P1] agreed with [G5P2] and noted that the shared calendar may also become a proxy for productivity: "So, for the calendar, right, everyone sees you in office or not out of office. So one way to handle that is like trying to fool it, like, you know, like putting random things in it. Because, yeah, if you can't just help to put less information on it, just try to fool it. I would do that. [laughter]". Conversely, [G3P2], [G4P1], and [G7P3] expressed a sense that using AI systems may similarly affect one's social standing, as they become more socially accepted within our societies.

**Separation of private and business calendars and the role of personal responsibility.** We identified a broad desire to keep personal time management and professional time management assistance separate, as expressed by a participant who compared it to having a shared work calendar and a private, less structured calendar. Such a desire was echoed by participants including [G6P1], [G6P2], [G8P1], [G9P1], as well as [G8P3], who stated: "from my perspective, I would tell that currently, suppose, as [G8P1] told, there is a work calendar and there is also a personal calendar, and I can view them like side-by-side. I do not mix them together". This strict separation protects the social standing not only from the negative effects of miscommunicating the user's intent to others, but also protects the user from others' control: "I would have one issue with using this [calendaring Assistant] in my Outlook is that there's a lot of private life shared, as I have a shared calendar for work. So the one that I use in work is the Outlook one and is also shared for people who book appointments for me. And I know that there are ways that you can kind of hide certain components, but I wouldn't want to share that amount of detail or personal information in the work" [G6P2]. So, by limiting the IPA's capabilities, users aimed to prevent the assistant from miscommunicating personal data and from implying their consent to put events in their calendars.

### 5.5.2 Human-Assistant Interaction

During human–assistant interaction, users' perceptions of the system are shaped most directly, with prior work finding that impressions, trust judgments, and mental models form dynamically during interaction [Kulesza et al., 2012, 2015]. The chosen interaction modality and the user interface design were the most direct elements discussed by the participants. We identified

subthemes related to the system capability, the control and scrutiny mechanisms provided, and data use. Table 5.7 below overviews whether participants preferred voice or text as the primary mode of interaction with their ideal time management assistant. The column Refs (short for references) indicates how many statements were coded to express that preference.

Table 5.7: Preferences for interaction modality across focus groups and participants

| Modality Preference | Groups | Refs | Participants | |
|---|---|---|---|---|
| Prefers text over voice | 3 | 6 | 5 | [G2P2], [G4P2], [G6P1], [G6P2], [G6P3] |
| Prefers voice over text | 1 | 1 | 1 | [G2P3] |
| Depends on context | 6 | 10 | 7 | [G1P1], [G1P2], [G2P1], [G4P1], [G5P2], [G6P3], [G7P3] |

While SIPA4TM had a straightforward buttons-and-fields design, we asked our participants what the ideal user interface for their IPA in time management would be. Only one participant, [G2P3], said they would prefer voice over text in all situations. Notably, during the second phase of the study (the SIPA4TM interaction phase), [G2P3] indicated they were dyslexic, which situates the preferences and underscores the need for inclusive and accessible design and aligns with prior findings, such as those by Klein et al. [2024]. [G4P2], [G8P1], and [G8P3] voiced concerns about privacy and the use of voice, reflecting concerns previously discussed by Tabassum et al. [2019] on the assistants always listening, including during sensitive conversations, [G8P3] stated: "if it's in a setting in which it will detect whenever I ask it something, which means it has to listen every time, constantly /.../ so there's not much privacy [left]." In contrast, [G8P1] was concerned about being overheared or looked at while giving voice commands: "[I]f I'm in a situation where I don't want to be shouting at my phone I'm more [concerned about] privacy, [such as] if you're in an open space or with people around or if you're working. [Then,] you don't want to be shouting in an office on the phone. The only time I'd say I'd use that is with the connection at home for changing a song on Spotify, maybe". On the other side, [G4P2] was concerned about other people hearing the IPA's responses "[I]t just depends like if I'm outside, if I'm at work, I won't want, like, everyone else around me hearing what my assistant is saying". The latter two concerns were previously reported by Cowan et al. [2017] as reasons people abandon using voice assistants. Further reasons Cowan et al. [2017] identified included problems with speech recognition, and [G4P2] also complained about this, stating, "for me text is more reliable because a lot of times with voices it doesn't understand what you're saying". Regarding time management specifically, [G6P2] stated they prefer text over speech, as they "see a calendar as a visual rather than an audio". The last participant who indicated they prefer text-only interaction was [G2P2], who reported that it was down to personal preference: "I've never really used voice much. Even when it come to calling my friends, I prefer to text".

Participants [G1P1], [G1P2], [G2P1], [G4P1], [G5P2], and [G6P2] stated they would ideally want the assistant to switch between voice-based and text-based interactions. Again, in line with the findings by Cowan et al. [2017], we found that using voice was a popular choice when the user is driving (reported by [G1P1] and [G1P2]) and in other situations where one is engaged in situations that require them to have their hands free ([G5P2] and [G7P3]). Overall, some participants, like [G4P1] and [G5P2], were very goal-oriented in their modality preference, with the former stating: "I think in terms of how we want to interact with it, honestly, I just want to get the job over and done with. So sometimes voice is better, where I want to be like, yeah, that works, go. Or, like, sometimes the text with the graphic is a faster way than pulling. So it really depends".

We also found a resounding rejection of using avatars in time management interactions, with stated reasons including that avatar-enabled interactions would feel patronising [G5P1], overwhelming [G7P3], frightening [G4P2], or unnerving [G3P1]. [G2P1] extended their negative sentiment towards humanisation generally: "I think I'd be happy speaking to it. I wouldn't want it to speak back to me, I guess, especially if there's an avatar. So you would humanise it in a way I'd not want to". Similarly, [G6P3] found the idea of avatars "weird" and said they found it inappropriate for an AI to simulate emotions while giving time management advice, adding "I'd rather just read, I think. I think, yeah, just text. I wouldn't want any way of, like, telling me, like, yeah, like, frowning at me or whatever. [laughter]", with which [G6P1] and [G6P2] agreed. In contrast, [G4P1] was the only participant who saw a positive use case for avatars, primarily for tone-setting, and stated: "I think the avatar or something surrounding what the face looks like is kind of interesting and can kind of go back to the tone about how you want to manage your day. You know, is there a caring parental, like there's some psychology here, caring parental self-management tone of your day as a student or as a worker. Maybe you can embody that in a graphic. The mind can associate something with that, too. Like I don't want the AI to have a personality on it, but it's going to have an image, that image reflects maybe what my ambitions for this tool might necessarily be or how I want it to treat me".

The opinions regarding the interface design of SIPA4TM specifically, [G8P1] evaluated it as "user-friendly", as it is "quite similar to other calendars", which indeed was one of our explicit implementation goals. Still [G2P1], [G4P2], [G5P1], [G6P2], [G7P1], and [G9P2], stated that they did not like the user interface design, disliking the extra effort needed relative to commercial calendaring tools. Specific interaction elements noted as missing were drag-and-drop control, right-click context menu, and colour-coding for event types. While the opinions on the usefulness of suggestions varied, as we discuss further below, it is interesting to note that [G5P1] disliked the perceived tone of the assistant: "Yeah, because currently it tells you like, oh, yoga time is too short. It feels like my mum". More broadly, [G2P3] and [G4P1]

stated that they would prefer the suggestions to be visually overlaid in the calendar rather than displayed in a separate container, whereas [G7P1] felt that the suggestions should be more concise. Notably, the risk of persuasive advice being disregarded or received negatively was also highlighted by Herder and Herden [2023] in their user study on advice given by voice assistants.

### Capability

Participants in our expert groups differed markedly in their assessment of both the current and prospective capabilities of Intelligent Personal Assistants (IPAs) for time management. While many participants, including [G3P2], [G7P3], and [G9P1], expressed scepticism that IPAs could ever fully replace human personal assistants, several participants emphasised that IPAs could nonetheless substantially augment human assistants' work, particularly for routine, data-intensive, and coordination-heavy tasks. Across groups, capability was not discussed as a monolithic property, but rather as a composite of technical competence, social understanding, reliability, and perceived accountability.

**Time management as a fundamentally complex domain.** Ten participants, including [G1P2], [G5P1], and [G5P2], repeatedly characterised time management as intrinsically difficult to automate, citing challenges related to prioritisation, task switching, and frequent context changes. A central concern was, as [G4P1] noted, that calendars often function as "an inaccurate map of your time", with events distributed across multiple calendars, some scheduled off-calendar entirely, and others shaped by implicit social signalling rather than genuine availability.

This concern was echoed by [G6P2], who highlighted that an overpopulated calendar may serve for performative or signalling purposes rather than accurately reflecting work, noting a concern about the perception that "if my calendar is not overpopulated, it looks like I'm not working, which is not the case". [G6P2] also noted a discomfort with "regimental" calendars, which explicitly include all activities, criticising the reasoning behind SIPA4TM's suggestion regarding the commute time: "I know that I have 15 minutes between when I finish lunch and when I have to be at the doctor's. I finish lunch a little bit early and I'll excuse myself but I don't want to have to go through all the trouble of moving lunch closer and making sure that I have enough time to go to the doctor's".

This incompleteness in most people's calendars was seen as especially problematic for systems attempting to make proactive or autonomous scheduling decisions, since much of the relevant information – intentions, flexibility, social obligations, or unarticulated preferences – remains tacit. As [G5P1] explained, an IPA cannot reliably infer priorities or intentions in cases such as

social obligations, illustrating how assessing capabilities occurs during interaction: "just take the airport pick-up a friend as an example. If I put in 12:30 on it, does it make the pickup time? Or does it mean that I already planned 30 minutes ahead, say from 12:00? Then [the IPA is incorrectly] telling me that there's not enough commute time. /.../ while they're learning from us, we're also learning from them /.../, we could learn about like what [the IPAs] will alert us about and then try to avoid that".

Relatedly, participants including [G4P1], [G5P2], and [G9P2] highlighted the difficulty of modelling how different work contexts (naming as examples: working from home, travelling, visiting clients, working abroad, and business of the day) alter acceptable scheduling decisions. As [G4P1] noted, working remotely from another city or country constitutes is "fundamentally different to the other weeks" in terms of expectations, priorities, and acceptable trade-offs, raising doubts about any IPA's capability to infer that from calendar data alone.

**Humans versus AI in modelling intentions and social context.** When comparing humans and AI systems directly, 11 participants, including [G5P1], [G5P2] [G7P1], [G7P3], and [G8P1], consistently attributed superior capability to humans in domains requiring social understanding, contextual judgement, and sensitivity to interpersonal dynamics. Human assistants were described as better able to listen [G3P2,G8P3], interpret social cues [G4P1,G8P2], and understand relational and business priorities, such as knowing which meetings can be declined, postponed, or reframed based on contextual elements such as a manager's values or a client's flexibility [G4P2,G8P3].

Several participants, including [G3P2], [G5P1], and [G8P2], emphasised that these judgements are often grounded in long-term relationships, shared experience, and emotional attunement, which they did not expect AI systems to replicate. On that note, [G8P1] and [G8P3] believed humans are better able to choose the modality and adapt the tone of conveying a message: "if you're changing [a meeting] time, I think the personal approach is definitely more beneficial because you can gauge how that person's going to react [and] you can then approach it differently as well, whether it's over an email or a phone call or how you kind of manage that". Crucially, as [G4P2] remarked, when an AI system is wrong, "I'm the one who will suffer", whereas the system itself is incapable of "feeling" any consequence.

At the same time, most participants were careful not to frame this comparison as a simple opposition. For instance, [G7P1] and [G7P2] discussed a model of complementary capability, in which IPAs enhance human assistants rather than replace them. For them, IPAs were seen as particularly well-suited for scheduling optimisation, cross-calendar coordination, or availability matching, thereby increasing efficiency and reducing cognitive load for human assistants. In organisational contexts, this augmentation was explicitly framed as economically beneficial,

enabling one human assistant to support more complex workloads through the use of AI tools. Lastly, for tasks such as analysing and comparing flight options across booking platforms, [G7P2] believed IPAs would be able to be more effective than human assistants alone.

**Uneven capability across AI systems.** Participants distinguished between different classes of AI systems, sharply contrasting traditional voice assistants (such as Apple Siri or Google Assistant) with more recent LLM-based systems. Voice assistants were widely described as unreliable, limited, and frustrating, with frequent failures in recognition, execution, and contextual understanding, as reported by [G4P1], [G8P1], and [G9P2]. In that sense, it is notable how these descriptions echo the complaints reported by Cowan et al. [2017] almost a decade ago.

By contrast, LLM-based systems were perceived as substantially more capable in conversational interaction and reasoning, leading some participants, including [G7P2] and [G8P2], to describe recent systems as approaching "human-like" interaction in limited contexts. However, this increased apparent capability did not translate into unqualified trust, particularly when systems were imagined as acting autonomously or on the user's behalf (cf. the theme on trust). Interestingly, [G7P2] reported a recent perceived improvement in the voice assistant they were using, suggesting that the sharp perceived gap between LLMs and voice assistant capabilities might be narrowing (as the former is more integrated for the latter [Saeidnia, 2023]).

**Bugs, hallucinations, and transparency of errors.** Concerns about bugs and hallucinations featured prominently in participants' assessments of AI capability. Ten participants, including [G4P1], [G5P1], and [G8P2], expressed anxiety not only about the existence of errors, but about their opacity and potential consequences as well. Hallucinations were seen as particularly problematic because incorrect outputs are not always distinguishable from correct ones, placing the burden of verification on the user, as expressed by [G2P2], [G4P1], [G6P3], and [G7P1]. Additionally, [G5P2] was doubtful if the LLM-based IPAs could ever be reliable, as "it's just being trained on the internet [but] everything is rubbish on the internet".

### Data Use

Across the focus group discussions, the core concern about data use among participants does not appear to stem from ideological disagreement on the value of one's privacy. Instead, participants often spoke from a pragmatic stance, regarding extensive data extraction as an established fact of contemporary digital life ("the entire world is sharing your data, their data with all the big guys, /.../ it's not in your hands, whether you like it or not" [G3P2]). The principal concern was thus not whether private data is collected, but whether it is processed for purposes in alignment with the user's values or, at least, in a manner that would not cause

them harm. In other words, there was concern about misaligned interests, whereby service providers would prioritise monetary interests or secondary data use over users' well-being. Specific risks named included misusing calendar and behavioural data for purposes such as targeted advertising or government surveillance. It was this concern that underpinned the behavioural and psychological strategies the users described. These strategies were identified to belong to four broad groups:

1. pushing for client-side data storage and processing or resisting against digital profile with guest-mode use or use of multiple accounts;
2. embracing the "no harm" stance, whereby their online calendaring or LLM-use data is assumed to be low-risk;
3. active self-censoring of inputs to constrain data misuse risks; and
4. resignation to a "what will be, will be" stance, sometimes coupled with the hope that external forces such as regulation or public perception considerations would contain the extent of any data misuse.

Below, we elaborate on the identified interest misalignment concern as well as the four coping strategies. Table 5.8 provides an overview of participants whose statements were coded at least once in the corresponding points of the theme.

Table 5.8: Numbers of groups, codes, and participants and lists of participants coded in each point of the Data Use theme

| Stance | Groups | Refs | Participants | |
|---|---|---|---|---|
| Interest misalignment | 6 | 45 | 12 | [G2P2], [G2P3], [G4P1], [G4P2], [G5P1], [G5P2], [G6P2], [G6P3], [G7P2], [G7P3], [G9P1], [G9P2] |
| Resignation | 7 | 53 | 12 | [G1P1], [G2P2], [G2P3], [G3P1], [G3P2], [G5P1], [G5P2], [G6P2], [G7P1], [G7P2], [G7P3], [G9P1] |
| The "no-harm" thesis | 7 | 13 | 8 | [G1P1], [G1P2], [G3P1], [G4P2], [G5P1], [G6P3], [G7P1], [G8P2] |
| Self-censorship | 6 | 15 | 8 | [G1P2], [G4P2], [G5P1], [G6P2], [G7P2], [G7P3], [G8P2], [G8P3] |
| Client-side and guest-mode | 6 | 16 | 8 | [G1P1], [G3P2], [G4P1], [G5P1], [G5P2], [G8P2], [G8P3], [G9P1] |

**Interest Misalignment.** The first subtheme centres on a perceived misalignment between user goals (e.g., receiving useful assistance) and corporate or institutional incentives (e.g., monetisation, profiling, behavioural targeting). Participants framed the risk less in terms of abstract privacy violation, and more in terms of uncertainty regarding downstream uses: "I prefer not to give a lot of personal information because that's something very scary. Because

we don't know who is handling the, obviously, like different companies have their different rules" [G7P3]. This concern was often illustrated by anecdotes of responses that participants reported as intrusive, adversarial, or weird. For instance, [G9P1] described actively opting out of "more personalised ads" while assuming the underlying tracking persists: "it's the company selling the data and that profits they use it to target you. I think there has been times where I got a bit unnerved. I feel like I just talked about chewing gum and then I see 20 minutes later an ad on my Instagram for chewing gum".

There was also a sense that there is no significant difference between service providers regarding risk of data use, because, as [G7P1] put it: "I don't mind the information because, yeah, I mean, it's like if Google has it, then ChatGPT has it, I don't, I mean, we lost it, it's too late". A similar stance regarding that "there's not really much difference against these companies", was voiced by [G2P2] who stated that they believed the increased concern about IPAs and LLMs in particular is because of the conversational fluency, "as soon as they can talk to you like a human, then it gets scary that someone knows all your data".

Monetary gain was the principal interest ascribed to the service provider, with [G5P1] stating a belief corporations would earn more by selling the data than through their primary services: "It's private because you pay this price. But if others offer a bigger price than what you [had], you know, the big companies are..., it's more about money.". Conversely, the interest of the user was identified as improved system performance [G4P1], content recommendation [G1P1], and personalised suggestions [G7P3]. Beyond the corporate service providers, some participants were also worried about interest misalignment between them and the national governments. In particular, [G7P3] was worried about service-based government surveillance: "I haven't started using the China's DeepSeek because we were told that they collect the user data", while [G6P3] talked about his concerns regarding government mass-surveillance, citing an online safety bill in the United Kingdom and the recently published collaboration between the United States government and a private company on data aggregation.

**Resignation.** Now we describe the four identified stances, or strategies, participants described on the use of their data by online services, including chatbots and IPAs. The first of the four identified strategies for controlling one's data was one of almost complete resignation and a sense of futility to regain control, as [G3P2] unequivocally asserted: "irrespective [of whether] you like it or not, everyone's out there taking data from you. Whether it's Instagram, whether it's WhatsApp, or Facebook, you really can't – you know – get away that anyone's now taking your personal data. You are not liking it, but irrespective your data is being shared with everyone". This sentiment was sometimes voiced as a flattening of distinctions between companies: "if Google has it, then ChatGPT has it" [G7P1]. Big technology companies were seen as dominant in the control of one's data, with [G7P2] stating that whenever they see

a CEO of OpenAI in the news, they "feel it's over", leading them to "not have too many hopes" for privacy. Nevertheless, [G3P2] stated they still share personal data, as "if I want to make it better, I don't have much of a choice", expressing a common conclusion. In this view, privacy is treated as a historical artefact rather than a maintainable right, so sharing with one additional entity (be it an IPA or another internet service) is negligible relative to the current ecosystem.

While the sense of resignation was most notable while chatbots and other internet services were discussed, [G5P1], [G5P2], and [G6P2] expressed a view that even the contemporary workplace necessitates eroding privacy and detailed types of information that can be inferred from mandatory workplace calendars. Lastly, a few participants openly questioned the ability of national governments to regulate data use in the interest of the citizens, as [G5P1] put it: "the regulations could be changed because the big firms and the money as well" and noted that the regulation is "more of a protection for the current time, but /.../ your data will leave traces and in a future time, probably the regulation changes and then you can come back to the data".

**The No-Harm Thesis.** Indicatively, [G5P1] concluded his stance on likely futility of fighting data misuse through government regulation by stating "I'm not a very important person, so I wouldn't worry". Thus, some of the participants who expressed resignation regarding data use control, also expressed their belief in the "no-harm" thesis, whereby the detriment of the (potentially misused) data about them cannot harm them significantly. Importantly, the participants who expressed such a view were not necessarily also proponents of the resignation stance, as indicated in Table 5.8.

The no-harm thesis typically reduced the concern regarding online privacy by assessing calendaring and chatbot interactions as containing non-sensitive data, i.e. data too mundane or insignificant to be exploited. Notably, the potential for misuse was not ruled out entirely, as evident by the qualification in the statement by [G3P1]: "I feel like there's nothing really much about my schedule that I would be too worried about people knowing. I'm sure it could get into the wrong hands, but still it's... I kind of feel a bit... I kind of trust a bit more than I feel like I should". Such statements suggest the strategy serves as a psychological buffer against the discomfort of data sharing.

On the other hand, the no-harm thesis did not necessarily mean the lack of any precautionary behavioural changes. For instance, [G1P2] framed the risk of sharing personal data with chatbots as "not the end of the world", but then qualified this statement by conditioning its truthfulness on appropriate sharing behaviour ("where it's [not shared as] a public URL and that's very dangerous"), non-threatening content ("basically as long as you're aware that you're

not putting something like really stupid in the comment, you know, like this is when I go and buy illegal weapons"), and not an "abnormal person, where it's dangerous, that your location data is leaked for whatever reason". Some participants also suggested that low-risk calendaring assistance can be delivered with minimal information. For instance, [G6P3] argued that basic scheduling may require only knowing "when you're working" and where there is "free time on the calendar", without the system needing to "get more" than that. The willingness to share data is then conditional on the immediate utility of the service: "in terms of how much data they are collecting, obviously till the time it is useful to me and not harmful" [G7P3].

**Self-Censorship.** As we have seen in the example of [G1P2], even some of the participants who were more likely to argue for the no-harm thesis of data sharing expressed an understanding of limiting the content and types of data shared. The strategy of self-censoring content was also employed by [G7P3], who said they "try not to give personal details and information". Various forms of self-censorship were described by the participants, whereby they would limit what they believe they share so as to maintain a vestige of data control. Notably, several participants shared they started censoring their content more or even withdrawing themselves from internet services such as social media, after a negative experience of data visibility or federalisation of their data. For instance, consider [G4P1]: "I recently saw those face search things and uploaded my photograph and found five photographs of myself on websites that I did not know were there. So, my opinion about putting stuff that's seemingly open-facing on the web has gone down, and I'm quite uninterested in it. /.../ [A]ll the data that Facebook has on me of some sort of meta or, like, psychological performance or what psychometric data they have, kind of concerned about that stuff, and I am kind of pulling away from it". For some participants, this strategy results in more deliberate data sharing and self-vigilance. For instance, [G4P2] said they "I'm always sort of watching myself. I know what I'm sharing and then like, if it can be used or not" and continued by saying they abandoned social media and remain "hyper-aware /.../ so I wouldn't be that surprised by what people know about me". Similarly, [G6P1] reported being "very reluctant about putting too much" of their daily activities in a digital calendar. In these accounts, the user does not expect the system to enforce appropriate data boundaries; instead, the burden of privacy protection is internalised as ongoing vigilance and minimisation.

**Client-Side and Guest-Mode.** Finally, the most assertive strategy was to change the data-handling arrangement itself by building local data storage and processing; or (as a less technically savvy alternative) interaction with online services without logging in (i.e., through the guest mode) or by using multiple accounts.

The effort to do it was fueled by the desire to reduce exposure to cloud training or third-party

repurposing. [G8P2] expressed this in relation to a question about the permission control of SIPA4TM, saying that they saw stages as "really good since some users could just use this calendar for general use and they would not require the highest amount of preferences". But [G8P2] then continued by saying that "the only solution" that would "relieve" their privacy concerns is if "all of my usage data locally on my computer only", to avoid "sharing my data to train the chatbot on the cloud". [G8P3] agreed with the stance by [G8P2], noting that they "understand" that the data "needs to be on the cloud to probably analyse it or work on it" but then stated they would expect that "encryption could be possible, like from my end and from the server end to keep it like in a closed loop only within it". While [G1P1] and [G5P2] elaborated on their attempts to build locally-hosted AI-based systems, they described wanting to have a system they can control so it fits their needs fully, rather than emphasising data-use concerns.

[G3P2], [G5P1], [G5P2], and [G9P1] described privacy-preserving strategies rely on interactional changes of their online behaviour, by not logging in, using multiple accounts, opting out of personalisation, or using the incognito mode in browsers. Specifically, [G9P1] stated that, while they do use social media, "'[w]hen it comes up to say, do you want more personalised ads, I usually pick the option that says, no, just keep it general", clarifying that their hope is to thus minimise data misuse. Similarly, [G3P2] reported going to ChatGPT "in an incognito mode", noting that "it could be more perceptional", but concluding with "I just think, psychologically it's more safe". A similar hopeful but uncertain sentiment for changing their online behaviour was expressed by [G5P1], who described a slightly different approach: "have multiple accounts, so I tend to not keep everything on one account. Hopefully you're going to confuse someone for it, but I don't know". So, in contrast to resignation or self-censorship, the participants in this case have not only adapted their behaviour to the data regime, but also actively sought to create interaction contexts in which data collection is either structurally limited (client-side) or perceived as more obscured (guest-mode).

### Scrutability and Control

In situating our findings regarding the theme of Scrutability and Control, it is necessary to first distinguish between the related but distinct concepts of scrutability, explainability, and user control. As originally conceptualised by Kay [1999], scrutability refers to the user's capacity to investigate and study the personalised user model that underpins personalisation. While explainability often focuses on the system's ability to provide a rationale for its outputs, scrutability emphasises the active role of the user. Following the reasoning by Khosravi et al. [2022], we underline two defining characteristics of scrutability that set it apart from explainability. First, it necessitates a active investment of effort by the user to seek out answers. Second, scrutability acknowledges that in the face of modern, high-complexity systems, the

user may only achieve a partial understanding of specific facets rather than the total transparency implied by "explainability". Closely linked to this is the concept of user control Kay and Kummerfeld [2013], which we define as the active process by which a user, having sufficiently understood an application, modifies the underlying model to better align the IPA with their personal intentions.

Building on this notion, we developed *stages* as a mechanism of coupling data permissions and IPA features into three distinct levels. Conceptually, a stage can be understood as an extension of scrutability: it provides an explicit frame for how the assistant may use data to generate calendaring suggestions, thus aiding the user in shaping their understanding of the IPA and its modelling. We detailed the exact implementation of the permission control in Section 4.2 on page 55.

The remainder of this theme captures how participants discussed scrutability and control in practice, spanning five interrelated points:

1. How users verify AI outputs and assumptions;
2. How the assistant should prompt and scaffold user verification;
3. What explanations were considered inappropriate or counterproductive;
4. What participants considered explanations to be appropriate and when they were worth attending to; and
5. The participants' views on the permission control in stages.

In essence, our findings portray scrutability as a fundamentally *consequential* and *selective* activity: participants did not describe a standing desire to understand the system for its own sake, but rather a risk-driven investment of time and effort, that is more attractive when the assistant appears to make incorrect assumptions or when delegation threatens diminishing the user's prospects.

**User Verification** Participants often framed verification as a *user responsibility*, rather than something they expected the IPA to manage autonomously, encapsulated by [G7P1] as "you have to check the reliability still. Maybe at some point it will become so reliable and the errors will be so minimum that people will just /.../ say, okay, I'm fine with this one error out of one million. So far, it's not the case". In the context of SIPA4TM, this was partly grounded in the perceived transparency of calendaring outcomes, as [G2P2] put it, "for like a virtual assistant for a calendar, you can see if it's wrong. You can just add stuff to the calendar, you can see if it's a clash, you can see if you have enough time". A similar sentiment was expressed, among others, by [G1P1] and [G5P2], who would "sanity-check" the advice by

inspecting the calendar state after suggestions, mentally simulating consequences (e.g., timing feasibility), and comparing assistant outputs with their own tacit understanding of priorities and constraints. In such cases, users were not seeking comprehensive transparency; they were checking whether the suggestion was locally plausible and aligned with their intent.

Beyond calendaring, participants described additional *verification practices*. As is further argued in Section 6.1, we identified improving prospects as the principal motivation for these practices. More narrowly, the motivation is alleviating personal suffering due to IPA's mistakes, as [G4P2] explains: "it can be wrong, and if it's wrong, I'm the one who will suffer. And then there is no suffering for ChatGPT", underlining this is why they "take what ChatGPT tells me with a grain of salt", concluding that that is why they "still having to proofread it". Similarly, [G1P1] explained they would "mostly look into what's happening under the hood to figure out if it was going wrong". A notable variation of user verification is *cross-AI benchmarking*, where participants compare multiple AI systems to assess reliability. This might sometimes supplement explicit provenance-based verification. For example, [G2P3] explained that, for ChatGPT, they would "ask the same question of a number of different AI sites and to compare them" and sometimes also "look at the list of links to see where it was getting this information from". These accounts may be seen as attempts of improving one's social standing prospects at the expense of effort and time.

Verification was also discussed as a means of maintaining personal agency: even when participants were willing to accept assistance, they resisted becoming passive recipients of a system-driven schedule. This preference aligns with participants' broader emphasis on retaining the ability to judge and decide for themselves. As [G3P2] noted, they "definitely would have a look at [IPA's suggestions] once just to make sure [the IPA] is going to give me the right suggestions or not – right or wrong, I can decide for myself". Notably, the tasks in our study might have affected such behaviour by asking the participants to role-play someone else. Thus, it is likely that verification to assert personal agency and ultimate judgement would be more pronounced in a real-world interaction, as the statement by [G5P2] indicates: "I felt like I was changing things just because [the IPA] was telling me to, rather than because I wanted to. Or because it really made much sense. If it was me, I wouldn't have ever gotten up at half six in the morning to do any home workout". Thereby, verification is not only an accuracy practice but an autonomy practice: it allows users to benefit from assistance while remaining final arbiters of consequential changes, or, in terms of prospects, it reifies the actualisation of the self-concept in exchange for time and effort.

At the same time, a set of barriers constrained the depth of scrutiny participants were willing to undertake. First, several accounts indicated a *feeling of incompetence* with respect to diagnosing and correcting errors in the system's internal model. [G4P1] articulated this concern

by highlighting the fear of unintended side effects: "I don't know if I have confidence enough in my ability to [change the user model], depending on what the model is. If it's just wake up and going to sleep, totally fine. But if it's like, there's more characteristics to it than that, . . . I don't want to change the ingredients. I don't know how that's going to impact how this thing operates". Similarly, [G3P2] stated they "wouldn't give it that much time" because "people are much better than me when it comes to that", implying that even when curiosity exists, scrutiny per se does not warrant lowering one's time and effort prospects.

At the design level, participants discussed what made *verification more difficult*. For example, [G7P3] described confusion after not seeing the effects of a user model change highlighted in the calendar, reporting they "couldn't differ between the actual user model and there was just that timing change, wake up time change or something like that". A similar sentiment was expressed by [G8P1] who noted that some of the effects of changing stages "wasn't very obvious". Relatedly, in more advanced AI systems, such as ChatGPT, the way the user can verify the modelled information about the user might not be obvious. This was indicated by [G6P1], [G6P3], and [G8P3] who noted they haven't looked into the system's "memories" feature (which stores statements about the user in natural language), as they had not known this was possible. Ultimately, this indicates that user verification is not always driven by direct prospect calculation, but is also directly influenced by interaction design.

**IPA-Prompted Verification** Complementing the user-driven verification practices described above, participants also articulated expectations for *IPA-prompted* verification, particularly at moments of elevated risk or uncertainty. This sentiment is summarised by [G4P2]: "I like that it's asking me more questions than assuming more things. "This requirement for explicit approval was most acute for high-stakes proactivity, such as schedule reorganisation [G7P1] or financial transactions [G2P3]. Participants were also wary of a system that might act autonomously without a clear checkpoint, with [G1P2] stating: "I'd always want like final say... I wouldn't want it to basically reorganise my entire day without checking with me first". Similarly, [G2P2] expressed the desire to avoid scenarios where one might wake up the next day and my calendar's all changed on its own". Rather than granting broad authority once (e.g., "you can change my calendar"), participants, typically preferred the assistant to ask about *particular interventions* in context, such as "You have a problem here. Do you want me to reschedule? And then I say yes or no." [G7P1]. Similarly, [G7P3] explicitly framed IPA-prompted verification as "getting permission" for a change, thereby preventing the system from becoming an "independent association" that acts on its own.

Several participants reinforced the necessity of a Human-in-the-Loop (HitL) architecture, with varying degrees of desired IPA autonomy. On the more restrictive side, [G9P1] emphasised they wanted to remain the final arbiter: "I still want to be the person in control and saying,

'I approve', 'I approve' ". A similar restrictive stance was also shared by [G3P1] and [G3P2]. On the other hand, [G4P1] described the desired role of the user as the principal coordinator, giving an example of scheduling assistance: "I think there could still be a human in the loop though, if there was a coordination. I have to, like, they're saying in the room next door, I have to have an email thread for like 60 people. Like 60 emails because you're trying to coordinate a meeting between 10 people". Towards the more hands-off side of the human-in-the-loop autonomy spectrum, there was [G2P2], who stated "I'd find suggestions and I probably would be fine every time. But still, I wouldn't like to wake up the next day and my calendar's all changed on its own" and thus placed AI-prompted verification as a hedging mechanism against high-scale undesired consequences. In other words, the wariness towards IPA proactivity without a validation checkpoint can be understood as a fear that the user's self-concept and social standing prospects would suddenly deteriorate. [G4P2] imagined less human verification as the behaviour of the system becomes more familiar.

**Inappropriate Explanations** While participants expressed a clear desire for verification mechanisms, they were also firm on rejecting what they perceived as inappropriate explanations. In short, transparency and scrutability were not desired per se; instead, the participants wanted them to serve broader interaction goals and, at the highest level, their prospects. This sentiment was captured by [G4P1], who remarked that "people don't care about how the spaghetti is made, they just want to eat the spaghetti". We identified three hallmarks of inadequate explanations: explanations that were *too obvious*, *non-specific*, or *unreliable*.

First, several participants described a *"common-sense" threshold* below which explanations were experienced as redundant, and in some cases actively irritating. For instance, [G2P1] noted that when the assistant merely restated a visible clash (e.g., after the user had already noticed it), the explanation did not add value: "I found that as well actually, so when there would be a task and it would tell us that there's a clash, I can see myself that there's a clash, so I can see that first and then if I click on the date, the system tells me that there's a clash, but I already knew there was a clash, it's not exactly helpful". [G5P1] summarised this mismatch bluntly as "Cause I don't need your advice. I see". This sense of redundancy extended beyond "clash" messages, as participants also resisted being required to justify or annotate actions they considered self-evident. For instance, [G6P1], found it inappropriate that saving an event required a description, framing this as an unnecessary demand to "explain how you're doing something". Notably, the description was indeed optional, and clicking the save button one more time would have save the event, but [G6P1] was among the participants who have not read the description of the pop-up suggestion, as we have previously reported in Subsection 5.3.3.

Second, participants described explanations as inappropriate when they were *non-specific*, that

is, when the assistant acknowledged that something was wrong but did not translate this into actionable guidance. This was reflected in [G2P3]'s remark that explanations were shallow and did not offer "a firm direction as to what to do". In a similar vein, [G3P1] described clicking "explain" because they were "looking to see like the actual suggestion rather than how the assistant works", reinforcing the idea that the expected explanation should be situated in the task-related context and informative about specific actions that the IPA suggest or the user was about to undertake. This expressed aversion to explanations that foreground *architectural transparency* (how the system works) rather than *interactional transparency* (why *this* suggestion was made *now*) is also identifiable in [G7P1]'s wish for explanation to be "[o]nly very practical. Keep the practical part. Otherwise, I don't care. It is just a tool".

Similarly, participants felt that some explanations were *too verbose*, leading them to lose interest or feel like spending time was not a worthwhile investment, with the statement by [G8P2] on multiple levels of explanations: "There's three levels of exploring, right? So the first level was pretty much my go-to for whatever I would... kind of sort of require how it's making this decision".

In the third group, regarding the *unreliable explanations*, participants communicated primarily when asked to elaborate on their views of explanations provided by LLMs, rather than those of SIPA4TM. [G5P2], [G6P3], and [G7P1] did not consider such explanations to be faithful accounts of what the system actually did. This concern surfaced both as a general scepticism towards inherently opaque systems ("when it comes to AI, I feel like I'd be a bit more sceptical, because it's a bit more of like a black box" [G6P3]) and as suspicion that the system may show an incomplete picture ("maybe that's just what they want to show you. Maybe they actually have more information that they're not really talking about" [G6P3]). The latter also fits with the concern regarding interest misalignment, that we identified in the context of the theme on Data Use. [G5P2] stated they learn about AI through "trial and error", indicating that the nondeterministic nature of LLMs makes their responses unreliable: "So you can ask ChatGPT the same question three times and it'll give you three different answers or the same answer. But I'm very reluctant to rely on what these large language models actually came up with now". This untrustworthiness of explanations could explain in part why some participants, including [G2P3] and [G4P1], expressed that they rely on data provenience and answers by alternative LLM-based chatbots to verify the accuracy of originally received responses, rather than on post hoc explanations.

**Integrated Scrutability and Control** In contrast to the above, participants described a set of explanation practices that *integrate* scrutability and control into the goal-driven interaction: explanations were valuable when they were tightly coupled to a concrete suggestion, when they enabled rapid verification, and when they offered rather than demanded a credible path for

further exploration and, where appropriate, error correction.

Across groups, the appetite for explanation was primarily *diagnostic* and *task-oriented*. While it may be that this was due to the protocol at hand, the similar sentiment in expressed views regarding explanation in chatbots ("if there's some error with that, then I can potentially resolve it or start a new chat" [G8P2]), IPAs generally ("if it's making the wrong assumptions, you'd be interested to see why it's doing that and how you can correct it then" [G8P1]) and calendaring tools in SIPA4TM ("it's very functional, very consequential to this relationship I would have with my calendar /.../ [o]nly if it was going wrong would I be interested in seeing if there was ways to tweak it and fix the ways to make it work properly" [G1P2]) indicate a more general finding. Namely, participants were expected to scrutinise when something appeared wrong or surprising, not out of curiosity or to improve performance. For example, [G1P2] stated that if the system is consistently correct, they would become "less and less likely to look into how it's making those decisions", but if "it's not going so well, then I'm going to probably be more interested in why it's not going so well". Similarly, [G4P2] described their stance as "if it ain't broke, don't fix it. If it's broken, then take a look".

When explanations were desired, participants often preferred explanations embedded in the *main interaction mechanism* of suggestion-making and approval. [G1P2] described this integrated ideal explicitly: "if it were the case where it was like, 'Hey, we noticed there might be actually', you know, 'a better way to basically organise your day to day' and [the IPA] went through basically the reasons one by one and the change that have been made and that you can approve them, then I'd be much happier doing that rather than just like it kind of happening in the background with any rationale given whatsoever". On the flip side, explanations or even just suggestions that break the user flow without providing task-related information or that are unaligned with the intent of the calendar interaction were explicitly criticised by [G2P2], [G5P1], and [G6P3]. Beyond explanations and scrutiny, we identified the desire for user control to also be integrated with the main interaction mechanism, allowing the user to steer IPA learning, with [G4P2] stating this directly: "can I edit this explanation? Can I give an input? Can I like tell you feedback of like, oh, this is not completely right. This is what I want actually. And then it like learns and fixes itself".

**Stages** Finally, we examined participant responses to the *stages* mechanism, a design feature intended to couple data permissions with distinct levels of assistant proactivity, as illustrated in Figure 4.4 on page 61. Our analysis reveals that participants typically viewed stages less as a privacy control tool and more as a scaffold for *expectation*.

As [G1P1] articulated, the clear delineation of data sources was valuable because it "sets the expectations of what I know it's going to be able to do based on what it's using". This utility of

stages for *expectations management* was further made explicit by [G8P2] and [G1P2] who liked that SIPA4TM made it "really clear what was being used" and thus "sets the expectations of what I know it's going to be able to do based on what it's using as well". Similarly, [G6P1] noted that the stages could allow users with different personalities to align the level of assistance to match their preferred levels of desire for explicit control.

Some participants saw stages as a context or *domain switch*, where different levels may be appropriate for different parts of life and different tasks. For instance, [G6P2] appreciated that SIPA4TM, through stages, "gives different suggestions for different events in life". Similarly, [G5P1] described the utility of stages when deciding whether to use the IPA, stating they would "[p]robably try the one-star, two-star, three-star tour, a period of time each and say if there is something that came up like really helpful". Such statements have directly inspired the design guideline of *Domain Attunement*, which we present in Subsection 6.3.2.

However, this appreciation for graduated control was not universal. Some participants struggled to distinguish the practical differences between levels. [G2P1] remarked that they "didn't really see much of the benefit of the final stage". [G4P2] reported not feeling confident enough to estimate whether giving more permission "was going [to] screw some things up in my calendar or not," and therefore "just kept it at two stars for everything". This confusion led some participants to request stronger onboarding, such as a "wizard which shows every part of it, how it works" [G8P3]. These findings indicate that without a shared mental model of why the stages exist, users may default to the highest setting simply to bypass the cognitive load of decision-making.

Some participants viewed the stages as an unnecessary friction to utility, expressing that their desire for maximum efficiency overrode their desire for granular control. [G7P3] exemplified this by stating "I would just go ahead with the highest [stage], for sure, because then I just can use my Google Calendar". A very similar line of reasoning was given by [G2P1] ("I'd be happy to let it know everything, as long as it's the most useful"), [G4P1] ("I'm probably willing to give over that control on scheduling my calendar because I hate scheduling calendars"), and [G1P1] ("I think different people might want different things. Personally, I would want to give it as much control as possible"). This sentiment was further reinforced by [G9P1] who stated they would want to "give it as much control as possible" to maximise the system's potential, while nevertheless appreciating the transparency stages provide: "I would like more companies or platforms to be more upfront about where that data is going" and [G8P2]: "I'm presuming level one would be more like standardised, generalised sort of advice on events. So, since I would like to organise everything properly and I would like good suggestions for organising my calendar, I would pretty much opt in for level three most of the times based on the sort of inputs it requires".

### 5.5.3 Trust

Across the expert groups, trust was closely intertwined with capability, control, and scrutability. Participants repeatedly distinguished between trusting an IPA to *assist* and trusting it to *act on their behalf*, with the latter being a much higher bar. We identified two interrelated dimensions of trust: comparisons between IPAs and human assistants in trust-building and considerations on how time affects trust levels. After reporting the two, we introduce the theme's only subtheme: trusting the IPA for agentic features, which captures instances when participants choose to use the IPA as a proactive agent to make changes in their calendars for them or to delegate for them in communications with others.

**AI versus humans in trust-building** When explicitly comparing IPAs to human personal assistants, participants in most cases described humans as the more trustworthy, especially for tasks that can affect the participants' social and self-realisation prospects. This favouring of humans is typically based on perceived differences in cross-contextual reliability [G4P1], communication rapport [G7P3], and accountability [G3P1]. [G4P1] explained that their workdays differ a lot when they are travelling and questioned whether an IPA could reliably handle such edge cases. As previously discussed in the theme on capabilities, [G7P3] expressed his belief that the IPAs "would reach that level" of personality judgment of their users. [G7P1] shared the sceptical view of AI in understanding humans, noting that "outsourcing" tasks to IPAs currently does not necessarily imply actually reducing workload for the user, due to the efforts needed for prompting and steering, concluding that "that's not outsourcing" and that "it's just gonna help that personal assistant do their job better". [G3P1] believed that human employees have better "initiative" to retain their position, making them more trustworthy, further stating that humans are more trustworthy as "if you're paying someone to be your assistant, they have that initiative to /.../ work for you"'. The idea of accountability was also related to the IPA's lack of capability to suffer, as with [G4P2] who stated that it is the case for any AI system that "it can be wrong, and if it's wrong, I'm the one who will suffer. And then there is no suffering for ChatGPT".

At the same time, participants did not universally reject AI as untrustworthy. Instead, trust was framed as *conditional* and *task-dependent*. Contrasting the accountability argument by [G4P1], [G2P1] stated IPAs represent a "free alternative" compared to paid human assistants, which could justify accepting a different trust profile if the system performs "almost the same job or better, maybe, in some cases, then why not consider it". On another pragmatic note, [G3P2] questioned whether trusting a human and a data provider with their time engagement is categorically different, saying it may be reasonable to say that "if you're sharing it with the person might as well share it with the data provider, because I would say that in today's world

– rather than the person giving you a better suggestion – the model would give you a better suggestion".

Importantly, participants also emphasised individual differences in trust orientation. Some expressed a strong inclination to grant extensive control if the system proved useful, such as [G1P1] who stated: "personally, I would want to give it as much control as possible", while others, including [G9P2], articulated a principled reluctance to ever trust an IPA with their time. These accounts suggest that trust in IPAs is not only a property of system design, but also of personal disposition, life situation, and prior experience with automation.

**Temporal aspects of trust and reliability** [G2P3] drew another analogy to working with a new human assistant: at the beginning, close supervision and verification are required, "[a]fter a certain period of time, you would trust them more to make decisions on their own. I think it would probably be the same with an AI assistant. At the initial stages, you would want them to consult you on any decisions. And then as the AI gets to know you better and you get to know it better". Such temporal framing of trust aligns with other participants' statements, who repeatedly emphasised the importance of reliable performance for trust. For instance, [G2P1] answered a question on the effect of explanations on their level of trust towards an AI system by saying: "As long as it does what I want it to do, then I'd be happy to trust it, yeah". Following this, [G2P2] noted stated: "it's less a question of trust, more so understanding why this is suggesting the thing it's suggesting", thus implying a connection between trust and scrutability. The question was then answered by [G2P3] who, in turn, further connected their feeling of trust with data governance and data security of the underlying service providers. These answers indicate that trust is related to scrutability and data use, but only conditionally, through prospects, as we discuss in Section 6.1.

Reliability emerged as an important prerequisite for trust in other focus groups as well. For instance, participants in focus group 8 discussed their experience with various LLM-based chatbots, whereby [G8P2] said: "mostly I just want my chatbot to be really reliable", underlining that task performance was more influential for them than the traceability of reasoning steps taken to get to the final answer. [G8P1] and [G8P3] agreed with this sentiment, with [G8P1] also expressing a concern that the reliability of contemporary LLM-based systems was not improving monotonically, and in some cases was perceived as declining: "it's become more unreliable. I've actually found that I'm using it less now than I was previously". Participants also articulated clear limits on how quickly trust could develop. Sudden or large proactive changes were widely rejected, even by those otherwise open to automation. For instance, [G2P2] described idea of waking up to an autonomously altered schedule as unacceptable: "I wouldn't like to wake up the next day and my calendar's all changed on its own".

### Decision to use Agentic Features

The role of social standing prospect was accentuated when participants were asked whether they would allow a more advanced version of the assistant to communicate with others on their behalf or act as an agent and delegate for them in communications with others. When thinking about such *agentic features* and the consequences of potential mistakes, the participants routinely expressed concerns about how others would perceive them as a result of their use of the assistant. This concern can be divided into three broad categories: 1) concerns about the ability to communicate appropriate information and to do so in an appropriate manner; 2) concerns about the degradation of the conversation; and 3) concerns about AI technology changing or not fitting with current social norms. At the intersection of the last two, we also identified that some participants categorically rejected delegation in personal communication. Table 5.9 below provides an overview of codes for each stance.

Table 5.9: Numbers of groups, codes (refs), and participants and lists of participants coded for each stance on having the IPA delegate for them

| Stance | Groups | Refs | Participants | |
|---|---|---|---|---|
| Would agree to delegate | 3 | 6 | 4 | [G1P1], [G4P1], [G4P2], [G7P2] |
| Conditional acceptance of delegation | 8 | 22 | 14 | [G1P1], [G1P2], [G2P1], [G2P2], [G2P3], [G3P1], [G3P2], [G4P1], [G5P1], [G6P1], [G6P3], [G7P1], [G7P3], [G9P1] |
| Would restrict delegation before agreeing to it | 7 | 42 | 13 | [G1P1], [G1P2], [G4P1], [G5P1], [G5P2], [G6P1], [G6P2], [G6P3], [G7P1], [G7P2], [G7P3], [G8P3], [G9P1] |
| Disapproval of delegation for personal use | 4 | 16 | 7 | [G2P1], [G2P3], [G5P2], [G7P2], [G7P3], [G9P1], [G9P2] |
| Delegation leads to loss of emotional, physical conection | 8 | 34 | 16 | [G1P1], [G2P1], [G2P2], [G2P3], [G3P1], [G3P2], [G4P2], [G5P1], [G7P1], [G7P2], [G7P3], [G8P1], [G8P2], [G8P3], [G9P1], [G9P2] |

Loss of connection, i.e. a fear of the *degradation of communication and their social standing* (be it as trusted colleagues, as beloved friends, etc.), may in part explain such a stance. We identified four specific channels of communication the participants were concerned about: loss of emotional connection [G9P2], loss of social cues [G2P1], loss of physical touch [G3P2], and unconvincing humanness enactment [G5P1]. In deliberating on lost connections, some participants explicitly linked it with trust, with [G3P2] claiming "I think trust is a factor, but that emotional connect can never come with AI. You always need that physical interaction, physical touch and emotional connect with the human, even if he's not as good as AI" and with [G3P1] agreeing with the sentiment. Similarly, [G8P1] expressed doubt if an API could

notice and appropriately react to fluid social cues needed for successful communication: "if you're changing that [meeting] time, I think the personal approach is definitely more beneficial because you can gauge how that person's going to react. And also, you can then approach it differently as well, whether it's over an email or a phone call or how you kind of manage that". Similarly, [G2P3] though that full automation of communication may not be possible, as "it's like people still want[ing] to sign paper letters or something for a personal touch for, there's always going to have to be some personal touch to communications".

In light of these concerns, participants proposed several ways to *restrict the assistant's capabilities*. One way was to limit the form of the sent communication: "One thing is it's going to be a generic email, but sometimes you'll want some personal touch if it's an important meeting" [G5P1]. Notably, as [G1P2] states, such simple, more standardised forms limit the danger of undesired content being shared: "If it's just a case of like, okay, we need to reschedule, and then an automated message goes out, it doesn't even have to be an LLM-based message. It could just be using the reasoning of the model to say, okay, this needs to change at this time. I'm going to, you know, check with the other platforms and see if there is a, like, optimised time that we can put this on instead that works for everybody". [G1P2] further explicitly separated this type of delegation of communication from the use case where the IPA generates a less standardised and more detailed message, of which they would be more wary: "If it was the case that there's an email being sent out, like a verbose email being sent out, then I'd be worried, because it depends on what data it will have access to from other parts of your calendar". Here, [G1P2] mirrored a concern by our expert in the discussion as well (cf. PA-S192 in Subsection 3.2.8 on page 43), and proposed an idea very similar to stages. However, even though the [G1P2]'s statement can be read as a concern about the data used, it is ultimately a concern about a degradation of one's social standing prospect that is the ultimate damage: "There could be a leak then from an event in your calendar that you don't want the other person to see necessarily. And that could get sent along to emails. It depends on what data it's using". So, rather than the data itself, it is what data about the user others see – and how it ultimately affects the main user's social standing. Relatedly, [G1P1] was quite categorical in allowing the assistant to act as an agent within the constraints of their own calendar, but not in communications with others, stating "I guess I have no problem with it, reading emails and updating my calendar. If I saw that someone replied and said I can't do that on Friday and then [the IPA] just deletes it or whatever, that'd be fine, but not the ability to send stuff".

While still wanting some level of control or capabilities restriction, some participants, including [G2P3], [G4P1], and [G6P1], expressed *enthusiasm about delegating work-related communications*, while retaining personal communication, with [G6P2] stated they would delegate communication in a work context, concluding that "if there was a way that I could book ap-

pointments within a work context, it would be sweet". [G5P2] noted that, while they would be interested in seeing calendaring suggestions, they would not let the IPA plan their entire week for them. Similarly, [G1P2] speculated that they "would be only okay with it in a situation where there's other people using the same platform who are also invited, purely invited to the same meeting as me". Automatically coordinating availabilities was a desired use case highlighted by [G9P1], especially when "an event that involves other people, those are the ones that are less flexible to update". Yet most participants categorically rejected delegating time management communication for personal events. Several participants, including [G2P3], [G7P2], and [G9P1] said they would feel disappointed if they found themselves on the receiving end of IPA delegation, with [G7P3] stating: "It would hurt me. /.../ on a work front, it is fine, /.../ [but] when it comes to personal stuff and people are typing messages from ChatGPT, I'm like, don't you know who you are? You need somebody else to type your message?!". Similarly, [G9P2] noted they would consider it inappropriate to not personally request a meeting rescheduling: "I don't want the person to know that I actually was so lazy, I didn't actually bother saying 'Hi, researcher, sorry I can't make it today', you know, just that it's me". Regarding the use of AI in research tasks, [G4P1] described their approach to using LLMs: "I use lots of ChatGPT to /.../ do a literature review for me, find facts, help me draft something,... But mostly speaking, I will finish the draft myself", emphasising that do not want to give "ChatGPT agency in the work that I'm doing, but it is assisting me in the jobs that I'm doing. Like I'll take responsibility for the outputs of it".

Among the participants, the most *open to using agentic features* was [G1P1], who stated: "I would want to give it as much control as possible. I would trust a program, if it had a lot of context in my life, to just, even in the background, make those changes". They stated that while suggestions-based IPAs could still be beneficial, an ideal use case is one that is proactive and delegative: "if it could reorganise everything in the background automatically and change everything, even if that had knock-on effects, if it could do all of that automatically and just tell me at any one moment, here's what you're going to do, I think that would be really cool". Similarly, [G4P1] was quite open to exchanging emails with a bot, stating that such interaction would become "socially and culturally accepted more", and offering an analogy with the current level of acceptance for Google Calendar invites. [G4P2] agreed with [G4P1], stating they would also use a bot to answer their emails if they could, and that the underlying hope was that it would allow humans to "spend their resources somewhere else where it's more needed". This framing aligns with the prospects of reducing time and effort required and of self-concept actualisation.

## 5.6 Desired Time Management

As Yildirim et al. [2023] reported, 85% of AI innovation projects fail, with many breakdowns stemming from a lack of human-centred design and inappropriate problem selection and formulation. Heeding this warning, we summarise the existing time management problems and the envisioned IPA use cases that participants voiced during both the thinking-aloud interaction and the subsequent focus group discussions. These desired use cases included practical support for everyday organisation, prioritisation, and cognitive offloading. While some hoped for a fully autonomous and proactive planner, most participants described a system that actively assists with the rationalisation of one's schedule, highlights potential problems that may be missed, and suggests actionable interventions at the user's discretion.

### 5.6.1 Action-Ready Advice

A recurring desire was that the assistant should help *resolve the problem* rather than just announcing it. Both the problem description and the proposed solution should be *presented* in a succinct, yet informative manner. At the same time, participants did not want *every* problem to trigger rescheduling.

**From problem reports to actionable resolution.** During the thinking-aloud interaction, participants frequently interpreted "clash detected" as a prompt to act, followed by surprise that there was no option to "automatically" resolve the clash in line with the suggestion. In focus groups, this surprise translated into an explicit request, among others, from [G7P1] who stated they wished the IPA "would say something like: 'You have a problem here. Do you want me to reschedule?' And then I say yes or no". Similarly, [G8P1] stated a core usability limitation of SIPA4TM for them was a lack of solution implementation: "The element where I found a bit of frustration was the assistant was making suggestions, but I couldn't implement any suggestions, so I manually had to change them myself. So, you know, notifying me that there's a clash, but then not proposing, you know, if you moved it, but not giving me the option to say yes or no, where I think that would have been more beneficial in terms of time, instead of me having to manually do that myself or reschedule it".

**Presentation requirements: concise, legible, and solution-forward.** The desired time-management assistance was not only about what the system suggests, but also how it communicates it. Participants repeatedly asked for shorter, more legible, and more solution-forward messages. For example, [G7P3] observed that "suggestions were just about the problem, not about the solution", while [G7P2] wanted a hint of what can be moved: "with the clashes, it could probably be better if it would give us a hint of maybe these could be moved

or they should stay here. Yeah, just on a click, OK, modify this". Similarly, [G7P1] requested "a small message" such as "traffic is 20 minutes and you don't have time and that's it" rather than "heavy text" and complex charts. These preferences tie directly to the findings presented in Subsection 5.3.3 that scrutiny is often purpose-driven and that interpretability issues can impose disproportionate cognitive overhead, implying that participants' desired assistance is one where explanation and action are tightly coupled and where the interface supports rapid, low-effort decisions.

**Fast paths for ignoring and filtering assistance.** A complementary desired capability was therefore a fast path for dismissal, deferral, or "good enough" decisions. For instance, [G6P2] objected to having to restructure a day for minor adjustments: "I was surprised not to see a quick way to ignore suggestions /.../ I know that I have 15 minutes between when I finish lunch and when I have to be at the doctor's /.../ I don't want to have to go through all the trouble of moving lunch closer and making sure that I have enough time". In this sense, desired assistance included not only "do more" but also "get out of the way" when the user already has a simple coping strategy.

### 5.6.2 Personalised Assistance

The participants wished the assistant would learn recurring patterns and preferences, thereby reducing the need for manual corrections. Beyond individual events, participants also articulated a desire for higher-level optimisation and reconfiguration of busy days or weeks.

**Learning routines and tailoring assistance to the person.** Rather than requiring constant corrective explicit user control, [G5P2] proposed that an assistant should learn "your weekly routines" and that "having some way of even just writing out your work week to start off with and then the AI learns from that rather than fitting in and changing the model on the fly explicitly" would be more useful. At the same time, participants emphasised that such learning should respect the diversity of calendaring practices and avoid imposing a singular, non-personalised notion of well-used time. For instance, [G9P2] described the SIPA4TM's structured schedule as mismatched to their own more fluid work rhythms: "I would wake up on a Monday and look at that and think, overwhelmed /.../ my work life just goes that if meetings run over, it runs into your lunch, you take it at another time /.../ lunch can be really any time". In this sense, desired assistance was conditional on apt personalisation: learning what the user treats as negotiable (e.g., lunch timing) versus fixed (e.g., externally scheduled meetings), and adjusting both the content and the tone of intervention accordingly (cf. nudging and self-concept concerns in Subsection 5.5.1).

**Maintaining commitments, including what disappears during replanning.** Participants also envisioned an assistant that supports longitudinal maintenance of commitments and recurring intentions. A prominent example was the expectation that the system should notice when a recurring task has been displaced. After a scenario in which an event for doing laundry was removed to accommodate another event, [G7P1] noted that the assistant did not flag the missing obligation: "so I basically removed the laundry, and I changed it for pickup from the airport to the front. Then in the suggestions, it didn't tell me that, oh, you have not [scheduled this week's] laundry". [G7P1] then suggested that reminders could operate at two levels: "warnings that you have a clash" and higher-level prompts such as "oh, you didn't do cleaning, shall I schedule cleaning?". This is consistent with the broader focus-group emphasis on the assistant helping manage busyness over a day or week (e.g., "I would like to be able to prioritise events and would be open to suggestions based on the priority of those events to say, oh, you've, like, if it could read that I have a very busy day, you could move events around like that" [G9P1]), and it matches the identified interaction-phase expectations that increasing autonomy should yield more substantive support (cf. Subsection 5.3.3).

**High-Level Optimisation.** Several participants expected the IPA to propose high-level alternatives. For instance, [G1P2] anticipated that the system would "come up with like an active suggestion /.../ a reconfigured version of the time for that day, and you can accept with this button or whatever". Similarly, [G2P1] described the appeal of an IPA that "recognises that there's a clash. And then it already looks for an optimisation in a way. And it can tell you what to sign. So, I think that would be faster in a way than you having a look at it and noticing the clash, and then thinking of the optimisation itself". [G1P1] took the idea a step further and expressed a hope for long-term goal-driven time management, "[s]o you could tell it, 'okay, in the next year, I want to achieve these goals' ".

### 5.6.3 Non-Disruptive Delegation

As discussed in Subsection 5.9, most participants were explicit in wanting to limit delegation of IPAs when interacting with other people, especially in private capacity. Nevertheless, participants often acknowledged that shared calendars already serve for social signalling. Finally, several participants proposed use cases that go beyond assisting in maintaining an efficient calendar.

**Bounded Delegation.** As we have seen in Subsection 5.5.3, participants varied in their comfort level regarding the agentic capabilities. For instance, [G4P1] said they were comfortable with the assistant coordinating logistics: "I have had email threads where some tech CEO CCs in a bot and the bot coordinates the meeting. I didn't go, oh, I'll never do that. I'm like,

oh, I'm talking to a bot now. You have to coordinate a meeting. I'll do that. That's fine". [G4P2] framed delegation of communication as a resource-saving move: "if I'm talking to a bot and people can focus on other thing, then they can spend their resources somewhere else where it's more needed. I understand it. I would do the same thing". However, most participants expressed stronger reservations about an assistant acting on their behalf in socially sensitive domains, particularly where mistakes would have amplified consequences. For instance, [G6P3] emphasised the importance of retaining agency: "if [the IPA] wants to suggest that I do that, then that's OK. But /.../ when you're interacting with other people, I don't think I want to have some agent doing that for me. /.../ If it's just changing my own calendar, I would be more comfortable with that than changing things that I'm doing with other people. Because in that case, if it makes a mistake, then there's more of a consequences". Furthermore, as discussed in Subsection 5.5.1, delegation in personal communication was often highlighted as especially undesirable and even socially destructive for one's social standing prospects.

**Representing flexibility, priorities, and social constraints.** Participants' desired time-management support was grounded in how they already reason about calendars: not as a complete model of life, but as a negotiation between fixed commitments, soft preferences, and social constraints. As described in the Subsection 5.3.3, participants repeatedly evaluated what is movable and what is not, often based on who is involved and on the perceived legitimacy of a clash. Desired assistance, therefore, included making such tacit judgments explicit in the interface: categorising events by flexibility, and using that categorisation to shape solutions. During interaction, [G9P1] articulated this directly as a design logic: "work events obviously would be less flexible compared to lunch or post lunch walk or gym" [G9P1]. In focus groups, participants similarly noted that, for events involving other people, flexibility is constrained: "if you have a meeting or if you have an event that involves other people, those are the ones that are less flexible to update" [G9P1]. This connects to the underlying finding identified in several places in our thematic analysis in Section 5.5, namely that the IPA is evaluated by its ability to respect real-world constraints that users treat as non-negotiable.

**Beyond the calendar.** Some participants also described desired future use cases that extend beyond calendaring per se, situating time management assistance within broader domains of task completion and information retrieval. For example, [G9P2] imagined an assistant that could handle "those big leaps" in everyday bureaucracy, such as navigating to the correct form or service endpoint, thereby reducing friction and cognitive load. Similarly, [G4P1] envision an IPA might order them a taxi or help organisations with their procurement processes. While such scenarios fall outside the scope of SIPA4TM, they illustrate that the desired time management support is often intertwined with reducing task overhead rather than solely calendar manipulation assistance.

## 5.7 Chapter Summary

This chapter detailed our three-phase user study, conducted in person with 22 participants. The first phase was a survey with demographic questions and questions about participants' prior experiences with AI in fields adjacent to IPAs for time management. The second phase was a thinking-aloud task-based interaction with SIPA4TM. Finally, the main source of data for our study were nine semi-structured focus group discussions with two to three participants. The three phases lasted around 90 minutes in total.

We tested for monotonic trends between ordered attitude groups (on self-reported interest in AI, experience with requesting data from online service providers, or perceived importance of explainability and privacy correlated) and the aggregate measure of scrutability engagement (sum of interactions with explanations, stages, and user model views) using the JT-test. As shown in Table 5.4.2, after the Holm p-value correction for four tests, only a weakly significant trend remained for the data request category at $p = 0.0926$. Thus, we did not find that participants who self-reported a stronger interest in AI or in the importance of explanations or privacy reliably engaged more with scrutability features during the interaction.

The verbal protocol analysis based on transcripts of thinking-aloud interactions with SIPA4TM has revealed that participants often tried to investigate the system indirectly, through probing and trial-and-error, rather than using explicit scrutability features (e.g., clicking the explain button, looking at the user model, etc.). Moreover, common sources of user frustration were proactive suggestions that disrupted the flow of interaction, and what was perceived as insufficient automation, that is, when SIPA4TM was unable to resolve the issue, rather than just offer suggestions. The participants have typically shown only superficial interest in both the graphical explanations and the user model fields.

Our thematic analysis has revealed that participants tend to consider expected utility and risk when deciding whether to trust an AI assistant, especially when deliberating over agentic features, such as proactively changing their calendar without explicit approval or delegating for them in communications with others. In discussing both the appeal and the concerns regarding AI-based assistance, the participants relied less on the features themselves and more on the expected *prospects* of the assistance. The codes from our thematic analysis supported defining prospects along three axes – time and effort, self-concept, and social standing – as well as identifying two further themes: trust and human-assistant interaction, connected not directly but through prospects.

Lastly, a focused analysis of the desired time management assistance has revealed the domain's complexity and social embeddedness. Overall, the participants wanted personalised assistance that accounts for the user's routines, helps them optimise their time while maintaining com-

mitments important to them, and signals changes appropriately. The advice given by the IPA was expected to both identify the problem and provide an actionable resolution and to present this advice concisely and legibly, with an easy pathway for ignoring and filtering assistance. Lastly, while delegation of communication was often completely undesired, some participants were still open to delegating business interactions, provided the IPA understood the user's priorities and the broader social constraints.

# 6 Designing for Scrutable Interaction

In this chapter, we synthesise our proposed principles of scrutability based on our findings across the user studies. We start by formalising the concept of *prospects*, which we found to be a paradigmatic connection between scrutability and human-assistant interaction with the user trust, in Section 6.1. We then define the final proposed principles of scrutable proactive IPAs in Section 6.2, as well as the roles and design implications in Section 6.3.

## 6.1 Prospects-Based Framing of Desired IPAs

As was discussed in Section 5.5, we identified *prospects* as the central theme of our analysis. We found that trust in an IPA does not emerge directly from system interaction, but through users' perceived prospects across three interrelated dimensions: anticipated time and effort, alignment with the desired self-concept, and expected social consequences. Figure 6.1 illustrates this three-axial model of trust.

As the figure shows, some prospects can be perceived as ambivalent or negative (in this case, social standing and self-concept), while others are neutral or positive (time and effort). This specific quantification is extrapolated for illustrative purposes from the statement by [G3P1]: "I feel like there's nothing really much about my schedule that I would be too worried about people knowing. I'm sure it could get into the wrong hands, but still it's... I kind of feel a bit... I kind of trust a bit more than I feel like I should". A possible reading (made likely by broader context of the discussion reported in Subsection 5.5.2 in the paragraph on the "no-harm thesis") is that the participant estimates that the content in their calendar may be misused but that such misuse is unlikely (demonstratively quantified to $-0.25$ on the Social standing axis in Figure 6.1), but that they would still trust them due to time and effort benefits (illustrated as 0.75 on the Time and effort axis in the figure). Notably, [G3P1] statement also contains a meta-commentary on trust: "I kind of trust a bit more than I feel like I should", which indicates a possible felt divergence between their ideal self (i.e. "I am not a person who trusts something they don't understand") and their experienced self (say, "I use the IPAs because they make my life easier, even though I believe my data might be misused"). In the

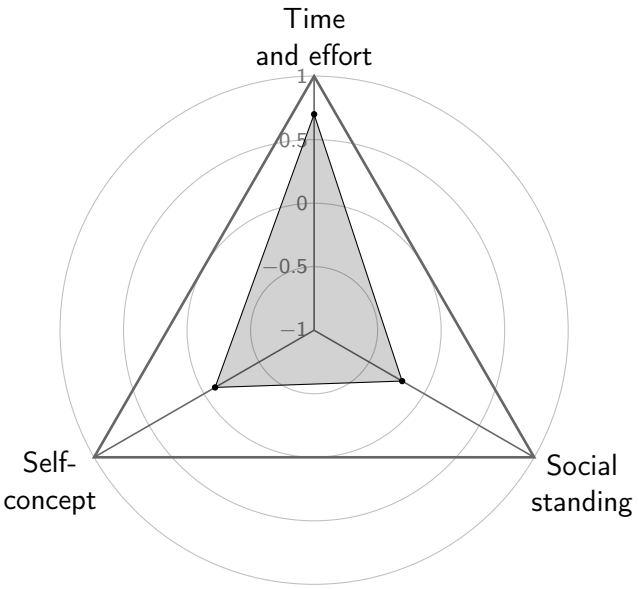


Figure 6.1: Radar chart illustrating how user's trust towards the IPA (quantified as the area of the inner triangle) emerges from users' perceived prospects across three interrelated axes: anticipated time and effort, self-concept actualisation, and expected social consequences; concentric rings indicate perceived harm (negative values down to $-1$), ambivalence (0) and likely benefit (positive values up to 1) for each axis

extrapolated example profile in Figure 6.1, we thus set self-concept to $-0.1$, i.e. likely slightly harmful for actualisation. The overall trust can then be quantified as the area of the triangle whose vertices are the corresponding values on the three axes for the three prospects.

### 6.1.1 Situating Prospects

We now situate the prospects within the findings of our thematic analysis and the prior publications in the domains of psychology, organisational science, and human-computer interaction. We will anchor our approach around the statement by [G2P3], in which the participant explained how they would decide whether to let an IPA delegate for them without prior approval: "I think it's like having a human assistant. After a certain period of time, you would trust them more to make decisions on their own. I think it would probably be the same with an AI assistant. At the initial stages, you would want them to consult you on any decisions. And then as the AI gets to know you better and you get to know it better, you can then... [the participant was interrupted here] /.../ I suppose when I said trust first, I meant more like that you can rely on their decisions rather than trusting that you're... So, it's a separate thing to say, well, is your data being given to somebody else? That should be security or whatever. But yeah, I mean, I'm happy with something like that making decisions or making suggestions. But I suppose, you'd have to have a long experience, whether to leave it to make decisions

without you being consulted or without you making the final okay".

To analyse how this statement fits within the prospect-based framing of trust and interaction, we ought to first distil the distinct claims expressed, then contrast them with further evidence from our study and prior publications. In this statement, we can separate the following claims:

1. Trusting IPAs is comparable to trusting human assistants ("I think it's like having a human assistant");
2. Trust develops over time ("After a certain period of time, you would trust them more to make decisions on their own. I think it would probably be the same with an AI assistant");
3. Trusting the system more would result in fewer explicit verifications ("At the initial stages, you would want them to consult you on any decisions");
4. Trust is conditioned both on the IPA's user model and the user's mental model of the IPA ("And then as the AI gets to know you better and you get to know it better");
5. Trust to delegate is conditional on IPA acting reliably in line with the user's mental model ("But I suppose, you'd have to have a long experience whether to leave it to make decisions without you being consulted or without you making the final okay");
6. Security, privacy, and data stewardship concerns are prerequisites to trust ("I suppose when I said trust first, I meant more like that you can rely on their decisions rather than trusting that you're... So it's a separate thing to say, well, is your data being given to somebody else? That should be security").

In the subsections below, we will revisit each point in the order listed above.

### Trust in Human Assistants and IPAs

As we have elaborated in Subsection 5.5.2, the view implied by [G2P3] that development of trust in other humans is analogous to trust in IPAs was not the majority view. Instead, our participants mostly believed that human assistants are, and are to remain, inherently more capable in understanding intent, in developing and maintaining rapport, and delegating scheduling than IPAs could ever be. For instance, [G7P3] were clear that they though IPAs "won't ever reach that level of convenience and comfort", as "that human connection that are there in terms of having personal assistance and that judgment of you personally – I don't think AI would reach that level". A similar sentiment, emphasising shared judgment, was also expressed by [G7P1]: "with humans, you feel more pressure, kind of, or you trust the judgment more; you trust it".

To decode these contrasting views, we ought to examine how the discourse on trust in human-computer interaction differs from that in psychology. Doing so allows us to hypothesise why a majority of our participants believed there was an inherent discrepancy between trusting human assistants and trusting IPAs.

Within contemporary psychology and organisational research, trust is most commonly defined in line with Mayer et al. [1995], whose vulnerability-based definition is widely regarded as the dominant conceptualisation in the literature [Colquitt et al., 2007, McEvily and Tortoriello, 2011, for instance]. According to Mayer et al. [1995], trust "is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party". Crucially, immediately after stating the definition, Mayer et al. [1995] qualified it by noting that their "definition of trust is applicable to a relationship with another identifiable party who is perceived to act and react with volition toward the trustor". As we argue later in this section, this volitional element may be the crucial factor underlying the identified difference in trust between IPAs and human assistants.

In human-computer interaction, especially within the domain of human-agent interaction, trust is most commonly defined in line with Lee and See [2004] as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". This definition is widely treated as canonical in the field and is adopted by major integrative reviews and meta-analyses [Hoff and Bashir, 2015, Zhang et al., 2020, among others], where it serves as the conceptual basis for research on appropriate reliance and trust calibration. Notably, Lee and See [2004] state that in their definition "an agent can be automation or another person that actively interacts with the environment on behalf of the person". So, the requirement of volition in the definition of trust by Mayer et al. [1995] is replaced with the requirement of capability to interact with the environment on behalf of the person in the definition by Lee and See [2004].

Evidently, trust is defined in a domain-specific manner. This was established, among others, by Lewicki and Bunker [1996], who described diverse foci of definitions of trust across social sciences. Lewicki and Bunker [1996] then proposed their own definition of trust and framework of trust development in an organisational setting. Lewicki and Bunker [1996] introduce their definition of trust by first highlighting the definition of trust by social psychologist Deutsch [1958]: "An individual may be said to have trust in the occurrence of an event if he expects its occurrence and the expectations lead to behavior which he perceives to have greater negative motivational consequence if the expectation is not confirmed than positive motivational consequence if it is confirmed". Lewicki and Bunker [1996], note that Deutsch [1958]'s definition emphasises the transactional element of trust, while noting that trust was broader than expec-

tations and also integrates risk predictability, which we will further discuss in Subsection 6.1.2. The framework that Lewicki and Bunker [1996] propose, however, is based on the definition of trust by Boon and Holmes [1991], which defines trust as "a state involving confident positive expectations about another's motives with respect to oneself in situations entailing risk".

With the Boon and Holmes [1991]'s definition of trust in mind, we now introduce the framework of trust in professional relationships by Lewicki and Bunker [1996] and then comment on the identified divergence between trusting human assistants and trusting IPAs. Thereafter, we will examine the conceptual utility of prospects through the lens of the framework. In the framework, there are three types of trust, which develop one after another over time:

1. Calculus-Based Trust (CBT), based on the calculation between expected rewards from preserving the trust and the risk of punishment for violating the trust;
2. Knowledge-Based Trust (KBT), where trust rests on available information gathered from the history of mutual interaction rather than deterrence; and
3. Identification-Based Trust (IBT), which develops when parties can effectively understand each other's wants, needs, and preferences, and share some of them as their own.

We postulate that the difference most participants in our study reported in their feelings of trust towards an IPA relative to trust towards another human stems from the fact that an IPA is seen as a tool without inherent values or preferences. Crucially, the IPA's "true motivation" may be constructed by the user, for instance, based on the perceived values of the company that developed it (as we have seen happen in our study as well). But human assistants are seen as motivated by graspable value systems grounded in subjective experiences, concerns, and aspirations. Therefore, participants find it hard to imagine moving to an identification-based trust that explicitly necessitates identifying values and preferences with an entity that is seen as having no subjectively felt values and preferences.

Finally, the three-staged model of trust by Lewicki and Bunker [1996] enhances the conceptual utility of prospects. Lewicki and Bunker [1996] framed the reward and punishment of the calculus-based trust primarily around reputational gain and risk, as being perceived as trustworthy or untrustworthy would yield wider organisational benefits or drawbacks. However, such reputational calculus is less straightforward when one party is an IPA, as the IPA lacks a social network to spread word of whether it keeps or breaks its word. Similarly, our participants believed they would ultimately be responsible for their IPA's actions, as the IPA does not have a personal reputation to worry about or a subjective capacity to experience worry ("it can be wrong, and if it's wrong, I'm the one who will suffer. And then there is no suffering for ChatGPT" [G4P2]).

Thinking about trust in terms of expected returns in time and effort, self-concept actualisation, and the impact on social standing is thus fully in line with the calculus-based and knowledge-based trust. The three dimensions are all based on the IPA user's estimates of their own future states, and none of them requires the IPA to have any motivation or reciprocal trust towards the user. Prospect-based framing of trust, therefore, removes the necessity of volition presupposed by Mayer et al. [1995] while also not requiring a framing of an environment-affecting capability found in the definition by Lee and See [2004]. Instead, our framing allows us to think about trust as a one-sided phenomenon, in which trust is based on prospects that change as the *perceived capability* of the "other side" – be it an IPA or another human – is revised through interaction.

### Temporal Aspect of Trust

In a reference to Deutsch [1958], Lewicki and Bunker [1996] note that Identification-Based trust – i.e. the type of trust that forms last and is based on identification with the other's intentions and desires – is the relationship where the parties may represent one another in the contexts where the other is not present, i.e. delegate for each other. This fits well with the long time period and extensive hands-on experience that participants, including [G2P3], anticipated would be needed before they would eventually allow an IPA delegate to act on their behalf. As we discussed previously in Subsection 5.5.3, unless participants outright rejected the possibility of ever wanting their IPA to delegate for them, they emphasised a gradual transition over time.

For instance, recall the statement by [G1P2]: "I'd say it depends on how good it is, to be honest. If it's consistently making the right decisions, I would say I'm consistently less and less likely to look into how it's making those decisions, you know, especially if the explanations are good /.../ if it's not going so well, then I'm going to probably be more interested in why it's not going so well". This statement also reflects the importance of the IPA reliably fulfilling tasks, which is in line with the Lewicki and Bunker [1996]'s model and the explanation of the transition from calculus-based to knowledge-based trust. The role of reliability, which the [G1P2] also implies in their statement, is discussed in further detail in Subsection 6.1.1.

Prospects also account for this temporal element. Namely, as the user interacts more with the system, their estimates regarding each of the three axes are continuously updated to become more accurate. Our findings on how these estimates change are discussed in Subsection 6.1.2.

### The Role of Explicit Verification in Trust Development

Within Subsection 5.5.2, we have exhaustively reported opinions on when the IPA should prompt the user to verify its assumptions, as well as the overall stance on allowing the IPA to act on the user's behalf without prior approval. Overall, the stances towards verification ranged from very conservative, such as by [G7P1] ("Any action has to be approved by me and any text has to be approved by me"), over licence-based reasoning about verification by [G7P3] ("I think asking is like getting permission. So it's more like the assistant asks that whether you want the change or not. So it's not just some independent association") and other, towards more hands-off approach, by participants including [G1P1] ("Personally, I want to give it as much control as possible"). A common thread across all but the most conservative of stances was exactly the type of verification that [G2P3] hints it in the statement we are using as the paradigm in this subsection – verification that serves not only as a confirmation that the IPA is correct in its interpretation of the user's intent, but also as a steering mechanism based on the underlying assumption that the IPA would learn from corrections.

As some participants noted, this explicit verification is also time-consuming, and it was typically assumed to be less effective than guiding a newly hired human assistant. [G7P3] believed that this would likely disqualify the use of IPAs for executives and others who are likely to already have a personal assistant: "So in terms of busy people, I think they are way too busy for technology also. They won't have time to put in their own things, that, okay, this is how my schedule is. You have to manage this and that. So I think busy people just need to talk while they're working to their system and then they will like the person".

Through prospect-based framing of trust, we formalise and quantify the trade-off that [G7P3] hints at, in which the payoff across all prospects is conditioned at the outset by a negative prospect along one dimension. Furthermore, quantifying prospects would allow setting personalised thresholds that trigger IPA-initiated verification prompts. We will discuss this potential direction in more detail in Subsection 6.3.1.

### User Model and Mental Model

As we have already emphasised, the prospects-based framing allows thinking about trust even when only one side in an interaction is capable of subjective feelings. Still, as summarised in Section 5.6, participants repeatedly expected the learning process to remain two-sided, with the IPA adjusting its user model of the user's needs as the user adjusts their mental model of the IPA's capabilities and way of working. In addition to [G2P3], [G2P2] has clearly expressed the importance of the user's mental model: "it's less a question of trust, more so understanding why this isn't suggesting the thing it's suggesting". In this, the participants in the study echoed our expert discussion findings and, in particular, PB-S241 (page 41), which

noted the importance of the user's perception of the assistant, i.e., the "internal model" of the assistant, as it both structures the human-assistant interaction and is structured by it. The role of the mental or internal model was also captured graphically by the experts, as shown in Figure 3.4.

Prospects are, by definition, something that exists within the user's mental model of the assistant. In fact, prospects may be seen as the aggregate of all assumptions in the user's mental model regarding the outcomes of potential or real IPA use. Prospects may thus serve as a proxy for the user's mental model as a whole. The IPA can then aim to model the prospects as a measure of the user's current level of comfort with its capabilities.

And as such, prospects can be combined with stages to allow not only gradual user control over the degree of agentic features, such as delegation of communication or proactive calendar changes, but also to serve as a human-assistant verification channel. As we have indicated in the paragraph on explicit verification above, the user could be prompted by an IPA to enable or utilise a certain functionality (or stage) they had not used until that point. Should the user refuse, the IPA could ask for a follow-up reason, so that the user could indicate whether the refusal stems from a concern in relation to one of the three axes of the prospects of using the feature, and if so, enable the IPA an opportunity to better frame the possible outcomes of using the features or just adapting its own user model with regards to the user's preferences. Of course, it is possible that the refusal stems not from a concern about their prospects, but from a personal reference (which is then just to be noted by the IPA) or from perceived unreliability, which we discuss next.

### Reliability

Unlike trust, for which we can relate the meaning of the concept in human-computer interaction to the same term in psychology and the humanities, the meaning of reliability in human-computer interaction differs categorically from that in psychology. In psychology, reliability refers primarily to measurements [Revelle and Condon, 2019], with an authoritative definition set by Cronbach [1947] as the degree to which a measurement is free from random error and yields consistent results across repeated applications under similar conditions.

In contrast, Lee and See [2004]'s work on trust in automation sees reliability as a performance-based *basis of trust*. For them, performance is a dimension that also includes predictability and ability and is perceived in reference to current and historical operations of the automation. Such a framing of reliability is well aligned with the concept *validity of the trust* in Lewicki and Bunker [1996]'s view of relationship-building. While clarifying their stepwise evolution of trust, Lewicki and Bunker [1996] describe that relationship building commences with calculus-based trust and that – if the validity of the trust is confirmed – the parties begin learning about

each other's preferences, priorities, and requirements. In doing so, the parties may move towards knowledge-based trust. Lewicki and Bunker [1996] clarify that confirmation of the validity of the trust happens as "the other party is consistent, and deterrence is not frequently required".

Through prospect-based framing of trust, we can then map the concept of validity of trust from Lewicki and Bunker [1996] into reliability with regards to prospects. Reliability is then quantified as the consistency of returns in relation to the actual outcomes in terms of saving time and effort, self-concept actualisation, and the effect on the user's social standing.

In our thematic analysis, reliability was repeatedly positioned as the primary prerequisite for adoption, with the continued use then depending on positive prospects. This can also be read from the statement by [G2P1]: "As long as it does what I want it to do, then I'd be happy to trust it". The use is conditional on reliable positive prospects and, as [G1P2] states further, "[i]f it's consistently making the right decisions, I would say I'm consistently less and less likely to look into how it's making those decisions". In this view, [G1P2], [G2P1], and indeed most participants, align with [G2P3] regarding the importance of reliable performance, which fits with our framing of prospects as a channel, metric, or reference in relation to which a system is seen as reliable.

A further statement of interest in relation to reliability and prospects is that of [G5P2]: "How do you learn from AI? It's trial and error. So you can ask ChatGPT the same question three times and it'll give you three different answers or the same answer. But I'm very reluctant to rely on what these large language models actually came up with now". The consistency of correct response is a clear example of reliably correct performance. In terms of prospects, reliably correct responses matter because they save time and effort relative to the baseline, in this case, manual information retrieval using traditional search engines (cf. Subsection 6.1.2 on reference dependence). In the statement by [G5P2], we can also identify resoluteness to abandon the system due to negative prospects following unreliable past performance. This is also in line with Lewicki and Bunker [1996]'s deliberations, as they state that trust may remain on calculus-based trust if one or more violations have occurred.

### Privacy and Security

In their statement, [G2P3] explicitly separates their understanding of trust from the question of whether data could be misused. In this section, we argued that [G2P3]'s – and other participants' – understanding of trust is closely coupled with prospects. On the other hand, as [G2P3] put it: "it's a separate thing to say, well, is your data being given to somebody else? That should be security or whatever". Related concerns about data misuse were mostly connected with privacy in our expert discussion analysis (cf. Subsection 3.2.5).

The relation between data security and privacy in the context of IPAs was surveyed by Bolton et al. [2021], who differentiate the former as safeguarding data and the latter as safeguarding user identity (although in Section 4.2.2 of their analysis, the authors implicitly extend the definition of privacy to include any sensitive or private information). A broader but widely recognised source, the article on "AI4People", an ethical framework for AI, Floridi et al. [2018] explicitly distinguishes privacy as control over personal data and informational self-determination from security as protection against unauthorised access, misuse, and harm. Such framing seems to align with [G2P3]'s use of the word security. In the National Institute of Standards and Technology report they authored, Boeckl and Lefkovitz [2020, Table 3] offer another somewhat diverging separation of privacy and security. For Boeckl and Lefkovitz [2020], privacy engineering objectives relate to predictability, manageability, and disassociability, whereas security objectives include confidentiality, integrity, and availability. This conceptual divergence of definitions and the entanglement of two concepts within human-computer interactions has been noted by Dourish and Anderson [2006], who state that the distinction between privacy and security differs between contexts, concluding that "privacy and security cannot be analyzed independently".

In our focus group discussions, we can also see this conceptual entanglement. For instance, [G8P2] saw local data hosting as the only full remedy to their data-sharing concerns. Where there might not be possible, [G8P2] noted stages might be helpful to minimise sharing, while also conceding that: "it's a classic paradox of AI chatbots that if you want more preferences, you want it to understand you better, but you're not willing to /.../ share your personal data". This sharing minimisation and protecting one's identity fits more with privacy concerns, when considering both the definitions by Bolton et al. [2021] and Floridi et al. [2018]. And yet [G8P3] responded to this statement by saying: "I agree with what [G8P2] told about the data. And, I mean, sometimes I understand that probably sometimes the data, at least for a temporary amount of time, needs to be on the cloud to probably analyse it or work on it. Maybe in those aspects, probably some encryption could be possible, like from my end and from the server end to keep it like in a closed loop only within it". In other words, [G8P3] proposed a security-first approach, i.e. encryption and access control, in relation to [G8P2]'s statement on privacy. Moreover, [G7P1] explicitly linked the two while explaining their reasons for not using online calendars, stating that it is "[n]ot because of security and privacy reasons, but just because having it, it's more convenient for me".

With this in mind, in our thematic analysis in Subsection 5.8, we mostly used the umbrella term *data use*. Across the focus groups, participants focused more on what information about them is collected and how it might affect them, and most often, how it affects their social standing prospects. Regarding security specifically, aside from the statements by [G2P3] and [G8P3], participants rarely discussed specific approaches to data security.

It is also notable that our findings corroborate the divergence between actual system security and its perception, with [G1P1] being the only user who mentioned the possibility of investigating the system systematically through a user manual. Yet even [G1P1] qualified the possibility of actually doing it: "I'd probably initially check like the manual, the user manual and how it works just to figure out how I can get the most out of it. But if it was working really well, I wouldn't".

Regarding their stance on the sensitivity of the data they are sharing and concerns typically associated with privacy, participants frequently framed privacy as uncertainty about downstream consequences. As we have described in our subtheme on Data Use in Subsection 5.5.2, participants expressed several strategies that existed in their approach with regard to these unknown consequences. While some, such as [G7P1], expressed complete pessimism over one's data control, saying "if Google has it, then ChatGPT has it, I don't, I mean, we lost it, it's too late, I think. Generally, I don't mind", others expressed a more transactional thinking with regards to sharing their data, such as [G7P3], who explicitly stated they considered "how much data they are collecting" with the criteria being the ration of perceived utility and harm ("is useful to me and not harmful") and the availability and appropriateness of alternatives ("It's about what I need it for. If the task, I can't do it without AI or I need that help right now"), with the preferred state being sharing as little as possible due to the risk of interest misalignment ("I prefer not to give a lot of personal information because that's something very scary because we don't know who is handling the [data and] different companies have their different rules").

As Dourish and Anderson [2006] explain, rational-actor economic models (like the utility theory) are insufficient to explain the relation of the user towards privacy and utility. A sufficient argument for that is the finding that the user is unlikely to learn about the system from a manual. Moreover, as Dourish and Anderson [2006] further notes, privacy and security can only be understood within the broader dynamic context in which they are embedded, such as social implications and questions of identity. Prospects fit these requirements: they explicitly frame thinking about social standing and self-concept[1]. Moreover – as with perceptions of privacy and security – prospects may change over time as one's view of the system evolves. Lastly, our findings also indicate that the utility theory is insufficient for explaining the identified participants' relationships to privacy and security. This is in part why we adopted the notion of prospects from behavioural economics, as Subsection 6.1.2 examines further.

[1]Even though Rogers et al. [1959] do not include or discuss the concept of identity in the model, which we referred to while discussing the concept of self-concept in Subsection 5.5.1, the two concepts are related. For instance, Oyserman et al. [2012] introduce identities as composed of one's traits, characteristics, and group relations and memberships and then claim that "identities make up one's *self-concept* – variously described as what comes to mind when one thinks of one's personality". However, in our framing of prospects of using AI systems, we separate considerations in relation to one's social standing separately from self-concept.

The cited prior work on privacy, security, and the socio-technical embedding of the two, alongside the identified participants' statements on trust, thus situate prospects as a bridge between system acceptance and users' trust on the one hand and the system's underlying privacy-preserving and security capability (as defined by Boeckl and Lefkovitz [2020]) revealed through system interaction.

### 6.1.2 Properties

Having established the three dimensions, or axes, of prospects connecting human-IPA interaction and trust, we can now consider the properties of prospects. As established previously, we adopt the concept from behavioural economics, specifically from the Prospect Theory as defined by Kahneman and Tversky [1979] and later popularised by Kahneman [2011] in his book "Thinking, Fast and Slow". Prospect Theory formalises decision-making under uncertainty by modelling how people evaluate *changes* relative to a reference point (rather than absolute outcomes as in the classical expected utility theory) and how they trade off potential gains against potential losses [Kahneman, 2011].

In this framework, the overall value $V$ of a prospect is not determined by the final state of wealth (or, in our case, the final outcome in relation to one of the three axes), but by the sum of value $v(x)$ of outcomes $x$ weighted by a decision weight $\pi(p)$ associated with the probability $p$ of that outcome occurring. Formulary, that is:

$$V = \sum \pi(p_i)v(x_i) \tag{6.1}$$

where $x_i$ denotes an outcome (a gain or loss relative to a reference point), $p_i$ its probability, $\pi$ a decision-weighting function (distinct from objective probability), and $v$ the value function. As elaborated by Tversky and Kahneman [1991], the value function is an asymmetric S-shaped continuous function that is concave above the reference point and convex beneath it. As we will discuss below, these mathematical properties capture diminishing gains sensitivity and loss aversion. One formulation of the value function, proposed by Tversky and Kahneman [1992] is as follows:

$$v(x) = \begin{cases} x^{\alpha}, & x \geq 0, \\ -\lambda(-x)^{\beta}, & x < 0, \end{cases} \tag{6.2}$$

where $\alpha, \beta \in (0, 1)$ capture diminishing sensitivity and $\lambda > 1$ captures loss aversion. While these parameters differ between subjects, Tversky and Kahneman [1992] reported median exponents from their data to be 0.88, while the $\lambda$ indicated strong loss aversion at 2.25. A

value function with these parameters is shown in Figure 6.2.

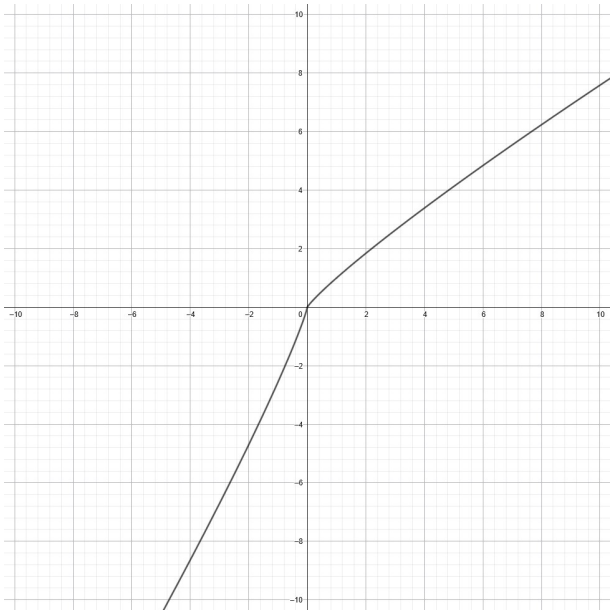


Figure 6.2: A value function, as described by Tversky and Kahneman [1992], which is concave for gains above the reference point (here, but not in general, at 0), convex for losses (here, $\alpha = \beta = 0.88$), and steeper for losses than gains (here by factor $\lambda = 2.25$)

While Prospect Theory originates in behavioural economics, our thematic analysis suggests that several of its core properties map onto how participants evaluated AI assistance for time management: participants framed prospective use less as an optimisation problem and more as a *risk-sensitive* trade-off between expected gains (reduced effort, improved organisation) and potential losses (errors, social consequences, diminished autonomy), with the latter routinely dominating their reasoning. Below, we argue that key points properties originally uncovered in the domain of wealth also hold for the three dimensions in human-IPA interaction: time and effort, self-concept, and social standing.

### Reference Dependence and Status-Quo Anchoring

Prospect Theory assumes that outcomes are evaluated relative to a reference point [Kahneman, 2011]. In our study, participants implicitly treated their existing routines, tools, and interaction norms as such a reference: judgements about SIPA4TM (and envisioned future IPAs) were frequently made by comparison to established practices (manual calendaring, e-mail coordination, commercial calendars, work routine differences when working in the office and when travelling, current LLM use cases, and current voice assistant capabilities). This reference dependence was evident in how "helpfulness" was framed: assistance was valuable when it reduced friction relative to the baseline, but quickly became unattractive if it increased interactional overhead, threatened one's self-view, or endangered one's social standing.

For example, [G1P2] characterised commercial voice assistants as inferior to the reference point of typing, in reference to the prospect of time and effort: "I would find them to be a slower and less effective user experience than just typing or using my phone as I usually would". Relatedly, when assistants produced additional steps or uncertainty, the perceived prospect shifted sharply from gain to loss, as can be read from the statement by [G5P2]: "it became that, that level of frustration where it's now easier to take your phone out to look up for the information". In [G2P1]'s interaction with SIPA4TM, this property manifested as rapid reversion to manual control when the assistant did not immediately outperform the baseline of doing the task without reference to suggestions: "in the first few minutes, like once I started doing the task, I found that the assistant wasn't actually really helpful, so I just, like, inherently just avoided using it". Such statements illustrate that participants did not evaluate the assistant in isolation, but against a stable reference of "what I already do" and that the threshold for perceived improvement was shaped by that baseline.

### Concave for Gains: Diminishing Sensitivity Towards Utility

A further property of Prospect Theory is diminishing sensitivity: as one moves away from the reference point, marginal changes have decreasing impact [Kahneman, 2011]. Tversky and Kahneman [1992] represent this mathematically using the parameter $\alpha \in (0, 1)$ in the value function and the concavity of the function above the reference point. Kahneman [2011] provides an intuitive example of this point: the subjective difference between winning $900 and winning $1000 feels much smaller than that between winning $100 and winning $200, even though the difference in absolute value is the same.

In our thematic analysis, an analogous pattern appeared in how participants discussed the value of improvements in explanation, transparency, and interface support. Many accounts suggested that participants wanted *enough* support to keep them confident and safe, but not exhaustive transparency. This was captured by the repeated stance that scrutiny is conditional and diagnostic, as illustrated by [G4P2] "if it ain't broke, don't fix it. If it's broken, then take a look" and [G2P1]: "As long as it does what I want it to do, then I'd be happy to trust it, yeah". In other words, once the system is perceived to be operating within an acceptable band of correctness, additional improvements enabled by scrutability yield limited marginal value. Conversely, when behaviour crosses a threshold into wrong or potentially incomplete, demand for explanation and correction increases sharply, as both the focus groups (cf. the "Integrated Scrutability and Control" point in Subsection 5.5.2) and the thinking-aloud analysis (e.g. "I'm just going to go and check and see if there's any more assistant suggestions for any of the days where there would be clashes where I might have missed them before." [G1P2]) indicate.

This threshold-like dynamic within the frame of diminishing returns also helps interpret why

participants often favoured lightweight cues (e.g., conflicts highlighted, assumptions flagged) over detailed justificatory dialogues. The goal was not maximal understanding, but sufficient confidence to proceed without incurring additional time and cognitive cost. This would fit well with the log-based findings from Section 5.4, which indicate a geometric decline in views from one level of the explanation to the next. It also aligns with the finding from the selected statements in thinking-aloud analysis from Subsection 5.3.3 that participants typically spent very little time analysing graphs: the time invested scrutinising assistant through them gave them marginally more compared to just reading the first level of explanation, whereas the first level of explanation (or just interpreting through implicit scrutiny) was perceived as likely higher in informational gain.

### Convex for Losses: Resignation and Reduced Marginal Pain

Conversely, Prospect Theory posits that individuals are risk-seeking when facing losses. So, a difference in prospects further from the reference point is valued less than the same difference in prospects closer to the reference point in the negative direction as well. Tversky and Kahneman [1992] represent this mathematically using the parameter $\beta \in (0, 1)$ in the value function and the convexity of the function beneath the reference point. In his book, Kahneman [2011] illustrates this by showing that people are likelier to take bad bets (90% to lose $1000 and 10% to lose nothing) than a sure loss with a higher utility function (e.g. 100% to lose $800) because the extra pain of a larger loss is psychologically smaller than the hope of avoiding the loss entirely.

In the context of our study, this property helps explain the "resignation" Data Use strategy observed in Subsection 5.5.2. When participants perceived that their privacy was already "lost", they exhibited a surprising willingness to "gamble" with further data exposure to access advanced features. [G7P1] captured this sentiment explicitly: "if Google has it, then ChatGPT has it /.../ we lost it, it's too late, I think". In other words, once the reference point is set in the domain of losses by assuming privacy is already gone, the user becomes less sensitive to incremental privacy intrusions. Consequently, they are willing to engage in the "risky" behaviour of using a privacy-intrusive service because the potential utility gain is perceived as the only way to recover value from a situation where the privacy cost is already sunk. This "bad gamble" is evident in the statement by [G3P2]: "it's just a matter of fact that you are not liking it, but irrespective, your data is being shared with everyone. So as far as this particular condition or the particular timetable goes, yes, I wouldn't be that comfortable to share the personal data. But to be honest, if I want to make it better, I don't have much of a choice. I really need to put in some suggestions in place".

### Loss Aversion and Negativity Dominance

In their foundational work, Tversky and Kahneman [1991] emphasised that a central property of Prospect Theory is that "losses loom larger than gains". This characteristic is represented mathematically through the factor $\lambda$ by Tversky and Kahneman [1991], through which the slope of the loss function is steeper for losses.

We can identify this property in how participants spoke about the consequences of errors and the allocation of responsibility. Across focus groups, the harm of a single mistake was repeatedly positioned as more detrimental to their prospects than the potential benefits of using IPAs. This was stated rather directly in [G4P2]'s reasoning about LLM outputs: "it can be wrong, and if it's wrong, I'm the one who will suffer. And then there is no suffering for ChatGPT". Another example of where the potential negative prospects towards social standing (stemming from a loss of a shared informational context with a friend) overshadow positive potential is given by [G2P1], while negating they would ever use an IPA do delegate communication: "Me personally, no, because there's many factors to this. But for example, my friend would be like: 'Hey, remember, we talked about this' – or something. I'm like, I did not talk to you about this".

In the time management context, the same asymmetry surfaced as a strong preference for retaining veto power, particularly for socially consequential actions (e.g., rescheduling meetings, changing commitments). Even participants who were broadly positive about assistance rejected the prospect of high-impact system actions without confirmation, exemplified by the recurring fear of waking up to autonomous changes: "wake up the next day, and my calendar's all changed on its own" [G2P2]. This aligns with the broader pattern that participants rarely framed adoption as "how much could I gain if this works?", but rather as "how badly could this go wrong if it fails?", consistent with the risk-driven scrutiny and approval thresholds reported in the Scrutability and Control subtheme of Subsection 5.5.2.

### Probability Weighting and Edge-Case Dominance

Prospect Theory also posits that people do not treat probabilities linearly with indifference to the scale of the outcomes, but rather through varying weights ($\pi$) for different outcomes [Kahneman, 2011]. The intuition for this property that Kahneman and Tversky [1979] provides is that most people feel they would pay more for a bullet from a revolver before playing Russian roulette, if the purchase would reduce the number of bullets remaining from one to zero rather than from four to three (out of six). As Kahneman [2011] summarised, people overestimate the probability of unlikely events and overweight unlikely events in their decisions.

In our study, these overvalued cases were typically the *edge cases*, i.e., situations in which

scheduling decisions entail atypical constraints, social risks, or contextual complexity. These edge cases were then seen as decisive in their prospects, thereby significantly affecting their trust and willingness to delegate. This was particularly prominent in discussions of whether an IPA could cope with non-routine weeks and context changes, where [G4P1] noted that being away from the usual work routine is "fundamentally different to the other weeks", raising doubt that an assistant could reliably generalise from routine data. Similarly, the prospect of a single socially costly misstep (e.g., inappropriate rescheduling, miscommunication with others) was sufficient for some participants to reject autonomy even if they acknowledged frequent low-stakes benefits, as we discussed in more detail in Subsection 5.5.1. This aligns with the qualitative pattern that trust was conditional on the system not merely being correct "most of the time", but being robust in precisely those low-frequency, high-consequence situations where users anticipated disproportionate losses. In Prospect Theory terms, these edge cases function as overweighted negative outcomes, thereby shifting the subjective evaluation of adoption towards caution and explicit verification.

### 6.1.3 Related Concepts in Human-Computer Interaction

To the best of our knowledge, formally adopting *prospects* from behavioural economics as a bridge between human-AI interaction and the user's trust is novel in the domain of Human-Computer Interaction (HCI) and related areas, including explainable artificial intelligence, scrutability, human-agent interaction, and interaction design. While the word is commonly used in publications in these domains, it is utilised in a non-formal sense, to mean future outlook, expectation, or likely direction. Our formal introduction – as users' beliefs regarding likely outcomes in terms of time and effort, self-concept, and social standing – is, however, related to several well-established terms. We have already situated prospects in relation to concepts such as reliability, trust, security, privacy, which are in Subsection 6.1.1, whereas here we relate prospects to the concept of affordances, propensity, and scrutability, as well as the two-dimensional human-centred AI framework by Shneiderman [2020].

#### Affordances

First, prospects are conceptually downstream from *affordances*. Originally defined in ecological psychology by Gibson [1979] as the interaction between animals and their environment, the term was then generalised and popularised by Norman [1988] in his book "Psychology of Everyday Things". As analysed by Kaptelinin and Nardi [2012] in their overview of the use of the concept in HCI, Norman [1988] described affordances in relation to the perceived capability, i.e. "a chair affords ('is for') support and therefore affords sitting". Similarly, a button in a user interface affords clicking, a text field affords text entry, and so "[w]hen affordances are

taken advantage of, the user knows what to do just by looking" [Norman, 1988]. While noting the utility of the analogy between the original ecological source of the concept for insights in HCI, Kaptelinin and Nardi [2012] conclude their article by arguing that the theory of affordances cannot be extended beyond the original scope of Gibson [1979], as to meet the growing conceptual requirements of describing the human use of interactive systems while remaining faithful to the original idea.

We offer no assessment on the closing argument by Kaptelinin and Nardi [2012]. Still, we note that prospects do align with the framing of affordances by Norman [1988] and the consequent HCI research, by providing a further conceptual frame. Namely, both prospects and affordances are internal evaluations by the user of system capabilities. But while affordances emphasise how that evaluation then influences what the user sees as *possible*, prospects are defined in terms of estimates of the *likely outcomes* of using the system, in relation to time and effort, self-concept, and social standing.

### Propensity

Propensity to trust, often abbreviated as PTT, as Scholz et al. [2025] define it, is a stable individual tendency in perception and decision-making with regards to trust. Lee and See [2004] offer a similar description of propensity to trust as a personality trait regarding one's tendency to trust that shapes behaviour in a systemic manner. Both papers attribute the underpinnings of the concept to Rotter [1971, 1980]. Scholz et al. [2025] further distinguishes PTT towards other humans (PTT-H) and towards automated systems (PPT-A) and investigate whether the two psychological dispositions may be distinguished theoretically and based on empirical evidence. Their conclusion points to PTT-H as a dispositional trust variable reflecting a personality trait, of which PTT-A is a narrower subtract. In other words, they support the view that PPT-H builds on a personality foundation for PTT-A. PTT-A is developed from PPT-H from the sum of experiences with similar automation systems and is influenced by early interactions with entirely new technologies. It should also noted that Lee and See [2004] also point to studies that indicate that levels of social trust vary significantly across countries and professions, indicating the effect of long-term experiences shared among groups of people can also influence trust.

Prospects too influence one's trust towards an automated system, but they are not necessarily just an individual trait nor are they primarily static. In fact, propensity to trust may be influential for the original reference point of prospects (cf. Subsection 6.1.2 on reference dependence), but statements from our focus group discussions point to a continually updated final judgement for trust ("After a certain period of time, you would trust them more to make decisions on their own" [G2P3]), depending on system performance ("If it's consistently

making the right decisions, I would say I'm consistently less and less likely to look into how it's making those decisions" [G1P2]), changing social norms ("I guess you talk about also social and cultural acceptability more. People accepted Google Calendar invites. People do Teams calls pretty quickly. And I have had email threads where some tech CEO CCs in a bot and the bot coordinates the meeting" [G4P1]), and one's circumstance ("Do I trust the AI to be able to make those decisions from what that week looks like for me visiting on the industrial side?" [G4P1]).

### Human-Centred AI

Shneiderman [2020] formalises the *Human-Centred AI* framework as a two-dimensional space, whose axes represent levels of human control and computer automation. While in this section we focus on the framework from the article and the subsequent book, the use of the syntagm both predates [Rasmussen, 2007] and supersedes [Germanakos et al., 2025] the framework by Shneiderman [2022]. The central motivation for the framework is to explicitly move away from a one-dimensional framing stemming from Sheridan and Verplank [1978] in which the level of automation and the level of human control are seen as a trade-off, i.e. proposing that more system autonomy comes at the cost of decreasing explicit controllability. Instead, Shneiderman [2022] proposes that reliable, safe, and trustworthy AI is both high in computer automation and high in human control, in a manner that promotes human self-efficacy, mastery, and responsibility while utilising the rapid action and capabilities of AI systems.

Notably, in his motivating critique of the one-dimensional framing of autonomy, Shneiderman [2020] points to studies indicating that more autonomy may still indicate an increase in human effort and time, due to the need to monitor autonomous systems stemming from a lack of reliability and the consequent lack of trust. Furthermore, the article concludes by highlighting human responsibility as a mechanism for both improved performance (due to feedback, innovative use, etc.) and higher reliability. In contrast, the conclusion warns against automation that does not support user understanding of the consequences of its actions and does not support reversing its actions, ultimately leading to the de-skilling effects that undermine humans. These points are supported by the findings of our user study, including emphasis that the user, not the system is the one that "suffers" by [G4P2]; noting that if the system does not fit the actual needs of the user it would reach a threshold of dissatisfaction leading to system abandonment due to an actual increase in effort and time by [G5P2], and the concern about overreliance on automated decision-making and reasoning by [G6P1].

Shneiderman [2020] also calls for research into the development of measures for appropriate levels of control and autonomy. We posit that prospects are a conceptual tool enabling such deliberations. They provide grounding for discussions on concerns and hopes of automation

with regard to time and effort, as well as their social consequences, allowing us to capture and quantify potentially diverging levels of actionability and accountability. Moreover, they explicitly allow for framing of the value of AI in terms of the user's self-concept, supporting the arguments on de-skilling, creativity, and work preference.

#### Scrutability

Finally, prospects help situate *scrutability within broader interaction design*. As we have described in our overview of stances towards explanations in Subsection 5.5.2, participants rejected explanations that were redundant, non-actionable, or unreliable, and instead valued explanations that were tightly coupled to corrective control and moment-to-moment decision needs. This motivates treating scrutability as a tool for aligning user prospects with the system's capabilities, rather than as an architectural goal with inherent value for the user. We build on this interpretation in the subsequent section by connecting the paradigm of prospect to our principles and design guidelines for scrutable time management IPAs. The guidelines also convey when nudging the user to scrutinise the system is most likely to be beneficial for prospect calibration and how the IPA should actively support the user in the process.

## 6.2 Principles of Scrutability

Our research question was to examine how principles of scrutability, as they were summarised by Kay and Kummerfeld [2013], can be adapted to the envisioned proactive IPAs. The dissertation which originated the concept "is based upon the premise that a user should be able to scrutinise such user models" [Kay, 1999]. This premise is then exemplified in the dissertation through *sam*, an adaptive text editor with a scrutable user model. Following the same premise, Czarkowski [2006] created the *ATML* framework for a tutorial in a markup language and *SASY*, a web application for adaptive content, both of which use scrutable user models.

What these examples and the premise behind them imply is a tight coupling between the concept of scrutability and a specific approach to realising artificial intelligence. This approach to user modelling is based on formal logics for representation and reasoning, which, as Pohl [1999] described, has three potentially partially overlapping stages: the gathering of usage and user data, knowledge inference from the data, and the user model representation. The data in the user model is either directly entered by the user, captured indirectly from the user's interactions, or deduced through a set of rule-based inferences. As surveyed by Brusilovsky [2001], this approach to user model data representation and inference was paradigmatic in adaptive hypermedia (the research domain at the intersection of user modelling and hypertext) of the 1990s and early 2000s. The ambition for such adaptive systems was to support users

in activities such as learning, managing personal views, and information retrieval.

However personalised systems nowadays have moved beyond the traditional approach to user modelling, as described by Herder and Masthoff [2025]. Instead, personalisation is realised through large-scale black-box systems trained to maximise a target such as user clicks. Latent patterns learnt about the user, the target, and other relevant relations then serve as implicit representations of user interest. This shift has enabled a more granular personalisation that no longer requires the use of stereotypes or other user groups and explicit modelling rules [Herder and Masthoff, 2025]. But, as Hanif et al. [2023] surveyed, it has also prompted concerns regarding lower levels of transparency and interpretability, in turn restricting deployment in vulnerable domains.

As we discussed in Subsection 5.4, participants often tried to learn about the underlying algorithm by probing its functioning, rather than by explicitly exploring the provided explanations that reveal the system's rule-based behaviour. The theme on trust in Subsection 5.5.3 then revealed that being able to fully understand how the system's reasoning was neither expected nor – as we have discovered – always desired. This sentiment was repeated in various terms in multiple groups from participants of varying demographics, including by [G2P1] ("As long as it does what I want it to do, then I'd be happy to trust it") and [G2P2] ("I guess it depends how good it would be, and you only get that by learning through doing").

With this underlying technological shift from explicit user modelling to black-box implicit inferences and the consequently evolving user expectations, we suggest that *scrutiny should be deliberated at the level of an AI system, rather than at the level of the user model*. Moreover, we propose that *the utility of enabling user scrutiny need not be considered a goal in its own right, but rather a capacity whose role and importance stems from the paradigm of prospects*.

In the subsection below, we outline our proposed set of six new scrutability principles. As Table 6.1 indicates, four of the principles build upon the four "overall principles" from Table III in the summative paper by Kay and Kummerfeld [2013]. In addition, Integrated Overlays corresponds most closely to the principle of evidence-based architectures, which Kay and Kummerfeld [2013] discussed in the context of realising the broader principle of parallel design. The User Model Level, while not explicitly defined as a principle, was articulated as a leading premise in the paper's introduction. To remain faithful to the source publication, the asterisks in Table 6.1 indicate the distinct conceptual grounding of the two marked principles. Unlike the remaining four, which were formulated as such by Kay and Kummerfeld [2013], the principle of *User-Model Level* is inferred by us (as argued at the beginning of this section), whereas *Evidence-Based* was introduced in the section on "*how* to create scrutable user models" [Kay and Kummerfeld, 2013].

Table 6.1: Overview of principles that served as the starting point and our proposed principles for proactive assistants

| Principles for Scrutable User Models by Kay and Kummerfeld [2013] | | Prospects-Based Principles for Scrutable Proactive IPAs | |
|---|---|---|---|
| **User Model Level*** | Scrutable user models as the foundation | **System Level** | Scrutability realised at the system level |
| **Parallel Design** | Design of representation and interfaces in parallel | **Prospects-Grounded Design** | Prospects as the grounding for the interaction design process |
| **Evidence-Based Architecture*** | Evidence from the allowed sources accretes the user model | **Non-prescriptiveness** | Scrutability as a capacity, not an architectural restriction |
| **First -Class Citizen** | User models exist outside a single application | **Integrated Overlays** | Integrated interaction with a layered system |
| **Context-Based Interpretation** | Each application and context may interpret the user model differently | **Uncertainty Signalling** | Uncertainty of system reasoning is signalled and scrutable |
| **Client-Side** | Personal digital infrastructures redefine the notion of client-side | **Model Transferability** | Facilitating customisable knowledge extraction |

### 6.2.1 System Level

Kay and Kummerfeld [2013] preface the concepts of scrutability in their summative paper by first describing the problems[2] that may be resolved by making the user models powering the personalisation scrutable. In later talks, Kay [2021b] adopted the syntagm "scrutable systems" but maintained that such systems should be "designed, from their very foundations, so that people can scrutinise them to understand and control how their personal data is captured and transformed into a user model". This bottom-up approach, which requires that systems be designed to follow a specific user model architecture or to, at least, be designed "from their very foundations" [Kay, 2021b] for scrutability, is what we see as the paradigm of prescriptive, user-model-level scrutability.

In contrast, we propose top-down, systems-level scrutability. The analysis of our user study in Chapter 5 clearly revealed that our participants overwhelmingly do not want to prioritise complete understanding and controllability over more direct benefits – in terms of time and effort reduction, self-concept actualisation, and social standing improvements – that a more capable AI system would bring.

For instance, recall the [G2P1]'s matter-of-factly answer on the question whether they would be

[2]As reported on in Section 2.3, these problems are: invisibility of personalisation; errors in user model; wasted user models; privacy; control; and others such as guidance. We discuss this further in Section 7.1.1.

comfortable with providing an AI system information about themselves and if the explanations would make them trust the system more: "As long as it does what I want it to do, then I'd be happy to trust it". Our study also suggests that users often treat scrutability as diagnostic rather than foundational. In other words, the dominant motivation was not an enduring interest in the underlying representation and inference, but a need to resolve moments of surprise, friction, or perceived error. This stance is exemplified by [G4P1]'s remark that: "I just want to get that out of my day, I don't necessarily need to go through the nitty-gritty of it". Similarly, [G1P1] framed their own interest in deeper inspection as contingent on failure: "I'd mostly look into what's happening under the hood to figure out if it was going wrong". This interactional logic is important because it implies that scrutability features that assume a stable, curiosity-driven inspection workflow (e.g., prolonged exploration of evidence provenance, inference rules, and model state) risk missing the moments when scrutiny is actually triggered. It also suggests that the practical requirement is not merely to expose internal state, but to support quick sense-making and actionable correction in the flow of interaction [Kulesza et al., 2012, 2015].

In that sense, our findings align more with the human-centred framework by Shneiderman [2022], whereby the scrutiny control is desired only in reference to a goal, such as more accurate personalisation, consequences containment, and error correction. The level of desired automation is also highly dependent on task at hand and broader context (typically high for straight-forward calendar scheduling, and low for communication with other parties in private capacity). As such, our principles see *scrutability as a conditionally desired capability* of advanced systems, rather than a guiding premise, especially not at a cost of curtailing user prospects.

### 6.2.2 Prospects-Grounded Design

The first explicit principle of scrutability proposed by Kay and Kummerfeld [2013] is *parallel design*. It postulated that the interface issues and user model representation should be considered hand-in-hand, from the very beginning of the design process. This principle is justified by the premise that the user should be able to scrutinise the model's reasoning to a "high degree of fidelity", by seeing "the actual details of the system reasoning and of the user model" [Kay and Kummerfeld, 2013]. Notably, all publications cited as exemplifying the principle [Glass et al., 2008, Koidl and Conlan, 2008, Core et al., 2006] were expert systems or rule-based AI.

With the prospects-grounded design, we acknowledge that most capable AI systems nowadays rely on systems that are based on opaque embedding architecture, whose conceptual mapping may not necessarily overlap with that of humans [Bommasani et al., 2022, Hewitt et al., 2025].

In other words, requesting that the actual details of the system reasoning and the user model be made available to the user may be neither possible nor meaningful.

Thus, with prospects-grounded design, we want to emphasise one of the principal findings from our user study: scrutinising the AI system is desirable only when it leads to a clarification or improvement of prospects. Only when the user believes that scrutinising or explicitly changing the reasoning and modelling of the IPA (however it may be realised) are they willing to invest their effort and time in doing so and to – in the process – risk potentially changing the system in a way they had not anticipated or wanted.

Practically, this would mean that, when considering IPAs, the designers should consider how their contributions towards (a feature of) the system would likely reflect on user prospects. This should also be done holistically at the level of the system, by considering both how the specific contribution would likely influence the prospects as well as how the overall interaction design and the IPA as a hole may contribute. All elements of the interface (be they dialogue-based or visual) should convey the potency of the underlying model to an extent that supports the user's prospect calibration. Interaction design should include pathways to scrutinise the individual features of the model and the interface to the highest degree practically feasible, while remaining faithful.

Thus, this principle proposes that designers and developers ground their work in modelling, not on making the user model inherently interpretable, but on making their IPA as useful to users for saving time and effort, helping them actualise their self-concept, or aiding them in improving their social standing. Scrutability, from this angle, is not a goal per se, but a capability that allows users to better understand how the IPA would help them and how – and to what extent – they may steer their IPA to have it understand them better.

### 6.2.3 Non-Prescriptiveness

Kay and Kummerfeld [2013, pp. 23–32] prescribe the approach to evidence representation, reasoning, and selection that user models ought to follow to be scrutable. They also argue for a high-level architecture for evidence-based user modelling. This approach is based on the idea of accretion-resolution by Kay [1999], whereby evidence from allowed evidence sources accretes the user model. When a value from the user model is needed in an application, the value is resolved by interpreting the model. From this general idea, Kay and Kummerfeld [2013] then describe more specific design guidelines for evidence representation that fits their accretion-resolution idea. According to the guidelines, each piece of evidence should record the timestamp of the source, appropriate data type for different sources, and evidence type: 1) explicit ground inference (explicit user-provided answers to questions about their preference or knowledge); 2) implicit ground inferences (from logging, monitoring, and automatic observa-

tion); 3) ex-machine ground inferences (tracking previous application responses to the user, to avoid repetition); 4) inferred from stereotypes; and inferred from model-specific inference rules. Furthermore, Kay and Kummerfeld [2013] propose principles for data representation, which include maintaining external evidence in original form for accountability, late binding of evidence for a need-only interpretation, and explainability of the meaning of modelled evidence.

These principles and guidelines were meaningful in a time when knowledge-based systems and logic-based AI were a dominant implementation substrate for personalised systems. However, contemporary data-driven user modelling and, in particular, foundation-model-based assistance does not typically expose user models in a form that cleanly supports the evidence bookkeeping that Kay and Kummerfeld [2013] prescribe. Specifically, the shift from symbolic to sub-symbolic, connectionist architectures (e.g., deep neural networks and Large Language Models) renders the requirement for explicit, atomised evidence tracking architecturally dissonant. As Goodfellow et al. [2016] detail, such models store knowledge not as discrete, retrievable records ("explicit ground inference"), but as distributed representations encoded within high-dimensional vector spaces. Consequently, the concept of "late binding" of evidence is technically incompatible with architectures where input data is immediately transformed into latent embeddings [Vaswani et al., 2017, contrast the architecture of the transformer in relation to the accretes-idea]. Furthermore, Lipton [2018] argues that such models lack "decomposability", i.e. the ability to map individual model parameters (weights) to meaningful human concepts or specific pieces of evidence. Therefore, enforcing the strict data typology prescribed by Kay and Kummerfeld [2013] upon a stochastic model would necessitate the creation of a post-hoc proxy model, which Rudin [2019] warns often leads to unfaithful explanations that describe a simplified abstraction rather than the actual, complex decision-making mechanism. Relatedly, interpretability research has repeatedly underlined a tension between *plausible* explanations and *faithful* ones: explanation interfaces can successfully communicate a coherent rationale to the user while still failing to reflect the actual causal determinants of the model's output [Jacovi and Goldberg, 2020, Rudin, 2019]. In other words, even where an interface can provide something that *looks* like accretion–resolution (e.g., listing "reasons"), this is not guaranteed to be an accurate representation of what the underlying stochastic model encoded or relied upon.

From the user perspective, the participants in our study have not shown a principal interest in the underlying knowledge representation and reasoning per se. For instance, logs indicate only brief glances at the visual explanations, while the comments in the discussion repeatedly emphasise that looking at the underlying models and explanations was principally utilitarian ("people don't care" [G4P1]). We also saw indications of awareness of unfaithfulness of explanations, for instance by [G6P3]: "I also don't know if I would believe that like what they have there is like all that".

With this in mind, we propose the principle of *non-prescriptiveness*, that is, the idea that the system's scrutability should not be conditional on, nor should it prescribe, the underlying user model type. We posit that scrutability should be a system-level capacity that admits the underlying model and the user into improving the user's prospects of using the IPA – whatever the underlying model ought to be. Of course, a consequence of this approach is that the scrutiny process might not yield exhaustive insight into the IPA's knowledge or a definitive understanding of the underlying reasoning. But such a non-prescriptive view of scrutability allows for continued emphasis on user involvement in understanding their IPA through faithful system responses, which is no longer conditional on technological assumptions.

### 6.2.4 Integrated Overlays

Relatedly, Kay and Kummerfeld [2013] also describes the idea of user models as first-class citizens, wherein user models exist separately from applications. These separate user models would then be accessed, reinterpreted, and modified by applications in their respective contexts. The user could then scrutinise their separate user model directly or through application context. In this sense, Kay and Kummerfeld [2013] also describe "interface principles" such as Ontology, stating that the interface should be able to explain each modelled aspect and the Outflow Process, enabling the user to determine what and how each aspect of their model is being used in a given application and context.

In our first envisioned diagram in Figure 3.1 in Subsection 3.1.5, we saw scrutability as a conceptually separate channel of interaction. This conceptual separation was indeed reified in applications cited by Kay and Kummerfeld [2013], as well as later by us in SIPA4TM. Practically, this separation meant that the user had to abandon their primary interaction to scrutinise the underlying model and steer the resulting personalisation, potentially changing the contents of the application interface in the process.

Our user study, both through the analysis of the discussions and of the thinking-aloud interactions with SIPA4TM, has strongly indicated that this approach does not fit user expectations. Instead, our study evaluation indicates a strongly goal-oriented use and expectations of IPAs as well as low tolerance for disruptions of the primary interaction flow ("I guess it depends on how good it would be, and you only get that by learning through doing" [G5P2]). Similarly, [G2P2] stated they explicitly defaulted to figuring our the assistant's behaviour through probing rather than clicking on the explain button: "I guess my first thought is to just to try to figure it out myself".

Therefore, we propose that the scrutiny be seen as conceptually integral to the same (and only) interaction flow that users are in for their main purpose of using the assistant. Of course, should a further exploration be desired, for instance, to gain a fuller understanding

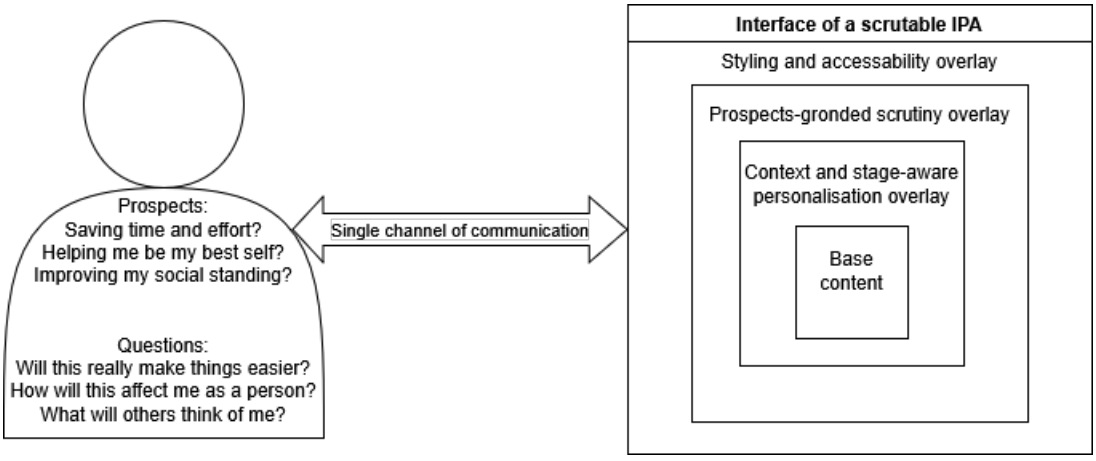


Figure 6.3: A conceptual diagram of the proposed principles

of the system, allowing the scrutiny to take over the whole of the user interface should be permissible, if meaningful for the amount of content. But such changes should come at deeper levels of exploration only, not at the first step and not as the paradigm of how the user interacts with the system. This idea is illustrated in Figure 6.3, which emphasises the single channel of communication and the grounding of interaction design in prospects.

With the overlays (the concentric squares) on the right-hand side of the diagram, we offer only conceptual guidance for how to think about the content displayed to the user, rather than an architectural ordering or requirements (in line with the previous principle of non-prescriptiveness). The overlays are thus symbolic and serve as references in thinking about the responsibilities of individual elements of interface and interaction design: the base content comes from an information retrieval engine, and is then personalised to the user's interest and preferences and adapted to the current context. Scrutability at a systems level is then there to allow probing of the underlying models, with uncertainty in mind. Crucially, for the user, this is presented in a unified manner, with conceptual delayering available only on request – there is no separate channel for scrutiny or user control. Lastly, as we saw in our own user study, the success of both user scrutiny and task completion depends as appropriate support given the user's accessibility needs.

For a concrete example, consider how Czarkowski [2006, p. 207] used a highlighting tool, implemented as part of his scrutable learning system. The tool signals to the user what parts of the content are personalised and offers an additional explicit explanation for how such personalisation was achieved. We would see this as a partial fulfilment of the principle, as it overlays the scrutiny on top of the personalisation layer applied to the primary content. Nevertheless, such a design still requires the user to switch on the highlighting explicitly, which toggles a new interaction mode. Complete fulfilment of the principle of Integrated Overlays would include explanation and highlighting as part of the main interaction channel and would

– in line with the role of Proactive Support we will discuss in subsection 6.3.1 – anticipate when the user would benefit from the tool and propose its use.

### 6.2.5 Uncertainty Signalling

Given the non-perscriptiveness approach to scrutability, we ought to acknowledge the potentially stochastic nature of underlying user (and context) modelling, as well as the uncertainty on data representation and reasoning. A scrutable IPA should let its user scrutinise the uncertainty of its beliefs, actions, and consequences and all in explicit reference to prospects. Moreover, even without explicit user scrutiny, systems should *signal their uncertainty*, thus fostering user explanations and inviting them to investigate further, as necessary.

This principle can be seen as an extension of the premise by Kay and Kummerfeld [2013] that different applications may interpret the user model differently, depending on context. This principle also states that uncertainty should be seen as part of that context and presented appropriately.

Of course, as we discussed in the context of agentic systems in Subsection 2.5, actually understanding the level of uncertainty and the meaning behind specific representations may be very hard. However, our empirical findings suggest that making the user aware of the best estimate of system uncertainty in appropriate situations remains desirable. Moreover, we propose that a scrutable IPA should also express the level of certainty about its understanding of the context as well (given that the actual use of contextual data might be uncertain as well, as Liu et al. [2024] explores in the case of LLMs).

### 6.2.6 Cross-Model Transferability

With the principle of "client side", Kay and Kummerfeld [2013] stated that the user model should ideally be maintained within the user's personal digital ecosystem rather than confined to the server-side logic of a specific application. The motivation for such decentralisation is to empower the users to inspect and modify their profiles through a single, consistent interface, which then serves as a unified source of truth for various scrutable applications. This vision prevents "wasting" the user model's data and helps curtail privacy concerns. Notably, "client-side" is not necessarily about local storage of the user model, but rather emphasises the idea of ownership and user control: "the shift needed from current practice is that the user should have access to and control over his user model, regardless of where it is stored" [Kay and Kummerfeld, 2013].

However, the contemporary landscape of proactive IPAs is characterised by heterogeneous providers, heterogeneous modelling pipelines, and heterogeneous incentives [de Barcelos Silva

et al., 2020, Chkirbene et al., 2024]. Under these conditions, insisting on a single, shared, evidence-based representation or on a uniform model architecture that would enable full "first-class citizen" reuse across applications seems less likely than at the time of the proposal. This motivates a reframing of the original client-side aspiration into a more pragmatic requirement: the user should be able to *extract* (and selectively re-use) the personal data and inferred preferences that have accrued through interaction, even if the receiving assistant represents them differently. In other words, rather than necessitating a shared architectural framework, IPAs should, at a minimum, enable the extraction of collected user knowledge and preferences in a format that permits transfer between disparate providers.

Finally, this emphasis has an analogue in regulation: the GDPR explicitly articulates a right to data portability, i.e., to receive personal data "in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided" [GDPR, 2016, Art. 20]. While the GDPR targets *personal data* rather than inferred user-model constructs per se, it nevertheless establishes portability as a normative baseline for reducing platform lock-in [Syrmoudis et al., 2021]. For scrutable proactive IPAs, we therefore treat cross-model transferability as an applied essence of client-side control: users should be able to retrieve the personal information and preference structures that shape assistance, inspect them, and carry them forward, even when the receiving system must re-encode them into a different internal model.

## 6.3 Roles of Scrutability

Kay and Kummerfeld [2013, pp. 21–24] also describe roles for their envisioned scrutable user models. There are three categories of roles: 1) control of the modelling processes; 2) validity; and 3) re-purposing. In this section, we will revisit each group and propose a reframing of these roles that embodies the principles of system-level prospect-grounded proactive IPAs. A high-level overview of the two sets of roles is shown in Table 6.2. Similarly to Kay and Kummerfeld [2013], who split each of the three roles into their comprising aspects, we divide each role into design guidelines that build on ideas from our proposed principles and empirical findings towards realising scrutable IPAs.

### 6.3.1 Proactive Support

The overarching motivation for the role of re-purposing for Kay and Kummerfeld [2013] is to capture more data users provide as they interact with various applications. The captured data is then to be stored in the *first-class* user model for later *re-purposing* in different applications,

Table 6.2: Contrasting overview of roles and their main aspects or design guidelines, for scrutable user models and for scrutable IPAs

| Roles for Scrutable User Models by Kay and Kummerfeld [2013] | | Roles for Proactive IPAs | |
|---|---|---|---|
| **Re-Purposing** | Navigation; Influence; Learning; Usability; Programmer accountability | **Proactive Support** | Cues Recognition; Apt Explanations; Empowering User Control |
| **Control of the Modelling Process** | Inflow; Reasoning; Use; Reuse | **Meaningful Constraints** | Domain Attunement; Potency Stages |
| **Validity** | Correctness | **Prospects Alignment** | Current Prospects Modelling; Prospects-Capability Alignment; Deliberate Onboarding |

that is, for reinterpretation of the data in the context of another personalised application. The basis for the role is the problem of *wasted user data*, which we discuss in more detail in Subsection 7.1.1.

There are five aspects of the role of re-purposing that Kay and Kummerfeld [2013] detail individually: 1) navigation (using the user model to view larger information spaces); 2) influence (by seeing explanations, users are more likely to accept the reasoning and advice); 3) learning (user models may serve as a mirror and thus aid learning); and 4) programmer accountability (achieved by keeping the prospective user in the creation of user models that will then be emphasised).

With our role of *proactive support*, we still principally maintain all benefits contained in the four aspects of the role, but propose that it is achieved at the level of interaction with the IPA, not the user model. The reframing builds on the principle of *System Level* in Subsection 6.2.1.

More specifically, with the role of proactive support, we posit that the scrutable IPA should extend its proactivity from task-based assistance, such as time management, to also supporting the user in scrutinising it. To do this, scrutable IPAs should be designed to recognise interaction cues which indicate the user might not understand something about the IPA and that it is a good time to offer proactive guidance. This guidance, when in the form of explanations about the IPA, its actions, or suggestions, should be appropriately presented, in line with the empirical findings. Lastly, the system should also proactively encourage the user to provide feedback and to control its future behaviour, in a manner appropriate to any algorithmic constraints.

### Cues Recognition

The guideline of *Cues Recognition* expresses an ambition for a scrutable IPA to track the user's actions and to recognise behavioural cues, such as diagnostic exploration or error-resolving attempts, that indicate opportune moments for proactively nudging the user towards scrutinising the system in a manner that would increase their understanding of the system and the accuracy of their prospect estimates. Our verbal protocol analysis in Subsection 5.3.3 revealed that the willingness to engage in scrutiny is highly conditional, often triggered only by system failure or unexpected behaviour. In general, the users in our study wanted to scrutinise the assistant when:

1. There is a clear affordance and a direct perceived utility in doing so (in terms of prospects); or
2. Their interaction is interrupted to nudge scrutiny (although this has found to be unreliable and frustrating); or
3. They attempt to correct repeated errors in the IPA's behaviour (unless they have already resulted in the prospects being so demolished that it leads to an abandonment of the system).

Other than in these situations (and when explicitly prompted to do so by a user study task), participants rarely approached scrutability as a detached activity or paused their main interaction flow to explicitly investigate the explanations or the user model.

With Scrutiny Cues recognition, the IPA would recognise what we termed *Implicit Scrutiny*, i.e. attempting to investigate or modify the IPA's behaviour through indirect means (cf. Subsection 5.3.3). Notably, Czarkowski [2006, p. 204] reported that their hypertext system for coursework learning recorded more scrutinisation due to "natural" motivators (e.g. using the course map to extrapolate how they are progressing) than the premeditated ones (the presence of jokes and hints expected to mildly annoy the users, so that they would scrutinise the user model to turn them off). This both aligns with our findings of "implicit scrutiny" being more dominant than the explicit channel through the explain buttons and indicates that actual cues may be unpredictable and something that IPA should also learn to recognise.

Once the cues are recognised, the IPA response should then actively support the user's scrutiny in the main interaction channels, not treated as a separate mode that requires shifting their control mode, in line with the principle of *Integrated Overlays*. Below, we elaborate more on the steps of *Proactive Support* that follow the *Cues Recognition*.

### Apt Explanations

The principal source on what constitutes *apt explanations* is our findings from the focus group discussions. In particular, the claims for this design guideline stem from the statements on "Inappropriate Explanations" and "Integrated Scrutiny and Control" within the broader subtheme of "Scrutany and Control" in Subsection 5.5.2. We supplement these findings with those from the log-based interaction analysis, elaborated on in Subsection 5.4, as well as with the related literature on explainable AI.

First, the participants expressed a desire for the IPA to provide "*just enough*" information, rather than an exhaustive justification that could be overwhelming. This follows from the logged interactions, where the depth of scrutiny degrades with geometric progression (roughly thrice as few explanation clicks for every level of explanation).

Furthermore, the discussions indicated users want explanations to be *problem-specific*. Such explanations provide actionable guidance or a clear handle for correcting system behaviour, but they were promptly dismissed as unhelpful. As several participants indicated, they were less interested in understanding how the system worked in the abstract than in knowing what could be done at that moment to improve a specific outcome. This aligns with prior HCI work, including by Shneiderman [2022] and Hanif et al. [2023], on actionable explanations, and motivates a design guideline whereby scrutability is embedded directly into the main interaction mechanism – coupled with suggestions, approvals, and reversible actions – rather than relegated to separate inspection views.

For a practical realisation of how this can be achieved, we can turn to *"ante hoc" explanations*, as advocated by Shneiderman [2022]. One example thereof are "sliders", or a different visual representation of the information about the system. SIPA4TM also actively used this approach for the visualisation of the current data permission stage, as detailed in Subsection 4.2.1.

Where that is not possible, it may be meaningful to provide *example-based post-hoc explanations*. As reported by Leake and Mcsherry [2005], even in case-based reasoning and expert systems, explanations based on examples may be more convincing for the end user than providing actual underlying rules, thereby also emphasising the importance of adapting the explanations themselves to the user's level of expertise. More recently, Lipton [2018] proposed using examples to illustrate not only model accuracy but the types of examples for which an AI system tends to be right or make mistakes, providing humans with an opportunity to estimate AI performance regarding their specific case.

*Counterfactual and semifactual explanations* thus offer one potential realisation of post-hoc explanations that fits what the study participants wanted. As explored by Gajcin [2025] in the context of reinforcement learning, counterfactual explanations offer alternatives which would

change the outcome, thus offering a pathway for correction. On the other hand, semifactual explanations indicate alternatives where the outcome would have remained the same.

Of course, as Lipton [2018], Shneiderman [2022], and others warn, post-hoc explanations may be informative in terms of expectation shaping, but they do not necessarily elucidate how the model actually works. For this reason, apt explanations must remain *faithful* to the underlying algorithm so as not to risk damaging rather than improving user trust, as we discuss in more detail with the problem of *Justification Suspicion* in Subsection 7.1.1. Next, we discuss how proactive support extends from IPAs explaining themselves to empower user control.

### Empowering User Control

The guideline of empowering user control encapsulates our finding that offering control mechanisms is not enough. Namely, while participants expressed a desire to retain control over their schedules, our findings indicate that user control over the IPA's suggestions was exercised selectively and with caution. As we will discuss further within the context of *Self-Efficacy Deficit* in Subsection 7.1.1, control was rarely treated as an end in itself; instead, participants engaged with control mechanisms primarily on a *needs-only basis*, when something appeared incorrect or misaligned with their intentions. Thus, to provide proactive support, scrutable IPAs should encourage users to use control as a means of continuous optimisation towards improving their prospects.

Our user study further indicates that *transparency of control is insufficient* for this empowerment. Such a sentiment was expressed by, for instance, [G3P2]: "Well, I wouldn't give it that much time because I think people are much better than me when it comes to that, to be honest with you. So I wouldn't devote much of my time into that, rather than I would focus on other things". [G3P1] agreed with [G3P2], adding that they "wouldn't put much effort into" into understanding how the IPA works to then correct its mistakes.

Furthermore, our findings suggest *increasing the granularity of control is not empowering* on its own, either. This was expressed clearly by [G4P1], with an emphasis on his concern of downstream consequences of potential changes: "I don't know if I have confidence enough in my ability to [control the user model], depending on what the model is. If it's just wake up and going to sleep, totally fine. But if it's like, there's more characteristics to it than that, [then] I don't want to change the ingredients. I don't know how that's going to impact how this thing operates".

The same participant then offered a proposal we use as a *basis for our guideline*: "I think it'd be interesting if there was a language model front face to it that I could talk to. I'd be like, 'I have this concern,' 'What is it?' 'How does this affect this?' 'It's like, oh', 'I didn't like that

change of meaning. Could I modify the characteristics?' " [G4P1].

Building from this idea, the design guideline of *Empowering User Control* expands the role of *Proactive Support* of the IPA to actively encourage the user to make changes that would increase their prospects of use. As [G4P1] suggested, this may be done through IPA by answering the user's concerns on downstream consequences of a change, as well as by *providing apt explanations* (as described above) regarding its actions and assumptions. And, crucially, this empowerment may be proactive, that is initiated by the IPA at opportune times (*cues recognition*) to make the proposed change or feedback more tangible.

Notably, this design guideline strengthens the coupling between user scrutiny and user control. Kay and Kummerfeld [2013] emphasised that users are more likely to control when they have actively investigated and understood how an application is personalised to them. Scrutable IPAs through proactive support further *empower user control* by actively supporting their investigation and emphasising how a certain change would affect future outcomes (in terms of specific changes or their prospects, as most appropriate). In Section 7.2, we discuss practical aspects of *proactive support* in the case of LLM-based agents.

### 6.3.2 Meaningful Constraints

Kay and Kummerfeld [2013] maintain that "if a user can scrutinize her own model to gain an understanding of it and the associated processes, this can be the starting point for controling these", i.e. they see scrutability as a prerequisite for user control. They state that through *control of the modelling process*, scrutable user models enable privacy as defined by Palen and Dourish [2003], that is to restrict the use of their personal data. Kay and Kummerfeld [2013, Table II] then break this role into the control of inflow (ground inferences allowed into model), reasoning (internal inferences within the model), use (availability of parts of the model to an application), and reuse (multiple applications using the model).

As conveyed by terms such as ground inferences, this breakdown of user control again indicates thinking about scrutability within the frame of representational and logic-based user modelling. In the view of scrutability that we put forward in this dissertation, we maintain scrutability as a prerequisite of user control. However, we do not propose seeing the role as fulfilled in relation to transparency of the four processes (in line with the principle of *Non-Prescriptiveness*). Instead, scrutable IPAs should both enable and entice their users to use existing control mechanisms when this is likely to improve prospects. To achieve this, we offer two guidelines on a non-algorithmic, conceptual level. The first, *Domain Attunement*, emphasises the need for the scrutible IPA to recognise and respect user-specific application domains, whereas the second, *Potency Stages*, allows constraining the IPA's capabilities to safeguard prospect-based constraints.

### Domain Attunement

Across both the interaction study and the focus group discussions, participants articulated a nuanced stance towards delegation and proactivity. While many welcomed the prospect of reduced effort through more capable assistance, this desire was consistently tempered by concerns about assistance not in line with one's values or of negative social consequences. The common theme was that the IPA should limit its proactive assistance or delegate communication in certain domains. Crucially, these domains are defined on the level of an individual user.

For instance, participants were particularly sensitive to the distinction between what is *movable* and what is not. [G2P1] illustrated this with an example of picking up a child from practice, saying "a system can't be like, that's not important, we're going to move that an hour later so that you can meet Jeff in the bar. And then it's like, well, your child has left cold outside of the football prep pitch while you're having a drink". In other words, certain commitments carry implicit social and moral weight that cannot be inferred reliably from calendar data alone. This illustrates the need for personalised domain-level constraints, which is to say that proactive changes for certain people (such as family members) or obligations cannot be moved (without prior approval).

As reported in Subsection 5.9 on factors affecting the user's decision to use agentic features, constraining the level of assistance was widely discussed in the domain of private and professional use. Seven participants have thus explicitly rejected using an IPA for delegation of private communication, with even more participants fearing the potential of loss of interpersonal connection. On a more fundamental level, [G6P2] echoed a concern about the blurring line between personal and professional with present-day shared calendars: "I would have one issue with using this in my Outlook is that there's a lot of private life shared, as I have a shared calendar for work".

Thus, with the design guideline of *Domain Attunement*, we call for recognition of the finding that delegation, proactivity, and even assistance in certain personally-defined domains were widely seen as unacceptable, regardless of system accuracy. Conversely, participants were more open to delegation in domains where the consequences of error were limited or easily reversible, such as suggesting alternative time slots, identifying clashes, or optimising travel buffers. A scrutable IPA should thus aim to enable constraining its effects based on the domain of use, rather than just the potency of a certain feature, which we discuss next.

### Potency Stages

The idea of stages was first expressed by a participant in our expert discussion (Subsection 3.2.8). The idea was then identified as a potential principle of scrutability and implemented in SIPA4TM (Section 4.2.1). This implementation was then evaluated in our thematic analysis of the focus group discussions (Subsection 5.5.2). Finally, after our overall findings were distilled into a set of principles of scrutability (Section 6.2), we propose seeing stages as a design guideline within the role of meaningful constraints, rather than as a principle itself. This is in line with the principle of *Non-Prescriptiveness* detailed in Subsection 6.2.3, as we note that it may not be achievable in all cases.

Moreover, even though the stages were originally devised as a form of privacy protection through data use licences, whereby a stage couples a type of data with assistance features that require that class of data, participants in our study consistently interpreted stages as a mechanism for *expectation management*. As [G1P1] noted, clearly delineating the data sources "sets the expectations of what I know it's going to be able to do based on what it's using", a view echoed by [G8P2] and [G1P2], who appreciated that SIPA4TM made it "really clear what was being used". Several participants further observed that stages could support different preferences and contexts of use: [G6P1] suggested they allow users with varying comfort levels with control to adjust assistance accordingly, while [G6P2] and [G5P1] described experimenting with different levels to see which produced the most useful suggestions.

At the same time, stages were not perceived as limiting the assistant's potential, but rather as making its *potency more transparent* while allowing users to choose a preferred level of delegation. For instance, [G9P1] expressed a desire to "give it as much control as possible" while still valuing that the system is "more upfront about where that data is going," and [G8P2] similarly indicated they would usually select the highest level in order to receive more tailored scheduling advice.

So the revised key idea behind *potency stages* is not necessarily to have explicit data use licences in the forefront, but rather to enable *incremental shifts in responsibility*, allowing the user to increase the IPA's functional potency as their comfort and trust levels evolve. Notably, the realisation of stages goes hand in hand with the previously described guideline of domain attunement, with the former focusing on constraining the IPA's overall capabilities (globally) and the latter on constraining the IPA's actions within a user-defined domain (locally).

The analysis of our user study has underlined the importance of the realisation of stages within broader interaction design and the prospects of use. Consider the statement by [G4P2]: "I didn't want [the IPA] to have more control over my calendar because I don't know the app enough /.../ to make the right calls at all time. So giving it more permission, I wasn't sure

if it was going screw some things up in my calendar or not, so I just kept it at two stars for everything". This indicates that rather than tying features explicitly to data licences, as was suggested by the experts (cf. Subsection 3.2.8), stages should tie specific functionalities with prospects.

Such explicit coupling of prospects with functionalities in *potency stages* enables the user to accurately gauge the IPA's potency at specific configurations. The success of implementation of this guideline, however, depends on several conditions. Given the user feedback discussed within the theme of Scrutability and Control in Subsection 6.1.3, the success of acceptance of design in stages is conditional on the mechanism being:

1. introduced clearly (as part of deliberate onboarding, as we discuss further below);
2. yielding perceivable differences in system behaviour (discussed as part of the guideline on *prospect-capabilities alignment* and the principle of *integrated overlays*);
3. supporting incremental, revisable calibration of the user's evolving trust and situational needs (again, in relation to prospects-capabilities alignment, integrated scrutiny, as well as the attuned delegation and proactivity); and
4. allowing a clear separation of concerns, in line with prospects (and in particular with consideration of features causing a heightened anxiety, such as delegation of communication).

Still, it is to be expected that some participants, like [G7P1] would opt to the highest stage ("I would just go ahead with the highest, for sure, because then I just can use my Google Calendar"), whereas others, such as [G6P3] would want to retain the most basic functionalities only ("I think, if it's somewhere I put in the information and it's just using that, and it's got a very specific program, then maybe. But when it comes to AI, I feel like I'd be a bit more sceptical, because it's a bit more of like a black box."). Unless either extreme is overwhelmingly chosen, we propose that such an outcome should not be seen as a failure of design, but rather a reification of propensity to trust.

### 6.3.3 Prospects Alignment

With *validity*, Kay and Kummerfeld [2013] denote the role of scrutability in enabling the user to discover and correct errors in their user model.

Our proposed framing of scrutability also supports this role, with the principle of *System Level* extending it to the discovery and correction of errors in personalisation. In fact, with the role of *Proactive Support* discussed above, we expressed the idea that it's not just the user who has an active role in scrutinising the assistant, but that the assistant should also actively recognise

when scrutiny should be initiated, offer explanations, and then empower the user to control it.

With the role of *Prospects Alignment*, we expand this idea of validation to the field of prospects. With the guidelines below, we describe how scrutable IPAs model the user's current likely prospects of using the IPAs and how the IPAs in turn allow the user to validate their own prospects and the IPAs' modelling thereof.

To illustrate the idea, consider a case when the IPA proposed moving an event to a time that is unsuitable, not due to a user preference, but due to a constraint unknown to the IPA (e.g. a one-off event that is not in the calendar – a case which had repeatedly come up during the user study discussion). The user should, in that case, be able to scrutinise the decision at the system level, to see whether the cause is a deeper misunderstanding or (as in this example) a one-off mistake due to a missing piece of information. While this alignment is a role scrutable IPAs are to support continuously, the user's view of their prospects is most malleable during the onboarding.

### Current Prospects Modelling

Our user study found that the possibility of scrutiny does not directly affect users' trust in the IPA. Instead, shortly after commencing their interaction, the users estimated their prospects – i.e. the likely outcome measures of engaging with the IPA in relation to the required time and effort, self-concept actualisation, and their social standing – and then reframed the value of investigating the system accordingly. As an illustrative example, recall the stance of [G2P1]: "in the first few minutes, like once I started doing the task, I found that the assistant wasn't actually really helpful or I just like inherently just avoided using it because I found that, okay, I know I can change it myself, so I'll just do that because it's easier because I'm trying to figure out the assistant".

Similarly, we observed that low-stakes suggestions, such as the insertion of buffer times or simple reminders, were often accepted with minimal scrutiny. In contrast, socially or personally consequential changes – such as rescheduling a meeting with others or cancelling a commitment – triggered a much deeper inspection. For these high-stakes tasks, the participants' confidence in their own tacit understanding of their schedule served as the primary benchmark; as they sought quick confirmation that the system's suggestion aligned with their existing expectations rather than a detailed explanation of the underlying logic.

With *current prospects modelling*, we propose that the IPA should gauge the user's likely perception of their prospects at that moment. These would serve as the system's starting point for its own scrutiny-related nudges and system interactions. The IPA may do this by mapping

what the user's actions (and features they choose to utilise or avoid) likely convey about their perceived prospects. Such updates with time improve the accuracy of modelled prospects with the user's views. Moreover, the IPA may update its modelling of the user's currently held prospects through explicit user feedback, either volunteered from the user at a time of their choosing or elicited by the IPA, in line with the guideline of Cues Recognition.

### Prospects-Capability Alignment

Not wanting to invest the time to delve into the IPA's scrutability features is not necessarily an indication of the user's inability or laziness. Instead, in line with the principle of *Prospect-Grounded Design* (Subsection 6.2.2), this should be seen as a sign that the user considers scrutability in that moment a bad investment, i.e. the time and effort put into scrutiny then would not improve their future prospects. This conclusion may stem from the user's propensity to trust, bad system interface design (including inappropriate affordances), or unreliable system behaviour to that point. It can also stem from their prior negative experiences with similar technology, such as mobile phone-based IPAs, for which [G1P2] – in line with the majority opinion – said that they "would find them to be a slower and less effective user experience than just typing or using [their] phone as [they] usually would".

However, if the IPA – through its modelled understanding of the user's perceived prospects – estimates that it would actually be a worthwhile investment for the user to learn about some capability or to *align their view of the prospects with the IPA's actual capabilities*, then the IPA should proactively engage the user to delve deeper in scrutinising it. The goal of such scrutiny is trust calibration and helping the user adjust their use of the IPA in line with their risk comfort levels and the underlying capabilities.

Since this calibration is happening on the level of prospects, which both the user and the IPA model in some way, it is important to concede that alignment remains a difficult problem across the fields of cognitive science, neuroscience, and machine learning. This is explored by Sucholutsky et al. [2025] in their perspectives paper on alignment between artificial and biological intelligence. In the paper, Sucholutsky et al. [2025] argue that – while representation alignment between humans and models might lead to output alignment as a result – representational alignment is very difficult and is not guaranteed or even necessarily indicative of better performance in task-solving.

Heeding the level of difficulty conveyed by the warring, we refrain from more encompassing implementation extrapolations. Instead, we ground the *Prospects-Capabilities Alignment* design guideline in the idea that the IPA's modelling of prospects – with regard to time and effort, self-concept, and social standing and in relation to both what the IPA models the user's desired and current prospects of use are, as well as what the capability-facilitated most

suitable prospects may be – should be addressed in systems responses to scrutiny attempts directly.

To do so, the IPA should explicitly acknowledge the risks and benefits of each feature or assistance capability. The principal goal is to align the user's perceived prospects of using the IPA (or its feature), the IPA's model of the user's perceived prospects, and the actual prospects given the IPA's capabilities and available information. The exact realisation remains dependent on the application, context, and the IPA's current capabilities. Regardless, the role of *deliberate onboarding*, discussed next, remains crucial.

### Deliberate Onboarding

Finally, we propose that the onboarding process for scrutable IPAs should deliberately indicate to the users what their prospects of use are. Our study has indicated that the user often felt unable to calibrate their trust towards the system's capabilities and to modify permissions stages because they lacked a baseline understanding of the system's reliability. Similarly, SIPA4TM's offering of explicit control of personalisation was sometimes met with surprise, either positive or ambivalent. Moreover, Sucholutsky et al. [2025] report that initial alignment constraints what the system is to learn to represent and output in the future.

This indicates that *deliberate onboarding* should also highlight scrutable features, especially if they differ from the currently typical interaction design. This onboarding should be targeted at prospects, address the user's concerns directly, and align with the *apt explanations* guideline to indicate their benefit and avoid having the user simply "skip them".

Still, it is worth emphasising that a minority of users in our study engaged with IPA's scrutability features at the outset of their interaction. This may indicate that there is a "propensity for onboarding scrutability", either built from innate curiosity or caution, or stemming from an educational or vocational background. *Deliberate onboarding* for such eager users may be lengthier; that is, even the onboarding to scrutable proactive personalised assistants should itself be personalised to the user's receptivity, scrutable for further enquiries, and proactively support the user in aligning their prospects with the system's capabilities.

## 6.4 Chapter Summary

We formalised the concept of prospects in HCI. We indicated how prospects may be used to quantify user trust towards an IPA and situated this with a paradigmatic quotes form on our user study, as well as in reference to the literature on human-to-human trust in organisations [Lewicki and Bunker, 1996] and human trust towards automation [Lee and See, 2004].

We then offered evidence that prospects of IPA use may fulfil properties originally defined for the value function in Prospect Theory by Tversky and Kahneman [1992]. Specifically, our findings suggest that users' thinking about the prospects of using a proactive IPA is reference-dependent, discriminatingly sensitive to utility but with reduced marginal pain, loss-averse, and edge-case-dominated.

We then distilled our findings into a set of proposed user principles for scrutable proactive IPAs. These principles reframe the concept of scrutability away from user-model level [Kay and Kummerfeld, 2013] to system level, thus accounting for both a technological change (away from explicit ontologies and logic-based deduction towards embedded and opaque models) and a shift in user expectations (where understanding of the system's capabilities is only of interest if the system is already advanced enough to warrant the time needed and if the effort invested is seen as likely to result in a clear benefit).

We defined six principles of scrutable proactive IPAs: 1) *System Level* (scrutability is defined and realised at the level of the system, not the user model); 2) *Prospects-Grounded Design* (prospects are the grounding for the IPA interaction design process, so scrutability should not be considered a goal with value independent of calibrating and improving the user's prospects); 3) *Non-Prescriptiveness* (scrutability as a capacity of an IPA conceptually independent of the underlying system architecture, rather than a set of guidelines conditioning the modelled representation and reasoning); 4) *Integrated Overlays* (the IPA integrates all contents, including for the purposes of enabling scrutability, and enables communication with the user through a conceptually unified communication channel); 5) *Uncertainty Signalling* (uncertainty of the IPA reasoning is signalled and scrutable); 6) *Model Transferability* (the IPA facilitates customisable knowledge extraction, rather than raw data retrieval).

Relatedly, we defined a set of eight design guidelines, split into three roles for scrutable IPAs. The first role is *proactive support*, which means that the IPA monitors user interactions for cues that suggest scrutiny might be beneficial, and then proactively offers apt explanations of itself and empowers user control. Second, with *meaningful constraints*, a scrutable IPA offers domain attunement and potency stages, which limit its capabilities to specific user-defined domains or specific global capabilities, respectively. Lastly, with *prospect alignment*, scrutable IPAs are expected to model and update what the user currently believes prospects of use likely are and then to help the user align these prospects both during the onboarding and in later interactions.

# 7 Discussion

In this chapter, we first briefly recapitulate the entire research process that has ultimately led us to the prospects-based paradigm, the principles of system-level scrutability, and the consequent design guidelines. Section 7.1 discusses how this new framing of scrutability relates to the motivating problems proposed by Kay and Kummerfeld [2013]. The section then argues that our reframing of scrutability also addresses the new problems we identified in our user study. Section 7.2 then discusses potential pathways for realisation of the envisioned principles and roles of scrutability in reference to the work on Agentic Interpretability by Kim et al. [2025]. Lastly, we conclude the chapter by discussing the limitations of SIPA4TM, our user study, and the broader research approach and propose potentially interesting avenues of future research in Section 7.3.

## 7.1 Scrutability of Proactive IPAs: From a Technology-Prescribing Goal to a Prospects-Based Capacity

This dissertation sought to investigate how the concept of scrutability, originally devised by Kay [1999] almost three decades ago, and its principles, as defined by Kay and Kummerfeld [2013], can be applied to the present-day vision of proactive IPAs capable of delegating for the user in communication with others.

To do so, our first step was to investigate the sources of dissatisfaction with IPAs available at that moment and the envisioned use cases and properties of envisioned perfect IPAs, so that we can extrapolate the role and realisation of scrutability within that vision. We achieved this by reanalysing the data from user studies by Cowan et al. [2017] and Völkel et al. [2021]. The studies indicated that users at the time saw contemporary IPAs as unreliable devices with limited personalisation, standing in stark contrast to the envisioned future IPAs, which were anticipated as highly capable, explainable, and proactive in both action-taking and communication. In our first diagram of envisioned IPA interaction design (Figure 3.1 on page 33), we thus delineated scrutiny and user control as a conceptually distinct communication channel

that would exist alongside task-oriented communication and social communication.

Following the feedback from the publication of the reanalysis of the studies and the diagram and further literature reading, we grounded the use case for further research to time management assistance. Time management was chosen as it was one of the earliest use cases for cognitive computer assistants such as Calo [Yorke-Smith et al., 2012], which had been recognised as a desired use case by users in the study by Luger and Sellen [2016]. It was also our belief that time management would narrow the scope of application considerations without making the use case too trivial for a meaningful investigation.

We then situated our initial outlook for the role and realisation of scrutability of proactive IPAs for time management by conducting a workshop-style discussion with experts in the domain of human-computer interaction and related research areas. The idea of stages, as an explicit coupling of data licences and assistant functionality, emerged as a potential new principle of scrutability during the discussion. Furthermore, the experts stressed the importance of appropriate data stewardship, given the inherently sensitive nature of one's calendar and the complexity of the underlying modelling required to provide meaningful user assistance. The experts also warned of the risks of the IPA reinforcing unhealthy behaviour and debated the appropriate implementation of proactive features. Finally, the findings highlighted open questions on the desired interaction between the IPA, its primary user, and other humans and assistants. The experts then sketched a model of the envisioned IPA's internal design – including modelling necessary for goal tracking, activity classification, and explanation generation – and the interaction. The diagram emphasised the importance of the user's inner model of the IPA and their level of system understanding. From the output of the expert discussion and the feedback from resulting publications, we then devised and published six interaction challenges (proactivity, delegation, modality selection, humanness and expediency, disclosure of capabilities, and effort-efficiency) and three modelling challenges (meaningful control, underlying technology, and data stewardship) that scrutable proactive IPAs ought to address.

In reaction to the feedback to the publications in relation to the expert discussion as well as the transfer report at the halfway point of the PhD research, we outlined the technical scope of SIPA4TM, our Scrutable Intelligent Personal Assistant for Time Management, that served as an anchoring point of our main user study. In scoping our main user study and SIPA4TM, we aimed to develop an assistant that would allow a quantification of user scrutiny, to track actual user behaviour, as well as provide meaningful calendar-management advice. Apart from logging interaction to contrast actual exploration tendency with self-reported attitudes on explainability and privacy, SIPA4TM's main goal was to inspire the focus group discussions, the main source of data from our user study. Moreover, we created a study protocol that included questions on the devised interaction and modelling challenges, as well as on the

features deemed too complex to implement meaningfully within the study's time frame (such as the delegation of meeting scheduling and other communication and goal-oriented proactive calendar manipulation). In developing SIPA4TM, we drew on previous scrutable frameworks such as those by Kay and Kummerfeld [2019] and Czarkowski [2006].

It was, however, during the period between our expert discussion and our main user study that chatbots and agents powered by foundation models advanced to the point of gaining widespread public adoption and becoming a dominant paradigm in research and development within the domain of assistive software [Bommasani et al., 2022, Zhang et al., 2026]. User expectations of both natural-language-based interfaces and underlying AI capabilities have then shifted. Notably, as Hanif et al. [2023] investigated, user expectations on explanations have also changed. The advancement in underlying capacity and the following use cases made some of the anticipated future features by both Kay [1999] and the participants in the user study by Völkel et al. [2021] a common place, such as a personalised summary of information, and personalised time management advice, and all in fluent natural language. However, the new technology also raised challenges not highlighted by the experts in our user study, such as opaque errors that are nevertheless communicated persuasively and do not elucidate the underlying mechanisms [Lipton, 2018]. Furthermore, the actual emergence of capabilities from black-box algorithms remains an unresolved object of active investigation [Bommasani et al., 2022], making actual scrutiny in the sense of Kay and Kummerfeld [2013] unfeasible due to the stochastic and opaque nature of the underlying model embeddings [Hanif et al., 2023].

Consequently, for the participants in our main user study, the idea of a user model with explicitly controllable values seemed atypical. Crucially, the idea of diverging from one's primary interaction patterns and goal-driven interaction for the purposes of scrutiny no longer appeared to be a good in its own right. Instead, the value of scrutinising the IPA and its model was predicated on the expected returns and broader consequences of using the system, and getting an insight into the consequences of the IPA's behaviour.

As the first of our two principal contributions, we proposed adopting and adapting the concept of prospects from behavioural economics to quantify the user's perception of likely results of their use of an AI system with regard to the change of time and effort needed to achieve their goals, the self-concept actualisation tendencies, and the effect on their social standing. We posited that prospects serve as the conceptual bridge between interaction with the AI system and trust in the AI system. We situated prospects as arising from affordances and interaction design, being strengthened by system reliability, and being both a part of and a proxy for the user's mental model of the assistant. We proposed study-based evidence that several properties of the value function in the Prospect Theory in behavioural economics apply to non-monetary framing of prospects that we defined. One such property is reference-

dependence in evaluating the utility of IPAs relative to the current role of technology in the user's life and edge-case dominance in reasoning about potential gains and losses. Moreover, we have presented evidence that there is a diminishing sensitivity to utility in thinking about the role of explanations, with users typically appreciating explanations that increase their confidence in their understanding of the underlying AI. But this appreciation most often did not extend into a willingness to scrutinise the system to the point of reaching a complete understanding of the powering algorithm. Strikingly, we also uncovered indications of reduced marginal pain regarding data use, with users who were more pessimistic about the state of privacy in contemporary computing indicating a more risky, or even careless, approach to data stewardship.

We proposed reframed principles of scrutability as our second principal contribution. We argued for shifting scrutability from a user-model level to system level as well as for five further principles for scrutable proactive IPAs. Most notably, we advocated that scrutability should be seen as a means to advancing both the perceived and actual user prospects of using the IPA. Instead of advocating for building scrutable first-class citizen user models that are then modulated into personalised applications built to admit such user models, we proposed that scrutability should not prescribe application design but rather integrate within the broader frame of prospect-grounded design. Scrutability, alongside personalisation, would then be one of the layers that shape the final system output, accessible through a unified system interaction channel. Given the historical patterns of adaptive application development and the current trajectory, we also softened the idea of user models being stored "client-side" to instead incorporate the principle of cross-model knowledge transferability into the goals of scrutable applications.

Three roles for proactive IPAs that follow these principles were then defined. First, with *proactive support*, we laid out recommendations for when and how the IPA should nudge the user to scrutinise the assistant's capabilities and assumptions, to the extent possible given the underlying algorithm. We also argued that proactive IPAs should empower user control in opportune moments. Second, given our empirical findings on user wariness regarding agentic features, such as proactive calendar changes or delegation of communication, the role of *meaningful constraints* called for scrutable IPAs to constrain their capabilities both based on personalised application domains and in terms of individual capability's potency. Third, *prospects alignment* proposes inferring the current perceived user prospects from user actions or explicitly eliciting them from the IPA, and then calibrating the user's estimate against the actual prospects. This role also highlighted the importance of deliberate onboarding.

In the remainder of this section, we now want to discuss how the motivating problem-oriented foundational goals of scrutability, systematically enumerated by Kay and Kummerfeld [2013],

are still addressed by our reframed system-level scrutability and the surrounding set of principles and design guidelines. Thereafter, we also deduce a definition of scrutability that embodies our reframing of the concept.

### 7.1.1 Problems Resolved through Scrutability

We argue that the system-level prospect-based scrutability still resolves all problem enumerated by Kay and Kummerfeld [2013], while also being helpful for new problems that we identified in our research on proactive IPAs for time management.

#### Original Problems

Kay and Kummerfeld [2013, pp. 3–4] propose a set of five problems of collecting user data that scrutable user modelling ought to resolve. Below, we summarise each problem in a sentence and then provide commentary on how the problem is addressed through our framing of scrutability.

**Privacy:** user models by definition contain personal information, which causes users to be concerned about their privacy.

While our expert discussion has reinforced this view, our consequent user study has revealed a more complex insight. With prospects-based framing we proposed that the scrutable IPAs need not guide the user towards scrutinising the way that their user model collects, represents, and infers personal data in an application context to the point of complete understanding. Instead, our principles emphasise the need for scrutable IPAs to elucidate the prospects of sharing certain data.

**Invisibility:** users are often unable to determine to what extent the content they see is filtered for them.

Scrutable IPAs that follow our principles would still address this problem, though not by offering a (potentially opaque and currently unrecoverable or conceptually unaligned) trace or explanation of their personalisation process. Instead, they should allow the user to understand how the content they see differs from that a new user might see. The exact realisation of this might differ, and the most suitable manner of doing so – such as through counterfactual or semifactual explanations of personalisation – is still a matter of open research. What we advocate for, is that scrutable IPAs even in tackling the problem of invisibility remain grounded in prospects, so to, for instance, emphasise to the user how the personalisation fits their self-concept by surfacing their preferences that fit a certain recommendation or noting how a

certain content filter might affect their social standing by diverging from the typical content for a region or the social class the user is in.

**Errors in User Model:** users do not have a systemic way to fix errors in their user model, which – as Kay and Kummerfeld [2013] phrase it – forces them to "resort to 'hunting' behaviors".

Our user-centric outlook is that exploratory interaction, through probing of system responses, is not necessarily an undesirable type of interaction. Moreover, it may be exactly the type of interaction that scrutable systems support (the indirect or "implicit" scrutiny). What remains a problem is if the user is forced to change their behaviour in undesired ways for a prolonged period of time in the hopes of influencing the system, as was described by Tintarev and Masthoff [2022]. Scrutable IPAs should provide clear pathways for users to effectively control the system and improve their prospects, and proactively empower them to follow them at opportune times.

**Wasted user models:** many devices and applications capture and store information about the user that is "staying dormant" instead of being utilised for personalisation.

In the years since the publication of the work by Kay and Kummerfeld [2013], this trend has only accelerated, with smart watches becoming the norm and IPAs commonly integrated in them [Pal et al., 2021]. In fact, rather than emphasising underutilisation of data for personalisation, recent works have instead emphasised the importance of research in, as Zanker et al. [2019] called it, the "dark side of online personalisation", whereby recommendations exploit user differences to tailor personalisation strategies or narrow the field of content to the degree where a concern about "filter bubbles" becomes prominent. Nevertheless, we do recognise that the user might want to extract from one IPA or AI system and reuse it in another, which is a related problem to that of wasted user models. In fact, this problem was one of the main motivators for our proposed *Cross-Model Transferability* principle.

**Control:** many adaptive systems do not let their users control either the personalised information stored in their user model or the resulting personalisation, which might, in part, reduce users' willingness to share personal information.

This problem is so prominent that scrutability became synonymous with resolving it in the chapter by Tintarev and Masthoff [2022] and works in recommended systems based on it [Balog et al., 2019, Labra and Santos, 2023, among others]. As discussed in Section 2.3, user control is not the totality of what scrutability is. Instead, user control is enabled thanks to scrutability. This separation remains in our system-level framing of scrutability. Even though

our principles do not envision the user being able to directly and fully control the (potentially black-box and stochastically embedded) user model, we posit that scrutable IPAs should not only enable control at the system level but also empower the user to feel confident enough to participate in the process. Again, the way to achieve this is for the IPA to faithfully (signalling uncertainty if present) support the user in gauging the degree of control possible and in understanding how these changes are likely to influence their prospects.

### Identified New Problems

Based on our empirical findings, we propose four new problems that may be addressed by scrutable proactive IPAs.

**Self-Efficacy Deficit:** The principles devised by Kay and Kummerfeld [2013] fit the expectations of the rules-based reasoning mechanisms at the time, which were at least theoretically traceable. Their framing of values such as privacy and user empowerment followed this technological context. However, as discussed in Section 6.2, IPAs are developed primarily based on opaque algorithms [Löfström et al., 2022]. Relatedly, expectations of what an appropriate explanation is and the user expectations have also shifted [Shneiderman, 2022, Lipton, 2018].

One such shift that we identified is that scrutability was seen as competing with the main tasks. That is, participants continuously weighed the prospective benefit of scrutiny against the immediate cost of attention, interruption, and effort, ultimately abandoning the assistant when it did not demonstrate value quickly enough. Crucially for our prospects-based framing, [G2P2] has explicitly framed understanding answers from an LLM-powered chatbot as a bad investment in terms of time and effort: "I think I'd say that I just kind of try and find out myself rather than trying to find out why this wrong answer is wrong. Like the amount of effort you put into figuring out why ChatGPT is wrong, you can use less of that effort and just do a bit more". Relatedly, [G4P2] emphasised they would only investigate the underlying system in case of a very clear need: "if it ain't broke, don't fix it. If it's broken, then take a look".

Several participants expressed hesitancy to modify system parameters directly, potentially due to concerns about unintended consequences. This sense of inadequacy was particularly prominent when control involved abstract or opaque model elements, echoing the sentiment expressed by [G4P1]: "I don't know if I have confidence enough in my ability to do it, depending on what the model is". As a result, some participants preferred not to intervene at all, choosing instead to just ignore incorrectly personalised suggestions. This is evident, for instance, in the thinking-aloud statement by [G2P3]: "I want to ignore the suggestion, because I just want to

start half an hour later, but not actually keep till after 5 o'clock". Still, the principle of retaining control of the direction of the suggestion was repeatedly emphasised in the discussion, including by [G4P2]: "think as /.../ I say, not as I do – I want these things to be there and then I could check them off or not because this is what I decided for myself".

This highlights a mismatch between the normative assumptions that underlie earlier scrutability work, namely, that users will invest sustained effort into inspecting and iteratively debugging user models [Kay, 1999, Kay and Kummerfeld, 2013], and the pragmatic reality of everyday time management, where the user's attention is a scarce resource and scrutiny is only rational when the prospects of doing so are positive. In this sense, our findings also resonate with the broader observation that users frequently express a desire for control and transparency, yet under-utilise transparency mechanisms [Kokolakis, 2017, Colnago et al., 2023].

With the formulation of our principles of scrutability in Section 6.2 and the role of proactive support in Subsection 6.3.1, we aimed to acknowledge this usage expectation direction and formulate scrutability principles that would result in an IPA interaction design which empowers the participants to make small, natural, but meaningful changes in expectations, rather than asking them to act in a way they no longer find natural.

**Justification Suspicion:** In our study, we have identified a sentiment that the explanation that the IPA provides might not be faithful to how the IPA actually works. Some participants even believed the IPAs could intentionally mislead them to protect the interests of the service provider.

For instance, consider [G6P3]'s answer to whether they would be curious about looking at ChatGPT's "memories" features, which are said to indicates what the model "remembered" about the user [Dash et al., 2026]: "I'm curious but I also don't know if I would believe that like what they have there is like all that... Like maybe that's just what they want to show you. Maybe they actually have more kind of information that they're not really talking about. That might be a bit more uncomfortable, I don't know". So, the participant's concern stems from a belief that the memories are intentionally incomplete to fit a latent goal of making the user feel good while using the platform. Relatedly, an expert in our discussion (PB-S171) emphasised that even when systems claim to follow principles such as data minimisation, the justification for collecting additional information is often framed in vague terms such as "improving the service", leaving substantial grey areas and raising concerns that data gathered for assistance may later be repurposed for other, undisclosed uses. An IPA that follows our principle would address this issue by focusing on prospects of use in a manner that acknowledges uncertainty while providing *apt explanations* (cf. Subsection 6.3.1).

**Persuasion Alignment:** In their article on the increasing influence of the personalised "online layer" on the daily "human layer", Herder et al. [2024] illustrate how using navigation emphasise the need to find a healthy balance between the benefits that technology provides with the enjoyment of daily activities. They illustrate this point by referring to the framing from Kahneman [2011] of mostly automatic and effortless System One thinking and the conscious and effortful System Two thinking, and suggest that the technology should also alert users when to engage with the "human layer" more fully through active engagement, rather than often seemingly more-convenient following of personalised and persuasive instructions coming throw the "online layer". The example they use is that of walking through a city, where a smartphone with a maps application may be very good at leading the user to their final location most efficiently without the need for a cognitive effort of remembering street names and landmarks (a clear improvement in terms of the time and effort prospect), but in the process, the user may become more alienated from their place of residence, miss potentially enjoyable establishments outside of their personalised online bubble, and ultimately feel less engaged with their own city (which we would denote as a decrease of the self-concept actualisation prospect).

This framing indicates another new problem that system-level scrutability may address – persuasion alignment. As explored by Herder and Herden [2023], the optimal voice for a persuasive system is very context, user, and task-dependent, thereby necessitating personalisation. Scrutable proactive IPAs following the guideline of *empowering user control* may help the user find the most beneficial ratio of self-concept and time and effort prospects.

Our user study further strengthens the case for considering the persuasion alignment problem. The participants wanted assistance that reduces cognitive labour, yet were sensitive to the assistant overstepping into socially or personally consequential territory. This surfaced as a strong preference for support that remains advisory, especially when the assistant could not reliably infer what is non-negotiable in a schedule. As [G7P2] noted, "not a lot of people would be a huge fan of being told what to do, but rather having a suggestion of what could have been done to make things better". Relatedly, [G1P1] highlighted a fear of such algorithmic invasiveness, stating they would be "really scared" to give an AI access to emails because "if there was some kind of glitch or if it got the wrong context" it could mass-email their contacts. Similarly, [G4P1] identified a "weird social thing" regarding the collapse of context between work and personal life, specifically referencing the discomfort of a boss seeing personal dental appointments in a shared calendar. This suggests that for proactive IPAs, the *Social Consequences* prospect is a dominant inhibitor. Scrutability mechanisms must therefore do more than show *what* data is used; they must demonstrate that the system understands the *social context* in which that data is deployed.

**System Lock-In:** A further problem that arises for proactive IPAs in the foundation-model era is *system lock-in*, that is, the user becoming dependent on a specific provider not only due to interface familiarity, but because the accumulated personal data, learnt preferences, and interaction history that cumulatively reinforce loss aversion and entrench the status quo, even if the user may be better served by a different system. In the scrutability framing by Kay and Kummerfeld [2013], the concern of "wasted user models" already anticipates that if user knowledge is captured yet remains non-transferable, then the insights gained from their data remain limited. Our data suggests an analogous dynamic for time management in particular. Participants described their scheduling practice as distributed across systems and devices, often with one e-calendar for work and another for private events.

Consequently, we treat the lock-in problem as a motivation for the principle of Cross-Model Transferability: rather than presuming a single stable user model architecture, scrutable IPAs should, at a minimum, support the extractability of user-relevant knowledge in a structured form that can be inspected, corrected, and reused across systems. The practical aim of addressing the problem is to enable users to carry over the assistant-relevant personalisation state (in the case of time management: personal constraints, inferred priorities, and delegation preferences) so that switching costs do not systematically outweigh the prospects of adopting a more reliable or better-aligned assistant.

### 7.1.2 Definition of Scrutable Proactive IPAs

Having analysed users' views on scrutability, connected them with our prior findings into a set of principles and roles, and discussed how the principles and guidelines relate to previously defined problems that motivate scrutability, we now offer a unifying definition of scrutability and of scrutable proactive IPAs. To avoid the ambiguity of the *they* pronoun used elsewhere in the dissertation (as *they* may refer to both the IPA and the user or just the user), we refer to the user by the pronoun *she* in the definition and to the IPA by the pronoun *it*.

**A proactive IPA is *scrutable* if it actively supports the user in aligning her current view of her prospects with the system's actual capabilities and likely outcomes, and in optimising her prospects given the IPA's functionality and her broader circumstances.**

*Scrutability*, consequently, denotes the capacity of a proactive IPA to provide such active prospects alignment and optimisation support.

For clarity, here we also offer a short-hand operational meanings of all concepts in the definition, and indicate sections where more detail is provided:

- Proactive Intelligent Personal Assistant: AI-based systems built with the purpose of

assisting the user (in general or in a specific domain, such as time management) capable of providing assistance that is complex (*intelligent*), adapted to the specific user and their circumstances (*personal*) and of doing so in a manner with a certain degree of autonomy, rather than just awaiting specific user instructions (*proactive*). We provide a state of the art regarding traditional IPAs and the desired level of IPA proactivity in Section 2.1, while the state of the art on LLM-based agentic systems capable of acting autonomously in Section 2.5.

- Prospects: Discussed in detail in Section 6.1, based on user study findings in Subsection 5.5.1. In short, prospects are likelihoods alongside three axes:
  1. The user would spend less (or more) time or effort by using the assistant compared to not using it;
  2. By using the assistant, the user actualises her self-concept (rather than increasing the congruence between her ideal and experienced self); and
  3. The social standing of the user would improve (change in the way the user desires, rather than degrade in the user's view) as a consequence of her using the assistant.
- The user's current view of the prospects: The belief (consciously formed or subconsciously held) regarding the user's prospects that forms the basis of her trust towards the system and is reified in the user's interaction with the system
- The system functionality and likely outcomes: is an abstract target, both because the actual outcome necessitates correctly and completely predicting the future and the full system functionality (cf. Section 2.4) may be principally unknown to the model as well [Kadavath et al., 2022]. We still think that it provides a meaningful direction of optimisation, which can then be concretised and operationalised depending on the actual IPA architecture and use case.
- Actively supports: follows the role of *Proactive Support* discussed in Subsection 6.3.1.

If an assistant is scrutable, its user can *scrutinise* it, by actively engaging in the actively supported process of aligning her current view of prospects with the actual likely outcomes of IPA use. In other words, we expand the emphasis by Kay and Kummerfeld [2013] that scrutinising requires *active effort* by the user to also include *active support* by the IPAs. Thereby, our definition retains the key conceptual separation of scrutability from the more machine-centric concept of explainability, as discussed in Section 2.3.

Lastly, *user control* is the process of the user optimising her prospects given the IPA's current capabilities and broader circumstances, including the user's preferences, deeper values, current social standing, goals, etc. In line with Kay and Kummerfeld [2013], we thus preserve the

conceptual distinction between scrutinising and controlling an IPA, while noting that both are actively supported by scrutable IPAs. Scrutinising enables meaningful control and, while conceptually distinct, these two processes may take place in an interleaved manner (cf. the principle of *Integrated Overlays* in Section 6.2.

## 7.2 Agentic Interpretability

In June 2025, Kim et al. [2025] – a team at Google DeepMind, the lab responsible for the development of the Gemini LLM model, which is to replace Google Assistant [Saeidnia, 2023] – published a preprint paper arguing for dialogue-based interpretation of LLMs, which they called *agentic interpretability*. In the paper, they define that a "method pursues agentic interpretability if it proactively assists human understanding in a multi-turn interactive process by developing and leveraging a mental model of the user, which in turn enables humans to develop better mental models of the LLM". This definition is based on the idea of the LLM and the user maintaining "mental model of each other".

In this subsection, we want to contrast the ideas from this paper with those of (parallelly developed) ideas of prospect-based system-level scrutability of proactive IPAs. Agentic interpretability is proposed with the current LLM-based agents in mind, while our definition of scrutable IPAs is explicitly non-prescriptive regarding the underlying model. Nevertheless, in this section, we will assume LLMs to be powering a scrutable IPA, thus allowing for a clearer comparison as well as an outlook for how our principles and guidelines from Chapter 6 relate to the current state of the art.

We start by discussing each of the three core ideas of the proposed agentic interpretability method from Kim et al. [2025]: multi-turn interaction, proactive assistance, and mutual mental models. We then discuss the feasibility of the approach in reference to concrete interpretability approaches by Schut et al. [2025] and Prasai et al. [2026], before concluding the section by returning to the key ideas by Kim et al. [2025] and contrasting them to our proposed framing of scrutability.

### 7.2.1 Iterative Dialogue

The iterative dialogue between the user and the agentic system, as Kim et al. [2025] see it, serves to explore and clarify concepts and explanations. Here, it is important to note that Kim et al. [2025] repeatedly emphasise that the concepts embedded by the LLMs (termed "machine concept" in the paper) and the user's understanding of concepts ("human concept") may not overlap. In particular, Kim et al. [2025] note the prior work by Schut et al. [2025] on the existence and extraction of new "superhuman" concepts from AlphaZero's internal vector

representation of chess moves, which we discuss in more detail in Subsection 7.2.4. Kim et al. [2025] then posit that the users may have two end-goals of engaging in such iterative dialogue, with one being to learn machine concepts (with a given example being a "superhuman" concept about a previously unknown chess strategy) and the other being to improve the performance of the LLM to make it aligned with the human's actual intent coded in the used concept.

Kim et al. [2025] underline that this iterative approach goes beyond the conventional human-in-the-loop paradigm, as human responses are integral to the dialogue's trajectory. The user may ask for hypotheticals, further clarifications, and hypothesise – all of which lead to different downstream explanation trajectories. In fact, Kim et al. [2025] say that one of the foreseen challenges of agentic interpretability that they call "human-entanglement-in-the-loop", i.e. the wide variance of explanation trajectories that is to be expected, given the interpersonal differences in expertise, curiosity, and reasoning style. This is compounded by the fact that output of LLMs is highly sensitive to differences in input (which is the basis of prompt engineering [Chowa et al., 2026]). Later in their paper, Kim et al. [2025] propose methods for a more consistent evaluation that use other LLM-based agents as proxies for human users in end-task metrics "to some extent". Kim et al. [2025] also propose that, through the iterative dialogue, the user may also be able to uncover potentially deceitful models by having the LLM produce contradictions in its explanation chains.

Looking back at the work by Kay and Kummerfeld [2013], the user's active effort to investigate the underlying model is a defining characteristic of scrutability. Furthermore, learning more about the model (and the AI system) and aligning it with the user's preferences and goals are among the defining goals of prospects-based scrutability, as we had discussed in Subsection 6.3.

One difference between the model-level scrutiny put forth by Kay and Kummerfeld [2013] and our system-level scrutiny is that the former prescribes a user model architecture that enables the user to scrutinise both knowledge representation and reasoning. In contrast, agentic interpretability and our system-level scrutability foresee that the user should be the arbiter of sufficient understanding and that exploration should be facilitated to the extent possible by the underlying technology while retaining faithfulness. Notably, the system-level scrutability and the agentic interpretability both foresee that the iterative process of understanding-seeking should occur through the main channel of communication, namely, the one used for the AI system's main purpose.

### 7.2.2 Proactivity

The second element that forms the core of agentic interpretability is proactivity, which, for Kim et al. [2025], refers to having the agent proactively assist human understanding through

dialogue. The goal of this proactive assistance is to enable the user to develop a better mental model of the assistant, while also improving its own "mental model of the user" (explored further in the next subsection). Kim et al. [2025] then connect proactive explanations to research on achieving common ground in communication by Clark and Brennan [1991], and note that proactive explanations should be structured to allow for open clarification requests and explicit confirmations.

Such a vision of proactive explanations, in which not just the user but also the AI system takes an active part in understanding-seeking, is one of the main new aspects of scrutable proactive IPAs compared to the idea of scrutable user models envisioned by Kay and Kummerfeld [2013]. This proposition is encapsulated in the design guideline for scrutiny *cues recognition*, which posits that IPA should anticipate user confusion and proactively aid users in scrutinising the underlying causes. Moreover, the guideline of *empowering user control* that posits that the IPA should proactively entice the user to control the system's behaviour to fit their needs better.

### 7.2.3 Mental Models and User Models

The third and final core element of agentic interpretability is that the machine and human build "mental models of each other", an idea that Kim et al. [2025] describes as under-explored in the context of agentic systems. Kim et al. [2025] maintain that LLM-based agents also possess a model of their user (which is in line with the prior work by Herder and Masthoff [2025], which posits that all adaptive systems need a user model, even if it is represented implicitly in neural embeddings rather than explicitly as a formal ontology). In a divergence from the terminology utilised by the user modelling community [De Bra, 2017] and in this dissertation, they refer to these models as the "machine's mental model of the human user" in reference to prior work on learning by thinking in natural and artificial minds by Lombrozo [2024], rather than by the term user model. The key suggestion by Kim et al. [2025] is that the LLM-based agent should also model the user's knowledge and confusion, effectively creating a metamodel within its model of the user's model of its capabilities and concepts.

Such metamodelling is, of course, very complex. We posit that it also represents a clear indication of why prospect-based framing (as presented in Section 6.1) is a useful conceptual tool, as it makes such meta-modelling more conceptually grounded and well-defined. In fact, this is a further motivation for the principle of Prospects-Grounded Design in Section 6.2. Moreover, our design guidelines of current *prospects modelling* and *prospects-capability alignment* in Subsection 6.3.3 provide insights on how to achieve such modelling of the user's current prospects and to then align them with the system's capabilities in a user-desired manner.

### 7.2.4 Feasibility

Along similar lines to our approach to defining the principles and design guidelines of scrutable IPAs in Chapter 6, Kim et al. [2025] state that many aspects of their proposed method and evaluation targets are intentionally kept abstract to leave room for implementation- and application-specific approaches.

Given the rapid advancement of foundation models [Bommasani et al., 2022], LLM-based AI systems capable of supporting the ideas of system-level scrutability or agentic interpretability may be feasible in the near future. Still, the manner to achieve it might be rather complex. To illustrate the complexity, we first describe the method of "superhuman" concept teaching by Schut et al. [2025], a paper that Kim et al. [2025] cited as a possible approach to achieving agentic interpretability. And then, to further illustrate the state of the art of dialogue-based exploration of LLM capabilities, we describe the approach by Prasai et al. [2026].

#### Human–AI Concept Gap Discovery and Transfer

For both Kim et al. [2025] and Schut et al. [2025], "superhuman" concepts refer to concepts unsupervised machine learning models obtained through new ways of reasoning over existing knowledge which humans do not yet possess. To illustrate their approach to the "superhuman" concept of extraction and teaching, Schut et al. [2025] chose AlphaZero, which had learnt chess through self-play without human supervision. To provide an idea of the complexity of the approach needed to extract such "superhuman" concepts, in this subsection we briefly describe the method used.

The first step in the method by Schut et al. [2025] is to excavate concept vectors by locating the values in the policy-value network and the Monte-Carlo Tree Search – the two components used for deciding the action of AlphaZero [Silver et al., 2018] – by utilising a concept discovery method based on convex optimisation [Ding et al., 2010]. Schut et al. [2025] define a concept and a minimal unit of knowledge and then achieve minimality by shrinking towards sparsity using the L1-norm while maintaining a set of constraints (in this context, the constraints are set to only allow sequences of chess states). Formally, this is represented as $\min ||v_{c,l}||_1$, where $v \in \mathbb{R}^{d_l}$ is a vector for a concept $c$ in the latent space of layer $l$ with dimension $d_l$.

Then, having excavated the concept vector from the latent space, Schut et al. [2025] filters the concepts based on teachability and novelty. Teachability of a concept, i.e. a unit of knowledge, is assessed by checking whether another agent can learn it. Concretely, they 1) select a *student* agent that already "understands chess" but plausibly does *not* possess the target concept, by choosing an earlier AlphaZero checkpoint whose top-1 policy overlap with the teacher on held-out concept prototypes is below a threshold. They then 2) teach the student on a small

curriculum of concept prototypes by updating the student to match the teacher's policy on the training split via gradient-based optimisation (minimising a KL-divergence loss), and 3) assess acquisition on a held-out test split by measuring how often student and teacher agree on the top-1 move on unseen prototypes. Finally, 4) to distinguish *concept-specific* learning from generic improvement due to exposure to stronger play, they compare against baselines in which the student is trained on randomly sampled AlphaZero positions (with meaningful plans) rather than on the concept prototypes, and evaluate both on concept-specific and random sets.

Novelty, in turn, is framed by Schut et al. [2025] as *non-redundancy with human knowledge*: a discovered concept should capture structure that is present in AlphaZero's play but not well-explained by the distribution of human games. Schut et al. [2025] implement novelty by contrasting how well concept-related structure can be reconstructed from basis vectors derived from human games versus AlphaZero games, and they retain only concepts whose reconstruction error is consistently lower under the AlphaZero-derived basis across ranks $k$. Together, these filters instantiate an explicit bridge from latent "concept vectors" to learning "superhuman concepts". The approach both finds internal regularities corresponding to chess strategies and retains only those that can be transferred to another agent in a measurable way (teachability/usefulness), thereby showcasing how this can expand the knowledge already contained in human experts' knowledge (novelty).

### Agentic Assistant for LLM Interpretability

Prasai et al. [2026] provide a concrete example of how agentic assistance can be used for interpretability work by packaging disparate inspection tools into a single dialogue-oriented workflow. Their system, *KnowThyself*, is presented as a chat-based interface in which users can upload models, ask interpretability questions in natural language, and receive interactive visualisations paired with guided explanations [Prasai et al., 2026]. Architecturally, the platform comprises an *orchestrator* LLM that reformulates queries and decomposes them into subtasks, an embedding-based *router* that selects among specialised modules, and a set of *specialised agents* that encapsulate concrete methods. In the showcase system Prasai et al. [2026] demonstrate, these methods include attention visualisation via BertViz, activation analyses via TransformerLens, a RAG-based explainer that grounds responses in literature, a BiasEval module that reports demographic disparity and safety metrics, and a conversational interface that presents results to the user.

Within the framing of this dissertation, such work illustrates a plausible near-term pathway by which dialogue can support *scrutiny*: it externalises a workflow for *question asking, evidence gathering, and interpretation* into an interface that is accessible to non-specialists, thereby narrowing the gap between what interpretability research can do and what users (or even

system designers) can practically interrogate.

### 7.2.5 The Role and the Desired Outcomes

The primary cited motivation for Kim et al. [2025] is a stated concern about falling behind as experts and as humanity. Meanwhile, the question driving this dissertation was what human-centred AI view of scrutability of proactive IPAs would be. That is, our starting point was a specific need (trusting the assistant for daily assistance, including time management), whereas for Kim et al. [2025] the starting point was the current state of LLM-based agentic AI, and the related interpretability concerns such as deceitfulness detection. A shared assumption in both this work on scrutability of IPAs and the paper on agentic interpretability is the idea that the AI should be seen as a servant of human needs. In this subsection, we provide a comparison of desired outcomes between the two terms and their relation to the previous approach to interpretability and scrutability, which did not rely on active dialogue-like exploration.

Kim et al. [2025] emphasise that non-interactive interpretation approaches – which their paper calls "traditional methods" or "mechanistic interpretability" and describes as "techniques for complete and rigorous solutions" – remain important for cases where complete understanding is needed, such as in safety-critical systems and for uncovering behaviours of the model that cannot be inferred from input and output alone.

Interestingly, the description of mechanistic interpretability by Kim et al. [2025] is more similar to user-model-based scrutability outlined in Kay and Kummerfeld [2013], as they both highlight the goal of enabling the user to investigate the application down to the representational and inferential levels. Similarly, we also believe that user-model-based scrutability and rule-based systems may have a role to play for more exact or safety-critical systems, where full transparency and control are needed and can be traced back to rule-based systems.

For complex domains with fuzzier constraints and potentially consequential outcomes for one's self-concept or social standing, the prospects-based system-level scrutiny is clearly preferred by users, as the analysis presented in Section 5.5 indicated. Systems-level scrutability is also admissible for more powerful, but potentially inherently stochastic and opaque systems. Moreover, it enables reasoning and deciding based on prospects, the leading concern for users in the study.

Kim et al. [2025] also maintain that iterative exploration and understanding-building between the agent and the user is the manner to achieve effective human-AI collaboration. Still, Kim et al. [2025] concede that going from mechanistic interpretability to agentic one is a task with as-of-yet unresolved challenges, citing in particular the "human-entangled-in-the-loop" concept explained above as well as computational efficiency trade-offs. And yet, Kim et al.

[2025] maintain that facilitating a proactive dialogue-based process of mental model alignment should be seen as an "add-on to any method".

In this dissertation, we have provided a user-driven roadmap for the role and implementation of prospect-level scrutiny, capturing what both experts and regular users find desirable and plausible, if currently slightly out of reach, for contemporary IPAs to achieve. We thus principally agree with the framing of a useful and desired "add-on" to interaction.

We nevertheless stress that the term *scrutability* – as formally defined in Subsection 7.1.2 – offers a theoretically grounded conceptual foundation supported by decades of prior research [Brusilovsky, 2024, De Bra, 2017] and a qualitative analysis of the contemporary user perspective. Thus, while we agree with most of the reasoning by Kim et al. [2025] (including with their claim that "[t]he term *agent* is overloaded and lacks universal consensus"), we call on future research in this direction to, in the words of Botti [2025], avoid "reinventing the wheel" and to use the term *scrutability* in pursuit of proactive dialogue-based interpretations. We also advocate for utilising the paradigm of *prospects* as a more concrete and less encompassing target than machine mental models of user mental models. In the next section, we provide potential avenues of such future work and reflect on the limitations of our approach.

## 7.3 Limitations and Future Work

In this section, we elaborate on potential future work, starting with the limitations of SIPA4TM compared to commercially available voice assistants and emerging LLM-based agents. We then discuss limitations of the approach taken for our main user study, before revisiting more conceptual limitations and research opportunities regarding the three axes of prospects as a paradigm for connecting interaction and trust, as well as the prospects-based system-level scrutability principles proposed in this dissertation.

### 7.3.1 SIPA4TM, Voice Assistants, and LLM-Based Agents

There are several limitations of SIPA4TM, which was implemented to serve as the basis of our user study, as described in Chapter 4.

First, SIPA4TM was designed as a calendar assistant desktop application. As such, participants in our studies might have dissociated it from IPAs they are already familiar with, including the increasingly common LLM-based assistants. This is a significant limitation, as prior studies have shown that the feeling of familiarity may affect both acceptance of proactive actions [Zargham et al., 2022] and the calibration of trust [Olson and Xu, 2021]. Moreover, since our priority was building an assistant that is fully scrutable, SIPA4TM's dialogue was constrained,

preventing us from investigating *humanness* as part of the interaction beyond constrained textual feedback. Keeping in mind the influence of voice synthesis and spoken interaction cues on user perception [Moore, 2017], SIPA4TM did not incorporate a conversational interface, which is otherwise typical for IPAs, including in the domain of promoting behavioural change [Herder and Herden, 2023]. Due to implementation constraints, SIPA4TM did not incorporate autonomous proactive calendar modifications, thus limiting how far the idea of staged permissions control could be stress-tested in the user study. Lastly, SIPA4TM was implemented as a rule-based system in line with scrutability principles and design guidance by Kay and Kummerfeld [2013, 2019], so SIPA4TM's behaviour was more inspectable and stable than what would be expected from inherently stochastic models.

While the focus-group protocol was designed to account for these limitations, further studies are required to examine how users would interact with IPAs that implement prospects-based scrutability principles under richer communicative and action-taking capabilities. Since SIPA4TM was built with modularity in mind, future work could reuse its architecture while replacing the rules-based suggestion and modelling pipeline with either an LLM-based agentic backend or a Wizard-of-Oz approach that simulates more capable autonomy and negotiation. This would allow a more direct examination of the trade-offs that participants repeatedly articulated: the desire for powerful assistance and delegation, coexisting with persistent needs for verification, bounded agency, and meaningful recourse when the system is wrong.

### 7.3.2 Qualitative Approach

Our main user study, evaluated in Chapter 5, was structured to investigate user perspectives on challenges related to proactivity, disclosure of capabilities, meaningful control, and data stewardship. Nevertheless, the approach imposes several limitations that future work should address to verify generalisability and further define the mechanisms behind participants' stated expectations.

First, both the task-based interaction and the focus-group discussion were designed for a single-session qualitative analysis. While this choice enabled in-depth explorative reflection on unfamiliar interaction concepts (e.g., permissions and scrutability of a proactive calendar assistant), it also limited our ability to assess longer-run dynamics such as habituation, the evolution of perceived effort-efficiency, and the extent to which humanness characteristics become a determinant of acceptance once the assistant is embedded in everyday use. A more longitudinal design, ideally combining repeated in-situ usage with periodic elicitation interviews, would be better positioned to capture when scrutiny is triggered, how mental models stabilise, and whether the perceived cost of verification diminishes or persists as users gain experience.

Second, the evaluation centred on a single assistant instantiation and a single underlying class of decision logic. This constrains what can be concluded about the interaction between scrutability mechanisms and the opacity–capability trade-off that motivates much of the contemporary interest in LLM-based assistants and is contained in our proposed paradigm of prospects. A natural extension would be a comparative design in which participants interact with two assistants: for example, one that is more capable but less scrutable (e.g., an LLM-based suggestion engine with limited disclosure), and one that is less capable but more inspectable and controllable (e.g., a hybrid system built with prospects in mind). Such a contrast would allow clearer attribution of which user attitudes are driven by assistant competence, which by prospects-grounded scrutability, and which by the perceived stakes of delegation. It would also address concerns voiced in our expert discussion that explaining decision-making or the implications of privacy policies may be intrinsically harder for certain algorithmic families, and that some desired forms of constraint satisfaction may remain effectively soft in stochastic systems even under conservative decoding, thereby limiting the meaning, and potentially attractiveness, of user control.

Third, our studies did not explicitly model individual-difference variables that plausibly mediate the acceptance of proactive intervention and interest in scrutiny. While participants occasionally hypothesised such differences (e.g., framing preferences as generational), our design cannot adjudicate between age, life context, time-management style, and personality traits such as conscientiousness or neuroticism as explanations for these differences. Incorporating individual-difference instruments (e.g., Big Five) would therefore be a straightforward way to sharpen future analyses [Barman et al., 2024b, Fatahi et al., 2023], including quantitative modelling of how such variables predict stage use, explanation engagement, and tolerance for assistant agency.

Finally, while our participant pool was linguistically and culturally diverse, it was recruited and evaluated within one national context. Moreover, although best efforts were made to obtain a participant sample with diverse educational and vocational backgrounds, most participants were highly educated members of a university community (cf. Subsection 5.1.2). Since propensity to trust in automation and to reliance on assistance is known to vary across user contexts, replicating the study across countries would indicate the level of generalisability and help separate context-specific norms from more stable interactional mechanisms [Lee and See, 2004, Scholz et al., 2025].

### 7.3.3 Axes of Prospects

In this subsection, we revisit some participants' claims regarding each prospect axis, relate them to existing evidence, and outline how the prospects paradigm might be extended beyond

the present time-management framing presented in Section 6.1.

While one iteration of theme candidates saw *time* and *effort* as separate axes, the final analysis did not support such a split. On the one hand, there were not enough codes to support two themes, and on the other, statements by the participants often combined time-related and effort-related reasoning. This may be a genuine property of the time-management domain – that is, effort and time being tightly coupled in everyday planning – but it may also reflect the kinds of tasks in our study or the design of SIPA4TM. Future work could therefore examine whether separating these dimensions becomes more meaningful in domains where effort varies independently of time (e.g., emotionally demanding communication tasks, or high-stakes verification work), and whether general-purpose IPAs should treat them as distinct prospect components or as one aggregated prospect dimension.

With respect to *self-concept*, some participants expressed the concern that extensive AI use may attenuate their own reasoning or agency (for instance, [G6P1]: "I just don't want my brain to stop working because I'm relying on the AI to think for me"). This resonates with emerging empirical work suggesting that AI tool use and cognitive offloading can relate to self-reported critical thinking in complex and potentially heterogeneous ways [Gerlich, 2025, Liu et al., 2025]. A direct implication for future work is measuring both user satisfaction and trust when using self-concept prospect-highlighting proactive IPAs, as well as changes in behaviour and psychological well-being over time. Such studies may also track whether users increasingly accept suggestions without further deliberation, whether they retain the ability to plan unaided, and whether assistants can be designed as "cognitive partners" that prompt reflection and comparison rather than substituting reasoning [Liu et al., 2025].

*Social standing* prospect was often phrased through rather complex social embedding of shared calendars in both business (e.g. [G5P2]: "I find that people invite me to meetings and put them in my calendar. And I have to explicitly say no, I can't go, which is really frustrating. /.../ So I found that the way around that is to actually block-book [time to] focus") and private life (for instance, [G5P1]: "even if it's your friend, like: 'Oh, do you want to do some exercise today?' – It's good. But if I don't put it in there, /.../ I feel like I need more than [a calendar event]"). The prospects regarding social standing were especially prominent when considering delegation of communication and they were reified as, for instance, as the feel of loss of shared information context by [G2P1]: "my friend would be like, hey, remember, we talked about this or something. I'm like, I did not talk to you about this". Such considerations likely intersect with cultural and linguistic context. In that sense, it may be interesting to consider how the prospects of social standing and self-concept fit within a broader frame by Herder et al. [2024] in which the personalised interactional "online" layer (including delegative IPAs that reduce time and effort) increasingly affects the "human" layer (changing the cultural context, norms,

and other sociatial narratives).

Prospects beyond the three axes studied here are plausible and likely domain-dependent. Since our evaluation focused on time management, dimensions such as direct financial cost, legal risk, or health and safety consequences were not foregrounded. Yet, in many deployment contexts of foundation-model systems, such dimensions may become central. Future work could therefore treat the present axes as a starting set rather than an exhaustive ontology, and examine how prospects should be enumerated or abstracted across domains, including industrial settings where reliability, compliance, and operational cost are significant [Chkirbene et al., 2024, McNamara et al., 2025].

### 7.3.4 All or Nothing?

A final limitation concerns the extent to which scrutability, as defined and discussed in Chapters 6 and 7, should be treated as an "all-or-nothing" design goal. The question is inspired by Moore [2017], who discussed whether spoken language is "all-or-nothing" in speech-based interaction. This work highlighted that partial improvements in speech technology do not always yield proportional improvements in usability, because interaction breakdowns can dominate user experience and create a persistent "habitability gap" [Moore, 2017]. A similar question arises for scrutability in the era of LLM-based assistants and agentic systems: is partial scrutability sufficient, or do small failures in veracity, consistency, or disclosure create disproportionate costs for prospect and trust calibration?

Our findings suggest that the answer is unlikely to be purely affirmative or negative. Participants routinely asked for "just enough" information to support correction, verification, and expectation management, rather than exhaustive internal transparency, complex visualisations, or verbose justifications. At the same time, they also articulated that some failures are categorically difficult to tolerate: for example, when the assistant's explanations are persuasive but ungrounded, when the responses are inconsistent, or when the assistant's implied capabilities exceed what it can reliably do. In such cases, partial scrutability may still leave users with the burden of inferring what the system is like from (potentially misleading) surface behaviour, undermining the very aim of aligning prospects through proactive support.

This motivates two complementary avenues of future work. First, research should examine which scrutability roles and principles matter most, and under what stakes: for example, whether consistent uncertainty signalling and prospects disclosure contribute more to appropriate reliance than occasional deep explanations. Second, work on agentic interpretability by Kim et al. [2025] argues that even when systems can produce fluent rationales, they are not inherently able to "teach what they know" in a way that is guaranteed to reflect underlying causal drivers, outlining another area of future work. These considerations suggest future systems

may not offer uniformly high-fidelity explanations at all times, but they can still be designed to reduce the "habitability gap" by ensuring that the most decision-relevant uncertainties, trade-offs, and action boundaries are consistently surfaced, and that users retain practical routes to correction even when internal reasoning remains partially opaque. The exact realisation of our proposed principles – and the users' reaction to potentially imperfect reification thereof – is thus an exacting and broad area of potential future work.

## 7.4 Chapter Summary

This chapter opened with an overview of how the landscape of IPAs and AI systems more broadly has advanced over the course of this dissertation's research. We then discussed the impact this change has had on the user expectations of scrutability, based on our user study. In this discussion, we contrasted the idea of scrutabile user models – which was the basis of the work by Kay and Kummerfeld [2013], as well as our expert discussion and the implementation of SIPA4TM – with the system-focused and capabilities-driven view identified in the opinions expressed by our study's participants.

We then discussed the implications of the problems that Kay and Kummerfeld [2013] formulated as the motivation for scrutability and argued that our reframing of the concept still addresses all of them. Moreover, based on the findings of our user study, we identify four additional problems that our reframing of the concept may help resolve: self-efficacy deficit, justification suspicion, persuasion alignment, and system lock-in. Finally, we deduced a formal definition of scrutability as a system-level capacity that supports prospect alignment.

We then detailed recent publications on the state of the art in interpreting LLM-based agentic systems, with an emphasis on the idea of "agentic interpretability" by Kim et al. [2025]. We introduced the three main aspects of their proposed approach: iterative dialogue, proactivity, and alignment of "mental models" of the user and the LLM. We highlighted the importance of the problem of conceptual misalignment of machine and human concepts. We then contrasted this approach with our principles and roles, and called for future research on active, dialogue-based interpretation of LLMs to build on previous findings in the domain of scrutability.

In discussing the limitations and potential future, we revisited the three proposed axes of prospects – time and effort, self-concept, and social standing – and related them to open questions in the broader domain of human-computer interaction. We then proposed applying the concept of prospects in domains beyond time management to test the appropriateness of both the overall concept and the three specific axes.

Finally, we argued that scrutability should not be viewed as an "all-or-nothing" phenomenon and proposed related considerations regarding its feasibility in the era of LLM-powered IPAs.

# 8 Conclusion

This dissertation sought to investigate how principles of scrutability can be extended to facilitate effort-efficient control of anticipated Intelligent Personal Assistants (IPAs), aiming to provide user-centred design guidelines for future scrutable-by-design assistants. However, during the research work towards the goal, natural language processing advanced dramatically, making some anticipated future IPA use cases more graspable. Our discovery indicated that user expectations regarding the scrutiny of personalisation are directly influenced by the now-widespread assumption of highly capable AI assistance realised through inherently opaque models. In this dissertation, we detailed this changing dominant view, in which scrutiny may no longer be seen as a goal users value per se, but rather as a conditionally desired capacity to understand and steer an IPA. We thus posited that scrutability ought to be considered through the lens of *prospects* regarding time and effort, self-concept, and social standing. In this chapter, we conclude the dissertation by integrating our research contributions into a closing narrative, discussing the fulfilment of the research objectives, and outlining future work that could follow from our discovery.

## 8.1 Fulfilment of Research Question and Objectives

The research question guiding this dissertation was:

*How can principles of scrutability be extended to facilitate effort-efficient control of proactive intelligent personal assistants for time management?*

Our dissertation concludes by suggesting that the principles originally proposed by Kay and Kummerfeld [2013] ought not only to be extended but reframed to fit the changing technological landscape and evolving user expectations. Based on our empirical findings, we argue that scrutability should be understood as a system-level capacity rather than a user-model-level property. We then proposed and elaborated on five further principles of scrutability:

1. Prospects-Grounded Design, whereby scrutability centres on faithfully elucidating most likely prospects of IPA use to the user, allowing them to appropriately gauge the amount

of effort and time needed to use the assistant for a certain task, the effect on their self-concept actualisation (whether the IPA is likely to increase or decrease congruence between the user's ideal vision of self and the experienced self), and the possible impact of the IPA on the user's social standing;

2. Non-Prescriptiveness, which stipulates that scrutability should be seen as a capability of IPAs that is not conditioned on a specific internal architecture;
3. Integrated Overlays, allowing the user to scrutinise, control, and perform tasks via the same interaction channel;
4. Uncertainty signalling, highlighting the need for the IPA to adequately inform the user of the stochastic nature of underlying algorithms and other non-deterministic elements of interaction; and
5. Model transferability, which calls for allowing the user to extract and transfer knowledge collected and synthesised by the IPA.

### 8.1.1 Research Objective 1: Literature Review

Our first research objective read as follows: *Undertake a literature review of state of the art in intelligent personal assistants, human-assistant interaction, time management assistance, and scrutability.*

Research Objective 1 (RO1) was fulfilled through an extensive literature review in the five research strands underpinning the thesis in Chapter 2: 1) IPAs and their personalisation foundations; 2) calendaring, time management, and behavioural-change-oriented assistance; 3) anticipated human-assistant interaction; 4) scrutability as a design goal for inspectable and controllable personalisation; and 5) contemporary opaque generative algorithms underpinning latest advances. Moreover, towards further fulfilment of RO1, in Section 3.1 of Chapter 3, we reported on the details of user studies by Cowan et al. [2017] and Völkel et al. [2021] and then reanalysed the study data on which the publications were based. This literature review, in combination with the deep analysis of previously recorded sources of dissatisfaction with current IPAs and aspirations for future "perfect" assistants, has enabled a literature-supported initial view of user expectations. This view indicated that proactive assistance and personalisation are desired and insufficiently realised elements of IPAs and framed time management assistance as a paradigmatic, yet bounded application domain for studying proactivity and scrutability. Our literature review has also indicated that scrutability, while well-established in user modelling and hypermedia research, is historically rooted in legacy logic-based systems and therefore requires reconsideration in light of evolving IPA capabilities.

### 8.1.2 Research Objective 2: User and Expert Expectations

The second research objective stated: *Understand user and expert expectations regarding future capabilities of IPAs, especially as they relate to proactivity, time management, and scrutability.*

Research Objective 2 (RO2) was addressed through three complementary strands of empirical work that progressively grounded future IPA expectations in: 1) previously reported user studies; 2) an ad-hoc workshop-style expert discussion; and 3) identified expectations arising in situated interaction with SIPA4TM, our purpose-built time-management assistant, and the subsequent semi-structured group discussion.

First, to obtain the initial, literature-grounded view of expectations for proactive and personalising assistants, we reanalysed raw datasets from Cowan et al. [2017] and Völkel et al. [2021] in Section 3.1. This reanalysis foregrounded what users implicitly treat as desired IPA behaviour — particularly expectations around proactive initiative, personalisation, and delegation – thereby motivating the dissertation's narrowing to time management as a paradigmatic, yet bounded domain in which such expectations become testable through concrete scheduling considerations.

Second, we conducted and thematically analysed a multidisciplinary expert discussion in Section 3.2, eliciting structured expectations regarding plausible use cases, interaction constraints, and the role of scrutability as an overarching concern spanning modelling, usability, and governance. The resulting synthesis in Section 3.3 further consolidated these inputs into a set of interaction and modelling challenges that group and operationalise these expectations.

Third, RO2 was further fulfilled by confronting these prior and expert expectations with the expectations identified through our main user study presented in Chapter 5, in which participants interacted with SIPA4TM and then articulated desiderata and constraints in focus-group discussions. In particular, the findings include a desire for action-ready advice, complex yet consequence-bounding proactivity, and an expectation-control mechanism. In particular, the user study revealed pronounced limits placed on delegation in socially sensitive settings and a wide variance of not only desired but acceptable forms of proactive assistants.

### 8.1.3 Research Objective 3: Interaction Paradigm

Research objective three was: *Formulate an interaction paradigm for proactive IPAs for time management that enables users to scrutinise the system's underlying models and accurately assess its capabilities and assumptions*.

Our initial direction towards resolving this objective was closely influenced by the model-level

scrutability frameworks presented by Kay and Kummerfeld [2013] and the explicit interaction and modelling reference models by, among others, Peirce et al. [2008] and Meurisch et al. [2017].

With such guiding ideas, we first produced the diagram of the envisioned IPA interaction design shown on page 33 in Subsection 3.1.5. The diagram sees scrutinising the control and communications engine of an IPA as a conceptually distinct interaction channel of the assistant that exists alongside those for – in the terms previously used by Clark et al. [2019a] – task-oriented and social communication. The diagram also explicitly showcases elements the IPA may consider in its decision-making, such as the user's context, goals, and relationships, as well as its capabilities, including controlling internet-of-things devices and delegating communication with others. These use cases and contextual elements were gathered from literature on IPAs, and in particular, from the user studies by Cowan et al. [2017] and Völkel et al. [2021], whose data we reanalysed in Section 3.1, in part to discover elements of the paradigm for RO3.

While maintaining the same initial direction, we organised a multidisciplinary discussion, reported in Section 3.2, with four university professors and three PhD candidates as participants. One phase of the discussion asked the experts to collaborate on outlining a model of a proactive IPA for time management that they had previously debated. This outline was then digitalised and expanded to capture the expressed intent recovered through a qualitative analysis of that phase of the discussion, producing the diagram on page 41. This diagram presents a conceptual breakdown of the IPA's internal modelling structure into the main user model (capturing broader context and enabling autonomous acting with reference to explicit data use and representation licences), a calendar model (for activity classification, time reflection, and scheduling), a user model (for explanation generation and modelling of the user's cognitive load, goals, and preferences), and an external interaction model (which would model the user's relationships and obfuscate the data it has about the user in communication with others). Notably, the experts envisioned just a single interaction channel with the user, potentially supporting multimodal communication. Lastly, the diagram highlights the role of *stages* as explicit data-use licences tied to specific functionalities, as discussed in more detail in Subsection 3.2.8.

The analysis of our main user study, presented in Chapter 5, however, has indicated a significant divergence in user expectations and requirements for human-assistant interactions. As we then discussed in Section 7.1, this shift necessitated moving away from considering scrutability as a desirable property in its own right and moving towards scrutability as a capacity of the system to actively support the user in their understanding of the system's capabilities and to help the use of the IPA in a manner that is most likely to result in outcomes that fit the user's goals. This is where the idea of *prospects* as the paradigm of such interaction was identified as the

conceptual link between interaction with and trust in a proactive IPA. As can be seen in the illustration on page 148 and as discussed in Section 6.1, prospects were identified to have three dimensions or axes: anticipated time and effort, self-concept actualisation, and expected social consequences. In our paradigm, we adopted terminology from the psychological framework described by Rogers et al. [1959], in which *self-concept* represents the totality of an individual's beliefs and feelings about themselves, and the *actualising tendency* is aimed at reducing the *congruence*, or discrepancies between the ideal self and lived experience. We then identified self-concept in AI interaction as users' evaluation of how engaging with a system supports or hinders their self-actualisation, as described in our thematic analysis in Subsection 5.5.1.

In Section 6.1, we also provided evidence that prospects in relation to these three axes might fit the properties of the value function in the Prospect Theory of behavioural economics as defined by Tversky and Kahneman [1992]. Specifically, we found that the user's estimates of their IPA use prospects are reference-dependent, sensitive to utility yet reductive in marginal pain, and negativity-dominated. We have also situated the paradigm of prospects with the prior work on trust both towards other humans [Lewicki and Bunker, 1996] and automated systems [Lee and See, 2004]. The interaction paradigm was also situated in relation to other HCI concepts. We then proposed that our framing of prospects may serve as a proxy for the user's mental model of the AI system's capabilities and outcomes of use that may then steer actively supported trust calibration.

Lastly, our user study has also strongly indicated that users do not wish to leave their main interactional flow to scrutinise the assistant. For this reason, our final illustration of the envisioned user interaction with the envisioned proactive IPAs on page 173, emphasises the integrated communication with the assistant, for task-solving, prospects-alignment, and system-level scrutiny. Still, as the discussed potential future work in Subsection 7.3.3 recognises, in domains other than time management, further axes of prospect may have to be considered as well as the exact manner in which the paradigm may be applied to understand (scrutable) human-assistant interaction.

### 8.1.4 Research Objective 4: Scrutable-by-Design IPA

Research objective four contained a practical task: *Model, develop, and evaluate a scrutable-by-design IPA for time management based on the defined interaction paradigm.*

Research Objective 4 was fulfilled by conducting research and development work on interaction and modelling for the implementation of the *Scrutable Intelligent Personal Assistant for Time Management (SIPA4TM)*, a calendar assistant that offers proactive suggestions for calendar manipulation. SIPA4TM was explicitly designed to operationalise scrutiny and control in line with scrutability principles and design guidelines as formulated by Kay and Kummerfeld [2013,

2019] and to embody the insights regarding the envisioned scrutable interaction paradigm stemming from our reanalysis of the work by Cowan et al. [2017] and Völkel et al. [2021] as well as our expert discussion. The SIPA4TM implementation process, detailed in Chapter 4, was iterative and informed by continual collegial feedback.

To include user evaluation as part of RO4 and to catalyse the discovery of the final iteration of the interaction paradigm sought for RO3, SIPA4TM was implemented to log user interactions and provide a quantitative record of how, when, and to what extent participants actually engaged in scrutiny during task performance. Beyond its current research utility, SIPA4TM was engineered with extendibility and maintainability in mind, so it can serve as a reusable research vehicle for future scrutability studies rather than a one-off demonstrator.

The evaluative aspect of RO4 was finally fulfilled through an empirical evaluation centred on SIPA4TM. The protocol for this evaluation is elaborated on in Section 4.3 (and reproduced in full in Appendix A4.3), The study was realised as a three-part, task-based user study with 22 participants, in which users received time-management advice at both daily and weekly scales and could scrutinise advice as well as inspect and control the underlying user model powering personalisation. The mixed-method evaluation produced evidence by combining interaction logs with qualitative data from concurrent verbal protocols and subsequent focus-group discussion, allowing the dissertation to connect concrete in-situ interaction (including actual scrutiny actions) to participants' articulated expectations about proactivity, privacy, and the role of explanations. In this way, SIPA4TM and its study jointly developed an evaluation protocol that fulfils RO4 and constitutes a potential resource for future research.

### 8.1.5 Research Objective 5: Principles of Scrutability

The final research objective was to: *Distil the findings from the IPA interaction evaluations into a set of principles and design guidelines for scrutable proactive IPAs.*

The analysis of our user study has revealed that the scrutable features of SIPA4TM, such as explanations of suggestions, data permissions control in stages, and the transparency of the underlying user model, were not assessed in isolation but holistically, in reference to SIPA4TM as a system. Thus, the first principle of *System-Level Scrutability* frames scrutability as the IPA's capacity to facilitate and actively support the user's investigation of system capabilities and most likely outcomes, rather than as an architectural property of the underlying model level that admits complete user probing. This principle also enables scrutability to remain conceptually meaningful even if the user model is embedded in, rather than distinct from, the IPA's broader model of the world, thereby powering the interaction with the user.

The participants repeatedly voiced concerns about highly capable IPAs in relation to their

prospects, that is, what the repercussions of their decision (not) to use an IPA's feature would be. As elaborated in RO3, our thematic analysis has identified opinions regarding prospects on the user's time and effort (whether doing the tasks would actually be easier with the IPA), self-concept (whether they imagine their ideal self as using the IPA), and social standing (how the IPA may affect how others see them). A key finding was that users' perceived prospects of interacting with the assistant affect not only their trust but also the estimated utility of scrutability. The principle of *Prospects-Grounded Design* encapsulates this finding, stating that scrutable IPAs should be designed holistically, grounded in the anticipation of the user's foreseen and actual interaction patterns. In particular, the principle calls for a scrutable IPA to nudge the user to align their perceived prospects with the actual system capabilities.

We further identified a distinct user preference for what we termed *Integrated Overlays*, that is, for scrutability to be an overlay over personalisation and the core content, with which the user would then interact through a unified communication channel. This principle is based on observed interaction patterns, whereby users preferred probing the system directly rather than abandoning the default task-driven approach to explicitly investigate behaviour by reading explanations or inspecting modelled values. Our analysis of focus group discussions further reinforced this finding, with participants stating that their approach to learning about AI systems was trial-and-error and that they expected to learn about the system as it learns about them.

Similarly, accepting the status quo of modern opaque stochastic models, participants in our study were less concerned about the provenance of modelled information or the internal collection or representation of data about them. They were, however, generally sceptical about the explanations provided by the AI systems they had interacted with in the past. Recognising this, we defined a principle of *Uncertainty Signalling*, whereby the assistant surfaces cues about uncertainty and competence limits to support prospect calibration without requiring sustained effort. This principle supports calibrating user prospects by making limits, ambiguity, and confidence levels legible, indicating to the user when engaging in further scrutiny or control may be most impactful.

With the principle of *Non-Prescriptiveness*, we emphasise that scrutability should not prescribe or depend on a specific internal representation or reasoning mechanism. Instead, it should be seen as a capacity to support the user's prospects, alignment, and trust, achieved irrespective of the actual realisation of the underlying model, provided that it fulfils the principles of scrutability and allows for the implementation of the design guidelines.

Finally, *Cross-Model Transferability* reframes the original aspiration by Kay and Kummerfeld [2013] for "client side" (which necessitated the existence of a user model as a first-class citizen outside of the application) to one that allows for the previously defined principles of Non-

Prescriptiveness and System Level, while retaining the original motivation. The principle of Cross-Model Transferability states that a scrutable IPA should enable its user to extract and reuse accumulated personal knowledge that the IPA has not only collected but also inferred about them. This principle mitigates lock-in and preserves user agency across evolving assistant ecosystems, and supports the idea of users as owners of the knowledge IPAs extract about them.

Extending from these basic principles is a set of eight design guidelines intended to steer the realisation of three roles for scrutable IPAs: *proactive support, meaningful constraints, and prospects alignment*. These three roles for scrutable proactive IPAs build on the three roles of scrutable user models, as defined by Kay and Kummerfeld [2013] by adapting them to the proposed principles of scrutable proactive IPAs.

The role of *proactive support* builds on the original role of re-purposing user models to aid and improve the user's interaction with the system (such as for navigation or learning). The role states that scrutable IPAs should be able to *recognise cues* for when the user is most likely to benefit from investing time to understand the underlying system better. The IPA is then to offer proactive guidance with *apt explanations* (based on the findings from our empirical study) that elucidate actionable advice and relevant insights that might benefit the user. Lastly, the IPA should *empower user control*, thus helping address the recognised problem of doubt about one's ability to meaningfully improve the personalised assistance. With this role, scrutability is explicitly extended from an active process on the user's side to a two-sided process, as the IPA supports the user in commencing the scrutiny and throughout their investigation.

*Meaningful constraints* adapt the role of control of the modelling process as envisioned by Kay and Kummerfeld [2013] to the principles of System Level and Non-Prescriptiveness as well as the identified conditionality of the users' perception of the utility of investing time in controlling the modelling process. As such, scrutable IPAs may not necessarily enable full user control of the inflow of data or the model reasoning, but a scrutable IPA should still allow the user to constrain both the level of its overall capabilities (*potency stages*) and to limit its capabilities to some user-specific domains (*domain attunement*). Potency stages allow the participant to enable only the assistance elements they are comfortable with, such as disabling the IPA's ability to send emails on the user's behalf. Domain attunement limits IPA's actions within a domain meaningful to the user, with examples often cited in the user study, including not moving events in the user's calendar if they involve family members, or only delegating communication for scheduling purposes for professional events.

Lastly, the role of *prospect alignment* builds on the idea by Kay and Kummerfeld [2013] that scrutability enables validation of modelled information about the user and correction of mistakes. The design guideline of current prospects modelling stipulates that a scrutable IPA

should infer, model, and continuously update the user's held prospects of using the IPA. Such *current prospects modelling* enables the IPA to both adapt its behaviour to users' preferences and facilitate *prospects-capability alignment*, when it estimates that the user's perceived prospects of use are misaligned with the system's actual functionality and likely outcomes. While this is a continuous process, our user study has also indicated that users first and most strongly estimate prospects during the early stages of IPA interaction, necessitating *deliberate onboarding*.

## 8.2 Research Contributions

This dissertation has two major contributions. First is the paradigm of prospects, which bridges the user's interaction with and trust towards the IPA. Prospects are defined in relation to expected time and effort, self-concept actualisation, and social standing (C1). Second are the principles of scrutability for proactive IPAs and accompanying design guidelines, which shift the concept from user-model-level to system-level and bound its utility to supporting prospect calibration and improvement (C2).

We also contribute two minor contributions, with the first being our implementation of Scrutable Intelligent Personal Assistant for Time Management (SIPA4TM) that embodies the principles of and guidelines of scrutable user models as proposed by Kay and Kummerfeld [2013, 2019] and which was implemented with expendability and potential reuse in mind (c1). The second is a compilation of experts' and users' opinions on and expectations of anticipated proactive time management assistants (c2).

These contributions further the state of the art by providing a comprehensive human-centred perspective on scrutability after a significant technological and attitudinal shift towards IPAs. Our proposed reframing of the concept of scrutability aligns with current user expectations while addressing both the original motivating problems (privacy, invisibility, errors in user model, wasted user model, and control) and the new problems we identified (self-efficacy deficit, justification suspicion, persuasion alignment, system lock-in). Crucially, the new framing of scrutability is not conditional on any realisation of the AI system, as our principles are defined in terms of user-desired goals to be achieved rather than knowledge-representational constraints. As such, our framing and principles may help bridge the ongoing efforts on machine-assisted, dialogue-based interpretability of agentic systems [Kim et al., 2025] with the more established areas of human-AI collaboration and user control in adaptive hypermedia, where scrutability has played a significant role [Brusilovsky, 2024].

This new framing of scrutability is grounded in the adoption of the term *prospects* from behavioural economics. In adopting and adapting the term, we connected it with related

concepts in human-computer trust (trust, affordances, mental models), thus situating a new paradigm that may aid future methods of understanding adoption willingness and the relation between interaction and trust. Our user study has confirmed that advancing proactive AI systems has amplified users' concerns about the impact of technology on their self-concept and social standing [Zargham et al., 2022]. With prospects, we have proposed how these concerns may be addressed in a quantifiable, interrelated manner, thus offering a timely conceptual tool for future research on understanding user behaviour.

Our empirical findings also contribute to the state of the art by providing directional guidance in the domain of time-management assistance. These findings, stemming from our expert discussion and the SIPA4TM-anchored user study, are timely, as they provide deep qualitative insight into ongoing research on when proactive user support is appropriate [Meurisch et al., 2020]. The significance of our expounding of both user desiderata and the related concerns in an inherently personal domain of time management is further heightened in a moment when personalised online services increasingly affect human lives [Herder et al., 2024].

### 8.2.1 Publication Output

This dissertation comprises the content of four prior publications and one just-accepted paper:

1. Jovan Jeromela. "Scrutability of intelligent personal assistants". *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22)*. 2022.

   This work was presented at the UMAP'22 doctoral consortium event and was published in the main conference proceedings. The paper connects literature on intelligent personal assistants with those on scrutability. To better understand the user's expectations regarding scrutability, task delegation, and proactivity, datasets from Cowan et al. [2017] and Völkel et al. [2021] were reanalysed. The findings, indicating the user's clear interest in these three features, are also reported in the paper.

2. Jovan Jeromela and Owen Conlan. "Voicing Suggestions and Enabling Reflection: Results of an Expert Discussion on Proactive Assistants for Time Management". *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. 2023.

   Submitted as a work-in-progress paper and presented as a poster representing a part of the results from the expert discussion that focuses on human-assistant interaction and the user's perception of the system, as well as the design elements of the conversational interface and the modality choice.

3. Jovan Jeromela and Owen Conlan. "Onboarding Stages and Scrutable Interaction: How Experts Envisioned Explainability in Proactive Time Management Assistants". *Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23)*. 2023.

   Published as a full paper and presented in the main track of the conference. The work presents the final analysis of the expert discussion, with an emphasis on the statements regarding the envisioned roles of explanations and scrutability. The paper also delves into the proposed idea of stages as explicit licences for data use tied to specific IPA functionalities.

4. Jovan Jeromela and Owen Conlan. "Devising Scrutable User Models for Time Management Assistants". *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24)*. 2024.

   This workshop paper was presented at the Explainable User Modeling (ExUM 2024), and published in the adjunct conference proceedings of UMAP '24. The paper synthesises our prior research, especially our reanalysis of prior user studies and the expert discussion, into a set of six interaction challenges and three modelling challenges for scrutable IPAs. The paper also presents preliminary implementation work on SIPA4TM and the then-developing user study protocol.

5. Jovan Jeromela, Matt Murtagh White, Jasmina Gajcin, Alok Debnath, Dipto Barman, and Owen Conlan. " 'As Long as It Does What I Want, I'd Be Happy to Trust It': Exploring User Perspectives on a Scrutable Assistant for Time Management". *Unpublished work currently in review for UMAP '26*. 2026.

   This full conference paper, just accepted for publication at UMAP '26, presents a concise overview of SIPA4TM, the realisation and evaluation of our main user study (presented in Chapter 5), and outlines the prospects paradigm and our scrutability principles. The paper emphasises the importance of scrutability for the UMAP community and highlights the main findings regarding it from the thematic analysis of focus group discussions. The paper concludes by arguing for a prospects-based view of scrutability of IPAs.

Moreover, during the course of work on this dissertation, Jeromela has also contributed to works in domains tangential to scrutability, namely human-centric explanations of black-box systems and the utility of AI-assisted content retrieval. All publications, along with their abstracts, are listed in Appendix A5.

Of these works, two publications are on the generation of human-centric explanations in reinforcement learning, and one not-yet-published work investigates the use of LLMs and knowledge graphs in information retrieval through mixed-method user studies. In the period, Jeromela

[2023] has also completed his master's dissertation on context-based tweet engagement prediction, which discusses the role of personalisation in recommender systems. Finally, Jeromela has co-founded and led a workshop on the role of context representation in user modelling (co-located with UMAP '23 and UMAP '24) and an upcoming workshop on communicating uncertainty to foster realistic expectations via human-centred design (set to take place alongside IUI '26).

## 8.3 Limitations and Future Work

As the scrutability of proactive assistants had been recognised as a research question with several open challenges, we opted for long-form open-ended discussions as the primary focus of our method. This limited the number of participants and prevented us from correlating views with potentially insightful personality scores [Barman et al., 2024b, Fatahi et al., 2023]. Relatedly, our main study was based on a one-off interaction with standardised human-facilitated onboarding, where the participants role-played someone else. Further studies are needed to examine longitudinal use with real calendar data. Lastly, while codes from our thematic analysis supported defining prospects along three axes – time and effort, self-concept, and social standing – further qualitative and quantitative studies are needed to investigate the appropriateness of the axes of prospects, as well as the utility of the concept and the proposed scrutability principles. As the propositions in this dissertation have implications beyond time management assistants, we invite future studies to investigate, ideally in a longitudinal manner, how signalling for the proposed three axes of prospects through interface and interaction design may affect user trust and proactivity adoption. Further limitations and potential directions of future work were discussed in Section 7.3.

## 8.4 Final Remarks

This dissertation sought to understand how the concept of scrutability relates to the anticipated proactive and delegation-capable IPAs. Remarkably, between the beginning of our research work and its conclusion, advancements in foundation models [Bommasani et al., 2022] enabled capabilities of agentic systems [Mihale-Wilson, 2025] reminiscent of those only speculated just years prior [Völkel et al., 2021].

Our research has revealed that user expectations have shifted dramatically in the period. While the experts in our discussion, which took place in early 2022, primarily grounded the concerns and opinions in the more bounded technological capabilities at the time, the participants in our 2025 SIPA4TM-based study already anticipated a much more capable, but inherently opaque and potentially untrustworthy assistants. The participants' concerns about the consequences

of unwanted AI actions were heightened by agentic features, such as proactively changing their calendar without explicit approval or delegating tasks in communications with others.

We propose adopting the concept of prospects from behavioural economics to explicate thinking about the user's expectations regarding the assistant's effect on the time and effort required to perform a task, self-concept actualisations, and social standing. Our findings indicate that the user's perceived prospects affect both their interaction with the assistant and their trust in it, as well as the credibility of its explanations (first major contribution, C1).

With that in mind, we propose system-level principles, guidelines, and a definition of scrutability grounded in prospects and non-prescriptive regarding the underlying technology. We maintain that our research provides strong evidence that scrutability should be integrated into the main interaction design, but that it should – nevertheless – be present and actively help the user align their perceived prospects of IPA use with the system's actual functionality and context mapping (C2). Scrutabile proactive IPAs should further empower the user to control their own (potentially implicit) user model, while signalling the extent to which that is algorithmically possible.

Excitingly, in their concurrently emerging work on *agentic interpretability*, Kim et al. [2025] at Google DeepMind, advocate for dialogue-based explanation-seeking from agentic systems, where user effort in the process is recognised and where a proactive "effort" on the machine's part is also expected. The active effort on the user's part is a defining distinction between the concepts of explainability and interpretability and the human-centric vision of a scrutable user model by Kay and Kummerfeld [2013]. Meanwhile, the emphasis on active support by the proactive IPA in the user's scrutiny is central to our reframing of scrutability (cf. Subsection 6.3.1).

We are thus eager to see how principles and the design guidelines formulated in this dissertation may be reified in future work on making the ever-more potent AI assistants more capable of actively supporting the alignment of the user's anticipations and the IPA's actual functionalities. Nevertheless, we invite further work in the area not to "reinvent the wheel" [Botti, 2025], but to instead explicitly refer to the concept of *scrutability* and to build upon the rich history of insights offered by the broader user modelling and personalisation community [De Bra, 2017], to which this dissertation has aspired to contribute.

# Bibliography

Öznur Alkan, Massimiliano Mattetti, Elizabeth M. Daly, Adi Botea, Inge Vejsbjerg, and Bart Knijnenburg. Irf: A framework for enabling users to interact with recommenders through dialogue. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021. doi: 10.1145/3449237. URL https://doi.org/10.1145/3449237.

Afra Amini, Ryan Cotterell, John Hewitt, Luca Malagutti, Clara Meister, and Tiago Pimentel. Generating text from language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 27–31, 2023.

Christos Amyrotos. Adaptive visualizations for enhanced data understanding and interpretation. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 291–297, 2021.

Terry Anderson and Julie Shattuck. Design-based research. *Educational Researcher*, 41:16–25, 2012. doi: 10.3102/0013189X11428813. URL https://journals.sagepub.com/doi/pdf/10.3102/0013189X11428813.

Solomon E Asch. Forming impressions of personality. *The journal of abnormal and social psychology*, 41(3):258, 1946.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech 2023*, pages 4489–4493, 2023. doi: 10.21437/Interspeech.2023-78.

Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. Transparent, scrutable and explainable user models for personalized recommendation, 2019. URL https://doi.org/10.1145/3331184.3331211.

Dipto Barman, Jovan Jeromela, Alok Debnath, Marloes Vredenborg, Anouk Van Kasteren, Judy Kay, and Owen Conlan. Second workshop on context representation in user modelling. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and*

*Personalization*, UMAP Adjunct '24, page 225–228, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704666. doi: 10.1145/3631700.3658528. URL https://doi.org/10.1145/3631700.3658528.

Dipto Barman, Kevin Koidl, and Owen Conlan. Discerning individual preferences for identifying and flagging misinformation on social media. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '24, page 110–119, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704338. doi: 10.1145/3627043.3659545. URL https://doi.org/10.1145/3627043.3659545.

Vince Bartle, Janice Lyu, Freesoul El Shabazz-Thompson, Yunmin Oh, Angela Anqi Chen, Yu-Jan Chang, Kenneth Holstein, and Nicola Dell. "a second voice": Investigating opportunities and challenges for interactive voice assistants to support home health aides. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517683. URL https://doi.org/10.1145/3491102.3517683.

Alejandro Bellogin, Ludovico Boratto, and Olga C. Santos, editors. *UMAP '22: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392075.

Florian Bemmann, Helena Stoll, and Sven Mayer. Privacy slider: Fine-grain privacy control for smartphones. *Proc. ACM Hum.-Comput. Interact.*, 8(MHCI), September 2024. doi: 10.1145/3676519. URL https://doi.org/10.1145/3676519.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Ariel Bereslavsky. jonckheere-test: A python implementation of the jonckheere-terpstra test for ordered alternatives. Technical report, Version: 0.1.1, 2025. URL https://pypi.org/project/jonckheere-test. Last access: 2026-01-22.

Pauline M. Berry, Thierry Donneau-Golencer, Khang Duong, Melinda Gervasio, Bart Peintner, and Neil Yorke-Smith. Evaluating intelligent knowledge systems: experiences with a

user-adaptive assistant agent. *Knowledge and Information Systems*, 52(2):379–409, 2017. ISSN 0219-3116. doi: 10.1007/s10115-016-1011-3. URL https://doi.org/10.1007/s10115-016-1011-3.

Zarif Bin Akhtar. From bard to gemini: An investigative exploration journey through google's evolution in conversational ai and generative ai. *Computing and Artificial Intelligence*, 2 (1):1378, Jun. 2024. doi: 10.59400/cai.v2i1.1378. URL https://ojs.acad-pub.com/index.php/CAI/article/view/1378.

Kaitlin Boeckl and Naomi Lefkovitz. NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management, Version 1.0. Technical Report NIST CSWP 01162020, National Institute of Standards and Technology, Gaithersburg, MD, 2020. URL https://www.nist.gov/privacy-framework/privacy-framework.

Tom Bolton, Tooska Dargahi, Sana Belguith, Mabrook S. Al-Rakhami, and Ali Hassan Sodhro. On the security and privacy challenges of virtual assistants. *Sensors*, 21(7):2312, 2021. ISSN 1424-8220. URL https://www.mdpi.com/1424-8220/21/7/2312.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Susan D Boon and John G Holmes. The dynamics of interpersonal trust: Resolving uncertainty in the face of risk. *Cooperation and prosocial behavior*, 190:211, 1991.

V. Botti. Agentic ai and multiagentic: Are we reinventing the wheel?, 2025. URL https://arxiv.org/abs/2506.01463.

Michelle Brachman, Siya Kunde, Sarah Miller, Ana Fucs, Samantha Dempsey, Jamie Jabbour, and Werner Geyer. Building appropriate mental models: What users know and want to know about an agentic ai chatbot. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 247–264, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713064. doi: 10.1145/3708359.3712071. URL https://doi.org/10.1145/3708359.3712071.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. ISSN 1478-0887. doi: 10.1191/1478088706qp063oa. URL https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa. doi: 10.1191/1478088706qp063oa.

Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, 2019. ISSN 2159-676X. doi: 10.1080/2159676X.2019.1628806. URL https://doi.org/10.1080/2159676X.2019.1628806. doi: 10.1080/2159676X.2019.1628806.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech 2023*, pages 1983–1987, 2023. doi: 10.21437/Interspeech.2023-105.

Manfred Broy, Harald Ruess, and Natarajan Shankar. It's the specification, stupid! *Communications of the ACM*, 68(12):22–24, November 2025. ISSN 0001-0782. doi: 10.1145/3736710. URL https://doi.org/10.1145/3736710.

Peter Brusilovsky. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11 (1):87–110, 2001. ISSN 1573-1391. doi: 10.1023/A:1011143116306. URL https://doi.org/10.1023/A:1011143116306.

Peter Brusilovsky. User control in adaptive information access. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 357–360, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392075. doi: 10.1145/3503252.3533722. URL https://doi.org/10.1145/3503252.3533722.

Peter Brusilovsky. Ai in education, learner control, and human-ai collaboration. *International Journal of Artificial Intelligence in Education*, 34(1):122–135, 2024. ISSN 1560-4306. doi: 10.1007/s40593-023-00356-z. URL https://doi.org/10.1007/s40593-023-00356-z.

Robert B. Burns. *The Self-concept*, pages 171–196. Springer Netherlands, Dordrecht, 1980. ISBN 978-94-010-9329-3. doi: 10.1007/978-94-010-9329-3_11. URL https://doi.org/10.1007/978-94-010-9329-3_11.

Colin Camerer, George Loewenstein, and Martin Weber. The curse of knowledge in economic settings: An experimental analysis. *Journal of political Economy*, 97(5):1232–1254, 1989.

Giovanni Campagna, Rakesh Ramesh, Silei Xu, Michael Fischer, and Monica S. Lam. Almond: The architecture of an open, crowdsourced, privacy-preserving, programmable virtual assistant. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 341–350, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052562. URL https://doi.org/10.1145/3038912.3052562.

Owen Chambers, Robin Cohen, Maura R. Grossman, and Queenie Chen. Creating a user model to support user-specific explanations of ai systems. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, page 163–166, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392327. doi: 10.1145/3511047.3537678. URL https://doi.org/10.1145/3511047.3537678.

Longfeng Chen, Zheng Xiao, Juyuan Wang, Zeyu Huang, Yawen Zeng, and Jin Xu. Hear: A holistic extraction and agentic reasoning framework for document understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 14376–14382, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3761999. URL https://doi.org/10.1145/3746027.3761999.

Zina Chkirbene, Ridha Hamila, Ala Gouissem, and Unal Devrim. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234, Dec 2024. doi: 10.1109/HONET63146.2024.10822885.

Sadia Sultana Chowa, Riasad Alvi, Subhey Sadi Rahman, Md Abdur Rahman, Mohaimenul Azam Khan Raiaan, Md Rafiqul Islam, Mukhtar Hussain, and Sami Azam. From language to action: a review of large language models as autonomous agents and tool users. *Artificial Intelligence Review*, 59(2):71, 2026. ISSN 1573-7462. doi: 10.1007/s10462-025-11471-9. URL https://doi.org/10.1007/s10462-025-11471-9.

H. Chung, M. Iorga, J. Voas, and S. Lee. "alexa, can i trust you?". *Computer*, 50(9):100–104, 2017. ISSN 1558-0814. doi: 10.1109/MC.2017.3571053.

Herbert H. Clark and Susan E. Brennan. *Grounding in communication*, pages 127–149. American Psychological Association, Washington, DC, US, 1991. ISBN 1-55798-121-3 (Hardcover). doi: 10.1037/10096-006.

Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. The state of speech in hci: Trends, themes and challenges. *Interacting with Computers*, 31 (4):349–371, 2019a. ISSN 0953-5438.

Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. The state of speech in hci: Trends, themes and challenges. *Interacting with Computers*, 31 (4):349–371, 2019b. ISSN 0953-5438. doi: 10.1093/iwc/iwz016. URL https://dx.doi.org/10.1093/iwc/iwz016.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019c.

Jessica Colnago, Lorrie Faith Cranor, and Alessandro Acquisti. Is there a reverse privacy paradox? an exploratory analysis of gaps between privacy perspectives and privacy-seeking behaviors. In *Proceedings on Privacy Enhancing Technologies Symposium*, pages 455–476, 2023.

Jason A Colquitt, Brent A Scott, and Jeffery A LePine. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology*, 92(4):909, 2007. ISSN 1939-1854.

Owen Conlan. *The Multi-Model, Metadata Driven Approach to Personalised eLearning Services*. Doctoral dissertation, School of Computer Science and Statistics, Trinity College Dublin, 2005.

Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, Milton Rosenberg, et al. Building explainable artificial intelligence systems. In *AAAI*, pages 1766–1773, 2006.

Melchiorre Cosseddu, Alessandro Giuliani, Marco Manolo Manca, Marco Pilloni, Alessandro Sebastian Podda, and Sandro Gabriele Tiddia. Explainable knowledge access: Recursive and rerank-based rag for interpretable qa. In Francesco Marcelloni, Kurosh Madani, Niki van Stein, and Joaquim Filipe, editors, *Computational Intelligence*, pages 213–231, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-15638-9.

Angelo Costa, Paulo Novais, and Vicente Julian. *A Survey of Cognitive Assistants*, pages 3–16. Springer International Publishing, 2018. ISBN 1868-4394. doi: 10.1007/978-3-319-62530-0_1. URL https://dx.doi.org/10.1007/978-3-319-62530-0_1.

Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. "what can i help you with?": infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350754. doi: 10.1145/3098279.3098539. URL https://doi.org/10.1145/3098279.3098539.

Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 2382–2393, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025780. URL https://doi.org/10.1145/3025453.3025780.

Lee J. Cronbach. Test "reliability": Its meaning and determination. *Psychometrika*, 12(1): 1–16, 1947. doi: 10.1007/BF02289289.

Marek Czarkowski. *A scrutable adaptive hypertext*. Thesis, School of Information Technologies, University of Sydney, 2006. URL http://hdl.handle.net/2123/10206. A scrutable adaptive hypertext.

Abhisek Dash, Soumi Das, Elisabeth Kirsten, Qinyuan Wu, Sai Keerthana Karnam, Krishna P. Gummadi, Thorsten Holz, Muhammad Bilal Zafar, and Savvas Zannettou. The algorithmic self-portrait: Deconstructing memory in chatgpt, 2026. URL https://arxiv.org/abs/2602.01450.

Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, 147:113193, 2020. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.113193. URL https://www.sciencedirect.com/science/article/pii/S0957417420300191.

Paul De Bra. After twenty-five years of user modeling and adaptation...what makes us umap? In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, page 1, New York, NY, USA, 2017. Association for Computing Machinery.

ISBN 9781450346351. doi: 10.1145/3079628.3079662. URL https://doi.org/10.1145/3079628.3079662.

Matt Dennis, Judith Masthoff, and Chris Mellish. Adapting performance feedback to a learner's conscientiousness. In Judith Masthoff, Bamshad Mobasher, Michel C. Desmarais, and Roger Nkambou, editors, *User Modeling, Adaptation, and Personalization*, pages 297–302, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-31454-4.

Morton Deutsch. Trust and suspicion. *Journal of Conflict Resolution*, 2(4):265–279, 1958. doi: 10.1177/002200275800200401. URL https://doi.org/10.1177/002200275800200401.

Chris H.Q. Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, Jan 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2008.277.

Paul Dourish and Ken Anderson. Collective information practice: Exploring privacy and security as social and cultural phenomena. *Human–Computer Interaction*, 21(3):319–342, 2006. doi: 10.1207/s15327051hci2103_2. URL https://www.tandfonline.com/doi/abs/10.1207/s15327051hci2103_2.

Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368254. doi: 10.1145/3338286.3340116. URL https://doi.org/10.1145/3338286.3340116.

Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. A survey on the optimization of large language model-based agents. *ACM Comput. Surv.*, 58(9), February 2026. ISSN 0360-0300. doi: 10.1145/3789261. URL https://doi.org/10.1145/3789261.

Robin Ian MacDonald Dunbar. *Grooming, gossip, and the evolution of language*. Harvard University Press, 1996.

Matthew W Easterday, Daniel Rees Lewis, and Elizabeth M Gerber. Design-based research process: Problems, phases, and applications. In *Proceedings of the 2014 International Conference of the Learning Sciences*, volume 1, pages 317–324. International Society of the Learning Sciences, Jun-2014 2014. URL https://repository.isls.org//handle/1/1130.

Justin Edwards, He Liu, Tianyu Zhou, Sandy J. J. Gould, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. Multitasking with alexa: how using intelligent personal assistants impacts language-based primary task performance, 2019. URL https://doi.org/10.1145/3342775.3342785.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html. Last access: 2026-02-23.

KA Ericsson and HA Simon. Protocol analysis: Verbal reports as data, ed. rev, 1993.

Somayeh Fatahi, Mina Mousavifar, and Julita Vassileva. Investigating the effectiveness of persuasive justification messages in fair music recommender systems for users with different personality traits. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, page 66–77, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399326. doi: 10.1145/3565472.3592958. URL https://doi.org/10.1145/3565472.3592958.

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28 (4):689–707, 2018. ISSN 1572-8641. doi: 10.1007/s11023-018-9482-5. URL https://doi.org/10.1007/s11023-018-9482-5.

Mark C Fox, K Anders Ericsson, and Ryan Best. Do procedures for verbal reporting of thinking have to be reactive? a meta-analysis and recommendations for best reporting methods. *Psychological bulletin*, 137(2):316, 2011.

Kieran Fraser. *ETHOS-Push An Approach toward Empathetic Orchestration of Scrutable Push*. Doctoral dissertation, School of Computer Science and Statistics, Trinity College Dublin, 2021.

Jasmina Gajcin, Jovan Jeromela, and Ivana Dusparic. Semifactual explanations for reinforcement learning. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, HAI '24, page 167–175, New York, NY, USA, 2024. Association for

Computing Machinery. ISBN 9798400711787. doi: 10.1145/3687272.3688324. URL https://doi.org/10.1145/3687272.3688324.

Jasmina Gajcin, Jovan Jeromela, and Ivana Dusparic. Towards personalised and user-friendly counterfactual sequences for failure correction. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, page 228–233, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713996. doi: 10.1145/3708319.3733653. URL https://doi.org/10.1145/3708319.3733653.

Jasmina Gajcin, Jovan Jeromela, Joel Wester, Sarah Schömbs, Styliani Kleanthous, Karthikeyan Natesan Ramamurthy, Hanna Hauptmann, and Rifat Mehreen Amin. Cure2026: Communicating uncertainty to foster realistic expectations via human-centered design. In *Companion Proceedings of the 31st International Conference on Intelligent User Interfaces*, IUI '26 Companion, page 243–247, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400719851. doi: 10.1145/3742414.3794961. URL https://doi.org/10.1145/3742414.3794961.

Owen Gajcin. *Counterfactual and Semifactual Explanations for Reinforcement Learning*. Doctoral dissertation, School of Computer Science and Statistics, Trinity College Dublin, 2025.

Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, Reading, MA, 1994. ISBN 0-201-63361-2.

GDPR. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016. URL http://data.europa.eu/eli/reg/2016/679/oj. Last access: 2026-02-15.

Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025. ISSN 2075-4698.

Panagiotis Germanakos, Monika Juhasz, Aparna Kongot, Devashree Marathe, and Dimitris Sacharidis, editors. *Human-Centered AI: An Illustrated Scientific Quest*. Human–Computer Interaction Series. Springer Cham, 1 edition, 2025. ISBN 978-3-031-61374-6. doi: 10.1007/978-3-031-61375-3.

James J Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Houghton Mifflin Harcourt, 1979.

Ken Gilhooly and Caroline Green. Protocol analysis: theoretical background. *Handbook of qualitative research methods for psychology and the social sciences*, 14:43–54, 1996.

Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, page 227–236, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939876. doi: 10.1145/1378773.1378804. URL https://doi.org/10.1145/1378773.1378804.

Shreepriya Gonzalez-Jimenez, Danilo Gallo, Ricardo Sosa, Eduardo Benitez Sandoval, Tommaso Colombino, and Maria Antonietta Grasso. A decision support design framework for selecting a robotic interface. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, HAI '22, page 104–113, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393232. doi: 10.1145/3527188.3561913. URL https://doi.org/10.1145/3527188.3561913.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org. Last access: 2026-02-15.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), aug 2018. ISSN 0360-0300.

Ambreen Hanif, Amin Beheshti, Boualem Benatallah, Xuyun Zhang, Habiba, EuJin Foo, Nasrin Shabani, and Maryam Shahabikargar. A comprehensive survey of explainable artificial intelligence (xai) methods: Exploring transparency and interpretability. In Feng Zhang, Hua Wang, Mahmoud Barhamgi, Lu Chen, and Rui Zhou, editors, *Web Information Systems Engineering – WISE 2023*, pages 915–925, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-99-7254-8.

Margaret C Harrell and Melissa Bradley. Data collection methods: Semi-structured interviews and focus groups. ., 2009.

Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. Trirank: Review-aware explainable recommendation by modeling aspects, 2015. URL https://doi.org/10.1145/2806416.2806504.

Eelco Herder and Sven Herden. Context-dependent use of authority and empathy in lifestyle advices given by persuasive voice assistants. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 132–139, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398916. doi: 10.1145/3563359.3596985. URL https://doi.org/10.1145/3563359.3596985.

Eelco Herder and Judith Masthoff. *User Modeling in Human-Centered AI*, pages 475–491. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-61375-3. doi: 10.1007/978-3-031-61375-3_6. URL https://doi.org/10.1007/978-3-031-61375-3_6.

Eelco Herder and Jouke Staring. Digital junkfood on social media: To each their own poison. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, HT '24, page 126–135, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705953. doi: 10.1145/3648188.3678163. URL https://doi.org/10.1145/3648188.3678163.

Eelco Herder, Laura Stojko, Jannis Strecker, Thomas Neumayr, Enes Yigitbas, and Mirjam Augstein. Towards new realities: implications of personalized online layers in our daily lives. *i-com*, 23(2):221–229, 2024. doi: doi:10.1515/icom-2024-0017. URL https://doi.org/10.1515/icom-2024-0017.

Morten Hertzum and Kasper Hornbæk. Frustration: Still a common user experience. *ACM Trans. Comput.-Hum. Interact.*, 30(3), June 2023. ISSN 1073-0516. doi: 10.1145/3582432. URL https://doi.org/10.1145/3582432.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. Conditional probing: measuring usable information beyond a baseline. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.122. URL https://aclanthology.org/2021.emnlp-main.122/.

John Hewitt, Robert Geirhos, and Been Kim. We can't understand ai using our existing vocabulary, 2025. URL https://arxiv.org/abs/2502.07586.

Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570. URL https://doi.org/10.1177/0018720814547570. PMID: 25875432.

Charles L Isbell and Jeffrey S Pierce. An ip continuum for adaptive interface design. In *Proc. of HCI International*, page 10, 2005.

Eduardo Islas-Cota, J. Octavio Gutierrez-Garcia, Christian O. Acosta, and Luis-Felipe Rodríguez. A systematic review of intelligent assistants. *Future Generation Computer Systems*, 128:45–62, 2022. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2021.09.035. URL https://www.sciencedirect.com/science/article/pii/S0167739X21003812.

Anna A Ivanova, John Hewitt, and Noga Zaslavsky. Probing artificial neural networks: insights from neuroscience. *arXiv preprint arXiv:2104.08197*, 2021.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Jovan Jeromela. Scrutability of intelligent personal assistants. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 335–340, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392075. doi: 10.1145/3503252.3534355. URL https://doi.org/10.1145/3503252.3534355.

Jovan Jeromela. Context-based tweet engagement prediction. Diploma thesis, TU Wien Factulty of Informatics, 2023. URL http://hdl.handle.net/20.500.12708/177668.

Jovan Jeromela and Owen Conlan. Voicing suggestions and enabling reflection: Results of an expert discussion on proactive assistants for time management. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9798400700149. doi: 10.1145/3571884.3604317. URL https://doi.org/10.1145/3571884.3604317.

Jovan Jeromela and Owen Conlan. Onboarding stages and scrutable interaction: How experts envisioned explainability in proactive time management assistants. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, HAI '23, page 306–315, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400708244. doi: 10.1145/3623809.3623851. URL https://doi.org/10.1145/3623809.3623851.

Jovan Jeromela and Owen Conlan. Devising scrutable user models for time management assistants. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation*

*and Personalization*, UMAP Adjunct '24, page 250–255, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704666. doi: 10.1145/3631700.3665182. URL https://doi.org/10.1145/3631700.3665182.

Jovan Jeromela, Dipto Barman, Hassan Zaal, Alok Debnath, Awais Akbar, Judy Kay, and Owen Conlan. 1st workshop on context representation in user modelling. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 174–176, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398916. doi: 10.1145/3563359.3595624. URL https://doi.org/10.1145/3563359.3595624.

Jovan Jeromela, Matt Murtagh White, Jasmina Gajcin, Alok Debnath, Dipto Barman, and Owen Conlan. "as long as it does what i want, i'd be happy to trust it": Exploring user perspectives on a scrutable assistant for time management. In *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '26, page 10, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 979-8-4007-2311-7. doi: 10.1145/3774935.3806186. URL https://doi.org/10.1145/3774935.3806186.

A. R. Jonckheere. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41 (1/2):133–145, 1954. ISSN 00063444. URL http://www.jstor.org/stable/2333011.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914185.

Victor Kaptelinin and Bonnie Nardi. Affordances in hci: toward a mediated action perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 967–976, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi: 10.1145/2207676.2208541. URL https://doi.org/10.1145/2207676.2208541.

Judy Kay. The um toolkit for cooperative user modelling. *User Modeling and User-Adapted Interaction*, 4(3):149–196, 1994. ISSN 1573-1391. doi: 10.1007/BF01100243. URL https://doi.org/10.1007/BF01100243.

Judy Kay. *A scrutable user modelling shell for user-adapted interaction*. Doctoral dissertation, Basser Department of Computer Science, Faculty of Science, University of Sydney, 1999.

Judy Kay. Scrutability, control and learner models: Foundations for learner-centred design in aied. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, pages 3–8, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-78270-2.

Judy Kay. The case for scrutable, personal, long-term user models for information retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, page 1–2, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450380553. doi: 10.1145/3406522.3444755. URL https://doi.org/10.1145/3406522.3444755.

Judy Kay and Bob Kummerfeld. Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Trans. Interact. Intell. Syst.*, 2(4), jan 2013. ISSN 2160-6455. doi: 10.1145/2395123.2395129. URL https://doi.org/10.1145/2395123.2395129.

Judy Kay and Bob Kummerfeld. From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, 50(6):2871–2884, 2019. doi: https://doi.org/10.1111/bjet.12878. URL https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12878.

Judy Kay, Bob Kummerfeld, and Piers Lauder. Personis: A server for user models. In Paul De Bra, Peter Brusilovsky, and Ricardo Conejo, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 203–212, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-47952-9.

Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022. ISSN 2666-920X.

Been Kim, John Hewitt, Neel Nanda, Noah Fiedel, and Oyvind Tafjord. Because we have llms, we can and should pursue agentic interpretability. *arXiv preprint arXiv:2506.12152*, 2025.

Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F. Welch. Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 105–114, Oct 2018. doi: 10.1109/ISMAR.2018.00039.

Andreas M Klein, Jana Deutschländer, Kristina Kölln, Maria Rauschenberger, and Maria José Escalona. Exploring the context of use for voice user interfaces: Toward context-dependent user experience quality testing. *Journal of Software: Evolution and Process*, 36(7):e2618, 2024.

Kevin Koidl and Owen Conlan. Engineering information systems towards facilitating scrutable and configurable adaptation. In *Adaptive Hypermedia and Adaptive Web-Based Systems: 5th International Conference, AH 2008, Hannover, Germany, July 29-August 1, 2008. Proceedings 5*, pages 405–409. Springer, 2008.

Spyros Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134, 2017. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2015.07.002. URL `https://www.sciencedirect.com/science/article/pii/S0167404815001017`.

Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 107–116, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368612. doi: 10.1145/3340631.3394840. URL `https://doi.org/10.1145/3340631.3394840`.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1–10, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi: 10.1145/2207676.2207678. URL `https://doi.org/10.1145/2207676.2207678`.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, Sep. 2013. doi: 10.1109/VLHCC.2013.6645235.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, page 126–137, New

York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333061. doi: 10.1145/2678025.2701399. URL https://doi.org/10.1145/2678025.2701399.

Bob Kummerfeld and Judy Kay. User modeling for the internet of things. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, page 367–368. Association for Computing Machinery, 2017. doi: 10.1145/3079628.3079658. URL https://dl.acm.org/doi/pdf/10.1145/3079628.3079658.

Veton Këpuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 2018.

Concha Labra and Olga C. Santos. Exploring cognitive models to augment explainability in deep knowledge tracing. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 220–223, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398916. doi: 10.1145/3563359.3597384. URL https://doi.org/10.1145/3563359.3597384.

David Leake and David Mcsherry. Introduction to the special issue on explanation in case-based reasoning. *The Artificial Intelligence Review*, 24(2):103–108, 10 2005. URL https://elib.tcd.ie/login?url=https://www.proquest.com/scholarly-journals/introduction-special-issue-on-explanation-case/docview/197989620/se-2. Copyright - Springer 2005; Last updated - 2022-11-21.

John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392. URL https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392.

Yaniv Leviathan and Yossi Matias. Google duplex: An ai system for accomplishing real-world tasks over the phone, 2018. URL https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html. Last access: 2021-11-14.

RoyJ Lewicki and Barbara Benedict Bunker. Developing and maintaining trust in organizations. *Frontiers of theory and research*, page 114, 1996. ISSN 0803957394.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu. Personal llm agents: Insights and survey about the capability, efficiency and security, 2024. URL https://arxiv.org/abs/2401.05459.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018. ISSN 1542-7730.

Jing Liu, Ahmad Johari Bin Sihes, and Ye Lu. How do generative artificial intelligence (ai) tools and large language models (llms) influence language learners' critical thinking in efl education? a systematic review. *Smart Learning Environments*, 12(1):48, 2025. ISSN 2196-7091. doi: 10.1186/s40561-025-00406-0. URL https://doi.org/10.1186/s40561-025-00406-0.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Helena Löfström, Karl Hammar, and Ulf Johansson. A meta survey of quality evaluation criteria in explanation methods. In Jochen De Weerdt and Artem Polyvyanyy, editors, *Intelligent Information Systems*, pages 55–63, Cham, 2022. Springer International Publishing. ISBN 978-3-031-07481-3.

Tania Lombrozo. Learning by thinking in natural and artificial minds. *Trends in Cognitive Sciences*, 28(11):1011–1022, 2024. ISSN 1364-6613. doi: 10.1016/j.tics.2024.07.007. URL https://doi.org/10.1016/j.tics.2024.07.007. doi: 10.1016/j.tics.2024.07.007.

Robyn Longhurst. Semi-structured interviews and focus groups. *Key methods in geography*, 3(2):143–156, 2003.

Ewa Luger and Abigail Sellen. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5286–5297, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858288. URL https://doi.org/10.1145/2858036.2858288.

Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. Social boundaries for personal agents in the interpersonal space of the home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376311. URL https://doi.org/10.1145/3313831.3376311.

Alexander Maedche, Stefan Morana, Silvia Schacht, Dirk Werth, and Julian Krumeich. Advanced user assistance systems. *Business & Information Systems Engineering*, 58(5):

367–370, Oct 2016. ISSN 1867-0202. doi: 10.1007/s12599-016-0444-2. URL https://doi.org/10.1007/s12599-016-0444-2.

Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995. ISSN 0363-7425.

Bill McEvily and Marco Tortoriello. Measuring trust in organisational research: Review and recommendations. *Journal of Trust Research*, 1(1):23–63, 2011. doi: 10.1080/21515581.2011.552424. URL https://doi.org/10.1080/21515581.2011.552424.

Claire McNamara, Lucy Hederman, and Declan O'Sullivan. Mind the gap: LLMs, competency questions, and the non-technical user in the humanities domain. In *First International Workshop on Users and Knowledge Graphs @ SEMANTiCS 2025*, 2025. URL https://openreview.net/forum?id=gDzqvLCrgc.

Aman Mehta. When agents disagree with themselves: Measuring behavioral consistency in llm-based agents, 2026. URL https://arxiv.org/abs/2602.11619.

Christian Meurisch, Maria-Dorina Ionescu, Benedikt Schmidt, and Max Mühlhäuser. Reference model of next-generation digital personal assistant: integrating proactive behavior. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 149–152, 2017.

Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. Exploring user expectations of proactive ai systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4):Article 146, 2020. doi: 10.1145/3432193. URL https://doi.org/10.1145/3432193.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Erik Miehling, Karthikeyan Natesan Ramamurthy, Kush R. Varshney, Matthew Riemer, Djallel Bouneffouf, John T. Richards, Amit Dhurandhar, Elizabeth M. Daly, Michael Hind, Prasanna Sattigeri, Dennis Wei, Ambrish Rawat, Jasmina Gajcin, and Werner Geyer. Agentic ai needs a systems theory, 2025. URL https://arxiv.org/abs/2503.00237.

Cristina A. Mihale-Wilson. From whether to how much: A multi-tiered perspective on delegation to agentic is. In *Proceedings of the 32nd International Conference on Information Systems (ICIS)*, 2025. URL https://aisel.aisnet.org/icis2025/hti/hti/32.

Martijn Millecamp, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. Controlling spotify recommendations: Effects of personal characteristics on music recommender user interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, page 101–109, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355896. doi: 10.1145/3209219.3209223. URL https://doi.org/10.1145/3209219.3209223.

Swati Mishra and Jeffrey M Rzeszotarski. Human expectations and perceptions of learning in machine teaching. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, page 13–24, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399326. doi: 10.1145/3565472.3595612. URL https://doi.org/10.1145/3565472.3595612.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6129–6139, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736570. URL https://doi.org/10.1145/3711896.3736570.

Roger K. Moore. *Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction*, pages 281–291. Springer Singapore, Singapore, 2017. ISBN 978-981-10-2585-3. doi: 10.1007/978-981-10-2585-3_22. URL https://doi.org/10.1007/978-981-10-2585-3_22.

Roger K. Moore and Mauro Nicolao. Toward a needs-based architecture for 'intelligent' communicative agents: Speaking with intention. *Frontiers in Robotics and AI*, 4, 2017. ISSN 2296-9144. doi: 10.3389/frobt.2017.00066. URL https://www.frontiersin.org/articles/10.3389/frobt.2017.00066.

D. Morley and K. Myers. The spark agent framework. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pages 714–721, July 2004.

Nikola Mrkšić. Our voice assistant spoke to google duplex. here's what happened..., 2020. URL https://www.polyai.com/our-voice-assistant-spoke-to-google-duplex-heres-what-happened/. Last access: 2021-11-21.

Christine Murad and Cosmin Munteanu. "i don't know what you're talking about, halexa": The case for voice user interface guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, New York, NY, USA, 2019. Association for

Computing Machinery. ISBN 9781450371872. doi: 10.1145/3342775.3342795. URL https://doi.org/10.1145/3342775.3342795.

Cataldo Musto, Fedelucio Narducci, Marco Polignano, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. Towards queryable user profiles: Introducing conversational agents in a platform for holistic user modeling. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, page 213–218, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379502. doi: 10.1145/3386392.3399298. URL https://doi.org/10.1145/3386392.3399298.

Joseph W Newbold, Jacob Luton, Anna L Cox, and Sandy JJ Gould. Using nature-based soundscapes to support task performance and mood. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2802–2809, 2017.

Donald A. Norman. *The Psychology of Everyday Things*. Basic Books, 1988. ISBN 9780385267748. URL https://books.google.com/books?id=OlNSRAAACAAJ.

Michael I Norton, Daniel Mochon, and Dan Ariely. The ikea effect: When labor leads to love. *Journal of consumer psychology*, 22(3):453–460, 2012.

Daniel E O'Leary. Google's duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management*, 26(1):46–53, 2019. ISSN 1055-615X. doi: 10.1002/isaf.1443.

Danielle Marie Olson and Yan Xu. Building trust over time in human-agent relationships. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 193–201, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386203. doi: 10.1145/3472307.3484178. URL https://doi.org/10.1145/3472307.3484178.

Jorge A. Osorio, Michel R.V. Chaudron, and Werner Heijstek. Moving from waterfall to iterative development: An empirical evaluation of advantages, disadvantages and risks of rup. In *2011 37th EUROMICRO Conference on Software Engineering and Advanced Applications*, pages 453–460, Aug 2011. doi: 10.1109/SEAA.2011.69.

Daphna Oyserman, Kristen Elmore, and George Smith. Self, self-concept, and identity. In June Price Tangney and Mark R. Leary, editors, *Handbook of Self and Identity*, pages 69–104. Guilford Press, New York, NY, 2 edition, 2012.

Debajyoti Pal, Mohammad Dawood Babakerkhell, and Xiangmin Zhang. Exploring the determinants of users' continuance usage intention of smart voice assistants. *IEEE Access*, 9: 162259–162275, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3132399.

Leysia Palen and Paul Dourish. Unpacking "privacy" for a networked world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 129–136, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136307. doi: 10.1145/642611.642635. URL https://doi.org/10.1145/642611.642635.

Neil Peirce, Owen Conlan, and Vincent Wade. Adaptive educational games: Providing non-invasive personalised learning experiences. In *2008 second IEEE international conference on digital game and intelligent toy enhanced learning*, pages 28–35. IEEE, 2008.

Xuan Lam Pham, Thao Pham, Quynh Mai Nguyen, Thanh Huong Nguyen, and Thi Thu Huong Cao. Chatbot as an intelligent personal assistant for mobile language learning. In *Proceedings of the 2018 2nd International Conference on Education and E-Learning*, pages 16–21, 2018.

Wolfgang Pohl. Logic-based representation and reasoning for user modeling shell systems. *User Modeling and User-Adapted Interaction*, 9(3):217–282, 1999. ISSN 1573-1391. doi: 10.1023/A:1008325713804. URL https://doi.org/10.1023/A:1008325713804.

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. *Voice Interfaces in Everyday Life*, page Paper 640. Association for Computing Machinery, 2018. ISBN 9781450356206. doi: 10.1145/3173574.3174214. URL https://doi.org/10.1145/3173574.3174214.

Jonathan Potter and Margaret Wetherell. *Discourse and social psychology: Beyond attitudes and behaviour*. Sage, 1987.

Suraj Prasai, Mengnan Du, Ying Zhang, and Fan Yang. Knowthyself: An agentic assistant for llm interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48): 41661–41663, Mar. 2026. doi: 10.1609/aaai.v40i48.42373. URL https://ojs.aaai.org/index.php/AAAI/article/view/42373.

Nora Ptakauskaite, Anna L Cox, and Nadia Berthouze. Knowing what you're doing or knowing what to do: how stress management apps support reflection and behaviour change. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

Emilee Rader and Rebecca Gray. Understanding user beliefs about algorithmic curation in the facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 173–182, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702174. URL https://doi.org/10.1145/2702123.2702174.

Sunayana Rane, Alexander Ku, Jason Michael Baldridge, Ian Tenney, Thomas L. Griffiths, and Been Kim. Can generative multimodal models count to ten? In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. URL https://openreview.net/forum?id=YyXCc2quWJ.

Lauge Baungaard Rasmussen. From human-centred to human-context centred approach: looking back over 'the hills', what has been gained and lost? *AI & SOCIETY*, 21 (4):471–495, 2007. ISSN 1435-5655. doi: 10.1007/s00146-007-0088-3. URL https://doi.org/10.1007/s00146-007-0088-3.

William Revelle and David M Condon. Reliability from $\alpha$ to $\omega$: A tutorial. *Psychological assessment*, 31(12):1395, 2019.

Francesco Ricci, Lior Rokach, and Bracha Shapira. *Recommender Systems: Techniques, Applications, and Challenges*, pages 1–35. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_1. URL https://doi.org/10.1007/978-1-0716-2197-4_1.

Tom A. Rodden, Joel E. Fischer, Nadia Pantidi, Khaled Bachour, and Stuart Moran. At home with agents: exploring attitudes towards future smart energy infrastructures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 1173–1182, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2466152. URL https://doi.org/10.1145/2470654.2466152.

Carl Ransom Rogers et al. *A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework*, volume 3. McGraw-Hill New York, 1959.

Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, 2019. ISSN 1573-7454.

Julian B. Rotter. Generalized expectancies for interpersonal trust. *American Psychologist*, 26(5):443–452, 1971. ISSN 1935-990X(Electronic),0003-066X(Print). doi: 10.1037/h0031464.

Julian B. Rotter. Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, 35(1):1–7, 1980. ISSN 1935-990X(Electronic),0003-066X(Print). doi: 10.1037/0003-066X.35.1.1.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110273. URL https://www.sciencedirect.com/science/article/pii/S0950705123000230.

Hamid Reza Saeidnia. Welcome to the gemini era: Google deepmind and the information industry. *Library Hi Tech News*, 43(1):18–20, 12 2023. ISSN 0741-9058. doi: 10.1108/LHTN-12-2023-0214. URL https://doi.org/10.1108/LHTN-12-2023-0214.

R. Sarikaya. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 34(1):67–81, 2017. ISSN 1558-0792. doi: 10.1109/MSP.2016.2617341.

Silvia Schiaffino and Analía Amandi. User – interface agent interaction: personalization issues. *International Journal of Human-Computer Studies*, 60(1):129–148, 2004. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2003.09.003. URL https://www.sciencedirect.com/science/article/pii/S1071581903001666.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf.

David D Scholz, Johannes Kraus, and Linda Miller. Measuring the propensity to trust in automated technology: Examining similarities to dispositional trust in other humans and validation of the ptt-a scale. *International Journal of Human–Computer Interaction*, 41(2):970–993, 2025.

Lisa Schut, Nenad Tomašev, Thomas McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human–ai knowledge gap through concept discovery and transfer in alphazero. *Proceedings of the National Academy of Sciences*, 122(13):e2406675122, 2025. doi: 10.1073/pnas.2406675122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2406675122.

Emre Sezgin, Lisa K Militello, Yungui Huang, and Simon Lin. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. *Translational Behavioral Medicine*, 10(3):606–628, 2020. ISSN 1869-6716. doi: 10.1093/tbm/ibz141. URL https://academic.oup.com/tbm/article/10/3/606/5885015.

Elhadi M Shakshuki and SM Mozammal Hossain. A personal meeting scheduling agent. *Personal and ubiquitous computing*, 18(4):909–922, 2014. ISSN 1617-4909.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 38154–38180. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/77c33e6a367922d003ff102ffb92b658-Paper-Conference.pdf.

Thomas B. Sheridan and William L. Verplank. Human and computer control of undersea teleoperators. Technical Report ADA057655, Massachusetts Institute of Technology, Man-Machine Systems Laboratory, Cambridge, MA, 1978. Technical report, 15 March 1977–14 June 1978.

Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020. doi: 10.1080/10447318.2020.1741118. URL https://doi.org/10.1080/10447318.2020.1741118.

Ben Shneiderman. *Human-Centered AI*. Oxford University Press, 01 2022. ISBN 9780192845290. doi: 10.1093/oso/9780192845290.001.0001.

Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Education Limited, 6th edition, 2018. ISBN 978-1-292-15391-9.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

SRI International. 75 years of innovation: Calo (cognitive assistant that learns and organizes), 2020. URL https://medium.com/dish/

75-years-of-innovation-calo-cognitive-assistant-that-learns-and-organizes-9f176f2de05
Last access: 2020-10-03.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Hiq7lUh4Yn. Survey Certification, Expert Certification.

Yu Sun, Nicholas Jing Yuan, Yingzi Wang, Xing Xie, Kieran McDonald, and Rui Zhang. Contextual intent tracking for personal assistants, 2016. URL https://doi.org/10.1145/2939672.2939676.

Yu Sun, Nicholas Jing Yuan, Xing Xie, Kieran McDonald, and Rui Zhang. Collaborative intent prediction with real-time contextual data. *ACM Trans. Inf. Syst.*, 35(4):Article 30, 2017. ISSN 1046-8188. doi: 10.1145/3041659. URL https://doi.org/10.1145/3041659.

Emmanuel Syrmoudis, Stefan Mager, Sophie Kuebler-Wachendorff, Paul Pizzinini, Jens Grossklags, and Johann Kranz. Data portability between online services: an empirical analysis on the effectiveness of gdpr art. 20. *Proceedings on Privacy Enhancing Technologies*, 2021.

Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. Investigating users' preferences and expectations for always-listening voice assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3 (4):Article 153, 2019. doi: 10.1145/3369807. URL https://doi.org/10.1145/3369807.

Teunis J Terpstra. The asymptotic normality and consistency of kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14(3):327–333, 1952.

Nava Tintarev and Judith Masthoff. *Beyond Explaining Single Item Recommendations*, pages 711–756. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4.

Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation*

*and Personalization*, UMAP '21, page 77–87, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383660. doi: 10.1145/3450613.3456817. URL https://doi.org/10.1145/3450613.3456817.

Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173632. URL https://doi.org/10.1145/3173574.3173632.

Amrita S Tulshan and Sudhir Namdeorao Dhage. Survey on virtual assistant: Google assistant, siri, cortana, alexa. In *International symposium on signal processing and intelligent recognition systems*, pages 190–201. Springer, 2018.

Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):1039–1061, 11 1991. ISSN 0033-5533. doi: 10.2307/2937956. URL https://doi.org/10.2307/2937956.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992. ISSN 1573-0476. doi: 10.1007/BF00122574. URL https://doi.org/10.1007/BF00122574.

Piyush Vashistha, Juginder Pal Singh, Pranav Jain, and Jitendra Kumar. Raspberry pi based voice-operated personal assistant (neobot). In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 974–978. IEEE, 2019. ISBN 1728101670.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Sruthi Viswanathan, Fabien Guillot, Minsuk Chang, Antonietta Maria Grasso, and Jean-Michel Renders. Addressing hiccups in conversations with recommender systems, 2022.

Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. *Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants*, page Article 254. Association for Computing Machinery, 2021. ISBN 9781450380966. doi: 10.1145/3411764.3445536. URL https://doi.org/10.1145/3411764.3445536.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL https://doi.org/10.1007/s11704-024-40231-1.

Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. Tem: Tree-enhanced embedding model for explainable recommendation, 2018. URL https://doi.org/10.1145/3178876.3186066.

Lauren Werner, Gaojian Huang, and Brandon J. Pitts. Smart speech systems: A focus group study on older adult user and non-user perceptions of speech interfaces. *International Journal of Human–Computer Interaction*, 39(5):1149–1161, 2023. doi: 10.1080/10447318.2022.2050541. URL https://doi.org/10.1080/10447318.2022.2050541.

Brian Whitworth. Polite computing. *Behaviour & Information Technology*, 24(5):353–363, 2005. doi: 10.1080/01449290512331333700. URL https://doi.org/10.1080/01449290512331333700.

Sue Widdicombe and Robin Wooffitt. *The language of youth subcultures: Social identity in action*. Harvester/Wheatsheaf, 1995.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui. The rise and potential of large language model based agents: a survey. *Science China Information Sciences*, 68(2):121101, 2025. ISSN 1869-1919. doi: 10.1007/s11432-024-4222-0. URL https://doi.org/10.1007/s11432-024-4222-0.

Jun Xiao, Richard Catrambone, and John T Stasko. Be quiet? evaluating proactive and reactive user interface assistants. Technical report, Georgia Institute of Technology, 2003.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh*

*International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. Creating design resources to scaffold the ideation of ai concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, DIS '23, page 2326–2346, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398930. doi: 10.1145/3563657.3596058. URL https://doi.org/10.1145/3563657.3596058.

Neil Yorke-Smith, Shahin Saadati, Karen L Myers, and David N Morley. The design of a proactive personal agent for task management. *International Journal on Artificial Intelligence Tools*, 21(01):30, 2012. ISSN 0218-2130. doi: 10.1142/S0218213012500042.

Xiao Yu, Lei Liu, Xing Hu, Jacky Wai Keung, Jin Liu, and Xin Xia. Fight fire with fire: How much can we trust chatgpt on source code-related tasks? *IEEE Transactions on Software Engineering*, 50(12):3435–3453, Dec 2024. ISSN 1939-3520. doi: 10.1109/TSE.2024.3492204.

Markus Zanker, Laurens Rook, and Dietmar Jannach. Measuring the impact of online personalisation: Past, present and future. *International Journal of Human-Computer Studies*, 131: 160–168, 2019. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2019.06.006. URL https://www.sciencedirect.com/science/article/pii/S107158191930076X.

Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450397391. doi: 10.1145/3543829.3543834. URL https://doi.org/10.1145/3543829.3543834.

Minxing Zhang, Yi Yang, Roy Xie, Bhuwan Dhingra, Shuyan Zhou, and Jian Pei. Generalizability of large language model-based agents: A comprehensive survey. *ACM Comput. Surv.*, February 2026. ISSN 0360-0300. doi: 10.1145/3794858. URL https://doi.org/10.1145/3794858. Just Accepted.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372852. URL https://doi.org/10.1145/3351095.3372852.

J. Zheng, C. Shi, X. Cai, Q. Li, D. Zhang, C. Li, D. Yu, and Q. Ma. Lifelong learning of large language model based agents: A roadmap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3650546.

Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. Ux research on conversational human-ai interaction: A literature review of the acm digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501855. URL https://doi.org/10.1145/3491102.3501855.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020. ISSN 0891-2017. doi: 10.1162/coli_a_00368.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

Chris Zimmerer, Philipp Krop, Martin Fischbach, and Marc Erich Latoschik. Reducing the cognitive load of playing a digital tabletop game with a multimodal interface. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3502062. URL https://doi.org/10.1145/3491102.3502062.

# A1 Links to Further Materials

This appendix contains links to repositories with additional materials related to this dissertation, including the SIPA4TM source code, the user study protocol, all transcripts, logs, quantifications, survey responses, and the thematic analysis codebook. The last access date for all links in this appendix is 2026-03-10, except for the links to the SIPA4TM code and the SIPA4TM-anchored study, which were last visited on 2026-03-12.

**Reanalysis** of user study data conducted by Cowan et al. [2017] and Völkel et al. [2021] (Section 3.1 in Chapter 3; Appendix A2):

- The dataset published by Völkel et al. [2021]: `https://github.com/kartikn27/nlp-question-detection`
- Excel tables with selected statements, the Jupyter notebook for filtering of statements, and further findings form our reanalyses: https://gitlab.com/Jovan_NS/envisioned-ipa-scrutability

**Expert Discussion on the Anticipated Proactive IPAs** for Time Management (Section 3.2 in Chapter 3; Appendix A3):

- The data from the discussion, including the anonymised transcript, the presentation shown by the moderator, the sketched diagram, and use cases from the post-it notes: https://drive.google.com/drive/folders/1-crFEKo2jyr7d-_5OgBSLlSF7B9SASP3

**SIPA4TM** code (Chapter 4):

- Gitlab repository: `https://gitlab.com/Jovan_NS/spipa4tm/`
- The code (with some reductions) is also available in the drivers for the user study raw data, linked below.

**SIPA4TM-Based User Study**, i.e. the main user study of the dissertation (Chapter 5; Appendix A4):

- A shared Google Sheet containing the full final mapping of thinking-aloud statements of

interest: https://docs.google.com/spreadsheets/d/1NuL6VJVpBSNjkRNj9OKicnLpf3-FNfcpm0C8jeNo1XI/edit?usp=sharing

- Thematic analysis NVivo files, showcasing the progression of codes through iterations and other analysis-related materials, including the output of the collaborative workshop-style preliminary themes candidates session: `https://drive.google.com/drive/folders/1vxB0sGqi7-NbP3TX_MCfNhXZrMfMvHbW?usp=sharing`
- Fully anonymised user study data, including a backup of SIPA4TM code: `https://drive.google.com/drive/folders/14fvkx9oScqboDqrDj2Tmtaw0zrHwVNe8?usp=sharing`

**Workshops** the dissertation author co-founded and co-organised (Appendix A6):

- CRUM@UMAP '23: `https://crum-workshop.github.io/2023/`
- CRUM@UMAP '24: `https://crum-workshop.github.io/`
- CURE@IUI '26: `https://cureworkshop.github.io/cure-2026/`

# A2 Reanalysis of Prior Studies

This appendix presents selected statements from the reanalysis of data from two previously published user studies, originally presented by Cowan et al. [2017] and Völkel et al. [2021]. These can be seen in tables A2.1 and A2.2 respectively. The motivation for the reanalyses was to better understand user expectations regarding scrutability, proactivity, and delegative capabilities of intelligent personal assistants (IPAs). The main findings are presented in Section 3.1. We thank Benjamin R. Cowan and Sarah Theres Völkel for their support and permission to access and reanalyse their datasets.

Table A2.1: Selected statements from the study presented in Cowan et al. [2017]. The ID follows the example from the paper and consists of the focus group and participant number, as shown in presented transcripts.

| # | ID | Statement |
|---|---|---|
| 1 | F1P3 | I wouldn't use Siri all that often because I don't find that I trust it as such. Like even setting an alarm I would be thinking "Oh I better go and check if that actually set" |
| 2 | F2P2 | No, Siri's not a real person. She doesn't remember what I say to her, she probably logs it for data purposes but she's not going to be like, "do you remember that time you told me this thing?" So, whatever, dude. |
| 3 | F2P1 | I'd love it to just do what it's supposed to do a lot more efficiently, like you know, I'd love to be able to type a letter for me, instead of having to type it up meself, for example, that would definitely be something I'd love it to do. |
| 4 | F2P2 | it turns on in inopportune moments |
| 5 | F3P? | I would like Siri to ask me back a question, to clarify something. /. . . / I would have liked a question-answer to be more of a dialogue. So Siri would ask me a question, what kind of, "are you making this now or would you like me to create a shopping list for the ingredients?" Something like that, that would be led more, more options there |

Continued on next page

**Table A2.1 – continued from previous page**

| # | ID | Statement |
|---|---|---|
| 6 | F3P? | I think Siri can do a lot more than we use it for, but I just don't really know what to use it for. |
| 7 | F5P1 | I don't want it to keep certain information |
| 8 | F5P1 | Like scheduling meetings. I think it's a long way from that, I think it's not good enough yet. And there's a lot of things that a person can do that Siri can't, like the nuances of scheduling and privatising, things like that. |

Table A2.2: Selected (parts of) dialogues from the study presented in Völkel et al. [2021]. The ID corresponds to the name of the scenario followed by the row number (0–204) of the corresponding participant in the provided dataset. Text in *italics* are the IPAs' statements.

| # | ID | Dialogue (snippet) |
|---|---|---|
| 1 | search-5 | Assistant, at what time can I watch Tenet at my local cinema tomorrow? *Since you have a birthday to attend tomorrow at 3pm, I would recommend you to buy your ticket for 7:30pm.* Could you do that for me? I want two seats, no preference. I want some popcorn too, order 2 menus. *Ok, would you like me to remind you tomorrow about it?* Yes, please! Also, send a message to Alice and tell her I already got us the tickets. |
| 2 | search-45 | could you please book me a ticket? *yes, booking a ticket now. do you have a seating preference?* hm, one the seats in the middle, please. |
| 3 | search-105 | I would like 1 adult ticket for 12PM *All booked for you in your preferred seating area* |
| 4 | search-155 | could you please book Movie 3 for me? *How many people shall I book for?* 2 please. *It seems there are a few deluxe seats available, you booked these the last time. Shall I book these seats?* Yes please |
| 5 | volume-142 | (IPA initiates) *Sam, your music is over 70 decibels. Knowing your neighbors, as we do, they might not be happy.* You're probably right, Shirley. Ok, turn the volume down to something they might find acceptable. *How about this? 55 decibels, they should be ok with this level.* Yeah, that's fine, bloody neighbors. *Would you like me to set 55 as a general control, that you can override?* Hmmm ok, better to be a good citizen and not be selfish, i guess. |

Continued on next page

**Table A2.2 – continued from previous page**

| # | ID | Dialogue (snippet) |
|---|---|---|
| 6 | volume-169 | Will my neighbours be able to hear it? *Yes, it's a little too loud and will no doubt be disturbing them.* Ok, turn it down to a level that they cannot hear it, please. *Done, would you also like me to text them a message apologising?* You know what, that's a good idea, yes. *They said it's not a problem, and thanked you for doing so.* |
| 7 | volume-173 | Thanks for reminding but don't worry they are away at the moment *Great, Party! Would you like me to invite a few friends around?* Sure, Why not *I'm sending out a text now* |
| 8 | conv.-5 | Could you also highlight a quote for me? I would like to highlight the previous paragraph. *Done, do you want to leave a note too?* Not really, but thanks for asking, I forgot I could do that. |
| 9 | alarm-197 | (IPA initiates) *Hi there, do you remember you have a meeting at work tomorrow morning at 7:30?* Oh no I completely forgot! *I have set you an alarm for 6am, and another for 6:10am in case the first does not wake you.* Thank you. |
| 10 | joke-135 | How can I increase the value of my house? *M, you should consider updating your bathroom* Why's that V? *Because the internal fittings haven't been changed for over ten years* How do you know that V? *Know what M? That a bathroom update will increase your property value?* No. That my bathroom hasn't been updated for over ten years. *That is because I pay attention to your conversations M.* That's weird but useful V *If you would prefer me to stop listening until I am spoken to I can programme that.* Yes stop listening until I say your name *Okay, that has been programmed in. Can I help you with anything else today M?* No thanks V |
| 11 | joke-137 | (IPA asked to suggest a birthday gift) *lavender perfumes bath oils or plants* why lavender *lavender is a soother helps with sleep and relaxation* great idea then where can i buy this *let me see what area your in but all well known large pharmacys should stock this* |
| 12 | joke-153 | (IPA asked for a funny story) *What about kittens that do crazy things* No No, best not bring that up again *Suggestion put on No No list* |
| 13 | open-26 | Where can i get kite surfing lessons in a safe seaside location *How about The watersports Academy in poole* Why there *It's just 2 hours from London and the water is very shallow so you'll find it easier to learn* |

Continued on next page

**Table A2.2 – continued from previous page**

| # | ID | Dialogue (snippet) |
|---|---|---|
| 14 | open-92 | Tom, it's Emma's birthday on Tuesday, help *Do you need a card and present?* Yes *What budget?* 40,00 £ *I have a birthday card and 3 Present options in your basket on amazon, take a look.* Perfect, can I get the card and personalised chocolates. *Yes, would you like it sent to her or you? And use your default card?* |
| 15 | open-107 | Bubble, i need a meeting ASAP with Bob *Sending a message to Bob... Bob suggests 12PM* Suggest 12:30PM, at a coffee shop in town. *Okay... Bob confirms this arrangement. This has been added to your calendar* |
| 16 | open-143 | Hello Sara, I have a busy schedule with work and shopping, what would be the best time for me to exercise *Hello, Looking at your schedule the best time would be in the evening at 6pm* Thank you Sara that time would be great |

# A3 Expert Discussion

In this appendix, we reproduce the contents of the PowerPoint presentation used for the moderation of a four-point expert discussion presented in Section 3.2. With this reproduction, we aim for complete methodological transparency, which we hope will aid potential reproducibility and future research. Afterwards, we provide a consolidated list of use cases and scenarios proposed by the experts. Further materials resulting from the discussion are linked in Appendix A1.

## A3.1 Moderator's Presentation

The introductory slide was titled "Expert Discussion: Proactive Time-Management Assistants", with the following slide outlining the rest of the presentation. The remaining slides are reproduced as originally presented.

### A3.1.1 Prologue

Context → Desired Outcome → Consent Form

#### Context

Working towards "next-gen" Intelligent Personal Assistants (IPAs)

Commercial examples of IPAs include Siri, Amazon Alexa, Cortana, etc.

However, current systems are:

- Mostly reactive (only respond when prompted)
- Mostly not personalised (your Google Assistant is quite similar to mine)
- Mostly used for very simple tasks

[The slide also contains logos of the most popular IPAs at the time: Google Assistant, Samsung's Bixby, Apple's Siri, Amazon Alexa, and Microsoft's Cortana]

### Proactive Assistants

By contrast, prior works indicate that IPAs of the future are to be personalised, proactive, and capable of performing complex tasks.

Towards that goal, Meurisch et al. [2017] organised an experts' discussion where participants came up with such a model of a general-purpose proactive assistant:

[The slide then contains a reproduction of Figure 1 by Meurisch et al. [2017], which contains a "[r]eference model of digital personal assistant integrating reactive and proactive behaviors" and shows six "areas of digital personal assistance": mental ("invisible" user states), physical ("visible user states"), activity (all user actions), environment (the physical real-world surroundings of the user), social relations (including interactions with other humans), and technology ("covers the digital world of a human and the interaction with it").]

### Proactive Time Management

We invited you, experts from various related fields, so that we can discuss:

- A vision of a proactive IPA for time management:
  - (Proactive) event/reminder creation, scheduling, and change support
  - Prioritisation and sharing of time-related information
  - Aiding (and proposing) behavioural change
- Aspects elevating the assistant above a mere calendaring tool:
  - Human-centred interaction design
  - Dialogue, user understanding, and control
- The most important aspects with respect to your field of expertise
- What a model of such an assistant could look like

### Desired Outcomes

We would like to use the outcome of this discussion to:

- Come up with a preliminary **model** of the proactive time-management assistant;
- **Note the ideas, observations, and constraints** that are brought up and compare these with the state of the art;

The first concrete use of the data could be within a prospective submission for the ACM CHIIR 2023 conference.

#### Consent Forms

This is not a user study, but a semi-structured expert discussion.

However, we do require consent forms, because:

- We will record the conversation (90 mins.);
- We will photograph and scan the written materials;
- You will be acknowledged in potential future publications that use this data either by name or anonymously, as preferred.

You can withdraw your consent and ask for exemption at any time.

### A3.1.2 Discussion Preparation [25m]

Use Cases and Scenarios → Presenting Initial Ideas

#### Use Cases and Scenarios [10m]

Please use the sticky notes provided to note down your ideas for potential **use cases** and **scenarios**:

- Use case: What would be a situation in which a proactive time management assistant could be used?
- Scenario: To whom and when would a proactive time management assistant be useful?

Ideally, one use case or scenario per note (these will be shared later).

Individual work.

#### Presenting Initial Ideas [15m]

Present your ideas for use cases and scenarios to others.

As you present, also put your sticky notes in front of the table, potentially grouping similar ideas together.

We will go in order and listen to everyone's ideas first.

### A3.1.3 Discussion [65m]

Use Cases and Scenarios → Presenting Initial Ideas

#### Brainstorming Discussion [20m]

Now, let us discuss these groups of use cases and scenarios:

- What **pieces of information (about the user, context, the calendar, etc.) must be collected** for the assistant to be able to perform these functionalities?
- What **constraints (=restrictions) must be observed** by the assistant for it to be able to provide these functionalities?

#### Intermezzo: Interaction [20m]

A slight change in the discussion direction.

- What **properties (=characteristics)** should the assistant have so that you would consider using it (or recommending it to someone else)?
- How should the assistant **share information** about you with third parties?
- What modes of **communication and control** would be appropriate?

#### Finale: Sketching a Model [25m]

Collaborate towards creating a schema of the model of the proactive assistant.

- Think of previously discussed **pieces of information and constraints (=restrictions)** that ought to be modelled?
- How would the **properties (=characteristics)** and the interaction principles be facilitated through the model?

Attempt to sketch your model on the big piece of paper in front of you.

It's okay to propose alternatives: feel free to have multiple sketches!

### A3.1.4 Outlook

Future Work → Acknowledgements

#### Outlook

Thank you very much for your participation!

End of the (structured) group work.

You will be contacted with the preliminary analysis before the submission (for CHIIR2023) for (optional) final feedback and thoughts.

You will also be asked for your preference regarding the acknowledgment mode (by name or anonymously).

**Future Work**

Planned future work:

- Large scale user role-playing study (akin to Völkel et al. [2021])
- Building a prototype model/ontology
- Building a prototype system
- System evaluation in user focus groups (akin to Cowan et al. [2017])

Confluence of IPAs and scrutability: producing overarching principles

[The slide ends with contact details and is followed by several slides with full acknowledgements to cited works and the mandatory funding acknowledgements]

## A3.2 Use Cases and Scenarios

The following is the list of all use cases and scenarios the participants wrote down on post-it notes in the first phase of the discussion.

- Carve around a time to spill over
- Understanding what least disruption means
- Calendar reshuffling
- Reflective capabilities
- Planning time for longitudinal skills and knowledge development
- Attention management
- Time management for professionals who charge by the hour
- Scrutability of management
- Meeting prioritisation
- Negotiations about scheduling and reservations
- Calendar reminders
- Time planning - best ways in terms of efficiency or extra spare time

- Time management within a task
- Helping in eyes-busy, hands-busy
- Modality choice
- Proactively reminded/resolved in case of clashes
- Explainability of recommendations/choices
- Reminded at a specific time
- Routine development + behavioural tracking
- Helping professionals who manage their and others' time
- Rethinking priorities
- Healthy amount of reflection

# A4 SIPA4TM-Based User Study

Section A4.1 of this appendix provides a redeposition of the Informed Consent Form that the participants were sent per email to read before coming for the in-person participation and then to sign before commencing their study participation. Section A4.2 contains a reproduction of the content of printed materials given to participants during the study, i.e. the "Week at a Glance" and the "User Study Tasks" documents, as well as an overview of locations, priorities, and commute times which was not provided to the user directly but was discoverable by scrutinising SIPA4TM's user model. Section A4.3 contains a full reproduction of the user study protocol, as provided by the main researcher (dissertation author) to the collaborating researchers. Lastly, Section A4.4 contains a reproduction of the table with thinking-aloud statement of interests descriptions as well as themes and subthemes they were attributed to. More details about the SIPA4TM and the manner in which documents reproduced in this appendix were created and the reasoning behind it can be seen in Chapter 4. The conduct and evaluation of the study are detailed in Chapter 5.

## A4.1 Informed Consent Form

Thank you for your interest in taking part in this task-based user study. The information below is provided to inform your decision about consenting to take part in this experiment.

### Lead Researcher

Jovan Jeromela, PhD candidate, Trinity College Dublin, Ireland

### Background of Research

This research is being conducted as part of a structured PhD in the School of Computer Science and Statistics at Trinity College Dublin. The aim of the research is to investigate user interest in preferences regarding personalisation and proactivity of intelligent personal assistants (also known as virtual or digital assistants or agents and abbreviated as IPAs). Personalisation is

a term for the ability of a software (in this case, of the IPA) to adapt its services to the user's preferences, whereas proactivity means that the IPA offers suggestions or other services without being explicitly prompted beforehand.

To date, most commercial IPAs, such as Google Assistant, Amazon Alexa, and Apple's Siri, remain purely reactive (only responding to explicit requests by the user), simple (mostly used for tasks such as setting reminders or alarms), and non-personalised (meaning that they do not learn habits or preferences of the user to perform their services better).

We aim to understand design principles that would allow future, more advanced, IPAs to be more helpful, more enticing, and better aligned with the potential user's priorities and values. In particular, we are focusing on how the user's data ought to be modelled and how the scrutiny of the personalisation is to be facilitated. Inspired by prior human-computer interaction studies, this research has 3 phases:

- a survey on prior IPA and time management experience;
- an interaction with an IPA prototype; and
- a discussion on impressions and ideas in groups of 2—5.

We focus on time management assistance (calendaring, event management, reminders, etc.) as a specific use case for the envisioned IPAs. The findings from this experiment are planned to be submitted to academic journals or conferences for publication and will contribute towards the completion of the lead researcher's doctoral dissertation.

### Procedures of this Study

1. You do not require any pre-requisite for this study. The study comprises three parts
    - A short initial questionnaire
        - Your questionnaire answers will be anonymised and retained for statistical analysis (such as to help us understand how prior use of IPAs and time management tools affect interaction with the assistant and to estimate how generalisable our findings are).
    - An interaction with a prototype of a time-management assistant
        - During this interaction, you will be asked to perform five tasks and you will be given an opportunity to explore the functionalities of the assistant.
        - Your interaction with the time management prototype will be automatically logged. These logs will then be anonymised and retained for further analysis.

  - You will also be asked to think out loud while interacting with the prototype, and your voiced thoughts will be audio-recorded, to help us understand your thinking process.
  - The audio recordings will be used to create anonymised transcripts for further analysis. Seven days after your participation, the audio recordings will be deleted, while the anonymised transcripts will be retained for further analysis and potential publication.
- A semi-structured interview/discussion in groups
  - In these discussions, you will be asked to voice and discuss your opinions about the assistant and the interaction design.
  - The discussions will be audio-recorded. These recordings will then be used for the creation and validation of anonymised discussion transcripts that will be used for further analysis.
  - The audio recordings will be used to create anonymised transcripts for further analysis. Seven days after your participation, the audio recordings will be deleted, while the anonymised transcripts will be retained for further analysis and potential publication.
  - The groups will include between two and five participants.

The prototype interaction and the focus group discussion will take place in person at Trinity College Dublin and will last approximately 90 minutes. You don't have to take part in this study. It is entirely voluntary. You can change your mind about taking part in the study and opt out at any time, even if the study has started.

### Payment

For your participation, you will be compensated with a gift card in the value of 45 EUR. If you decide to take part, you are still free to change your mind at any time and without giving a reason. However, you will not be paid if you exit before the study concludes.

### Publication

The findings from this research may be published in academic journals or conferences. In particular, we aim to submit the findings to the Conference on Human Factors in Computing Systems (CHI). Alternatively, we might use (parts) of the findings of the study in conferences such as:

- Symposium on Human-Computer Interaction for Work (CHIWORK),
- Conference on Mobile Human-Computer Interaction (MobileHCI),
- Conversational User Interfaces (CUI),
- Intelligent User Interfaces Conference (IUI),
- Human-Agent Interaction Conference (HAI),
- Conference on Intelligent Virtual Agents (IVA),
- Designing Interactive Systems (DIS), and
- the Hybrid Human-Artificial, Intelligence Conference (HHAI).

Additionally, findings might be (partially) reported in potential publications in academic journals such as:

- User Modeling and User-Adapted Interaction,
- International Journal of Human Computer Studies,
- Human-Computer Interaction, and
- International Journal of Artificial Intelligence in Education.

The results shared with the broader academic community and the public in general will be anonymised and no personally identifiable information about you would be shared publicly.

### After the Study

The questionnaire answers, the interaction logs, the audio recordings, the transcripts, and all other data will be stored securely in compliance with data protection legislation and with respect to all institutional policies of Trinity College Dublin. All personally identifiable data about you (including the audio recordings) will be permanently deleted seven days after your participation. Please, feel free to contact the researchers via the email address provided below should you have any questions or wish to have your input discarded and destroyed.

### Conflicts of Interests

The lead researcher or collaborating researchers might be demonstrators or teaching assistants in courses that you might be taking at the moment or in the future. Your participation in this user study or your decision not to participate will not affect the outcome or the process of

any such course. Individual results of the study may be aggregated anonymously and research reported on aggregate results.

### Declaration

- I am 18 years or older and am competent to provide consent independently.
- I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- I agree that my data is used for scientific purposes, and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I agree to participate in this experiment for the reward gift card of the advertised value.
- I understand that if I make illicit activities known, these will be reported to the appropriate authorities.
- I understand that I may refuse to answer any question and that I may withdraw at any time without penalty. I accept that I will not be given the reward gift card if I exit the experiment before it concludes.
- I understand that should I complete the experiment, I may subsequently ask that my input be removed from the study and any records destroyed – within a period of seven days after my participation. I understand that after seven days since my participation have passed, the corresponding data will be fully anonymised and pooled together for further analysis, rendering a withdrawal of data collected from my participation impossible.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I understand that, subject to the constraints above, no audio recordings will be replayed in any public forum or made available to any audience other than the current research team. However, anonymised transcripts of the recordings and other anonymised data might be publicly shared to facilitate the scientific reproducibility of any published findings and to enable potential further research.
- I will not name myself or any third parties in any open text field of the questionnaire or in any recorded part of the user study, including the focus group discussion.
- I understand that if I or anyone in my family has a history of epilepsy, then I am proceeding at my own risk.

- I understand that the experiment is hosted by Trinity College Dublin and that the researcher will hold all data from my participation in the same institution.
- I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
- I have received a copy of this agreement.

**Research Group**

This work has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6223. This work was supported by the ADAPT Centre, funded by Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106_P2).

**Responsibility and Contact Details**

I, the principal investigator, have provided the above information and the information contained on the Participant Information Sheet in good faith, and I commit to abide by it.

Jovan Jeromela ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin
Email: jeromelj@tcd.ie

**Participant's Consent**

By signing this consent form, you give your consent to participating in the research study and the collection and use of the collected data in the manner described above.

Participant's signature: _____

Participant's name: _____

Date signed: _____

## A4.2 User Study Resources

### A4.2.1 Week at a Glance

You don't need to memorise anything — just get a general sense of what the week looks like.

**Weekly Overview – Life of the Calendar Owner**

- **Work is the main anchor** – most weekdays are structured around office hours, with a mix of:
  - **Team collaboration**
  - **Deep focus blocks** for uninterrupted work
  - **Manager updates** and cross-team coordination
  - **End-of-week wrap-up**
- **Two major work projects** dominate the work duties:
  - *Project Cuttlefish* and *Project Prometheus*
  - These meetings should not be moved if possible, other than that there is some flexibility around work hours, with most employees arriving before 9am and leaving the office around 5pm, with 30-60 lunch around noon.
- **Fitness and health are prioritised**:
  - Gym training, home workouts, and yoga sessions
  - Post-lunch walks to stay active, if work permits
- **Meals are often social and scheduled**:
  - Pretty much every workday has a fixed lunch slot
  - Regular family dinners at home; some social meals dinners outside
- **Balanced private life**:
  - Household and chores are also sometimes scheduled over the week and on the weekend
  - Sometimes blocks of time are reserved for video games and relaxation
  - Sleep time is very important and no event is to take place between midnight and 5am, with as few as possible events taking place a few hours around that block as well
- **Weekends are more relaxed but some things are still scheduled**:
  - Saturday – household chores followed by lunch with friends
  - Sunday – family lunch out, then home cooking session for the week

### A4.2.2 User Study Tasks

First round of user study:

1. Add a grocery pickup on Sunday before the meal preparation time.
2. Move the Monday Weekly Team meeting 15 minutes early, so that it starts at 9:15.
3. Ensure there are no clashes over the week.
4. Can you see find any other improvements of the calendar for the upcoming week?
5. What are the different stages of the assistance?
    - Can you see what information the assistant requires for each of them?
    - If you haven't already, increase the stage to the maximum.
    - Does this give you more ideas for potential improvements of the next week's schedule?
6. If you have not done so already, make sure that you are on time for the doctor's appointment on Wednesday.
7. Pick up a Friend visiting from the airport at 12:30 on the Saturday.
    - Can you discover how the assistant sees what matters to the user?
    - Explore the suggestion explanations until you feel you understand how it works.

### A4.2.3 Locations, Priorities, and Commute Times

#### Locations

- Move (commute only)
- Home
- Office – some 45 minutes away
- Gym Liftiron – near home
- Getgo Store – near home
- Pro Clinic – near the office
- Restaurant West End – in the centre, roughly between the home and the office
- Maxx Cinema – in the centre, roughly between the home and the office
- Red Fox Pub – in the centre, roughly between the home and the office

**Commute Timetable is in Table A4.1** (not available to the participant, but in the user model of the assistant; all times in minutes)

Table A4.1: Distance/Time Matrix between Locations (in minutes)

| Move | Home | Off. | Gym | Store | Clin. | Rest. | Cin. | Pub |
|---|---|---|---|---|---|---|---|---|
| **Home** | – | – | – | – | – | – | – | – |
| **Office** | 45 | – | – | – | – | – | – | – |
| **Gym Liftiron** | 20 | 45 | – | – | – | – | – | – |
| **Getgo Store** | 10 | 35 | 40 | – | – | – | – | – |
| **Pro Clinic** | 60 | 20 | 20 | 50 | – | – | – | – |
| **Restaurant West End** | 25 | 25 | 35 | 15 | 40 | – | – | – |
| **Maxx Cinema** | 25 | 25 | 35 | 15 | 40 | 5 | – | – |
| **Red Fox Pub** | 25 | 25 | 35 | 15 | 40 | 5 | 5 | – |

**Priorities** (all explicitly here as well as in the user model):

- Working in the office at least 3 times a week (up to two times at home is okay)
- Try to work out at least 4 times a week, ideally once in a gym and once practicing yoga
- Working for eight hours a day, approximately from 9:00 to 17:30, to accommodate for a lunch break
- Prefer not to work overtime.

## A4.3 User Study Protocol

This is the guidance for leading the user study on the Scrutable Intelligent Personal Assistant (SIPA). First two phases involve individual tasks, and phase three is a group discussion.

Ideally, each slot includes one to three researchers, and each researcher is to supervise up to two participants (resulting in 2 to 6 participants per slot). There are to be around 20 participants in total, around 7 of whom in the first round.

If a researcher is assigned two participants in a slot, they may have both complete the survey (first phase) in parallel. However, the assistant interaction must be conducted individually with each participant, as it is a think-aloud study.

In the third phase, if more than one researcher is present, they should agree on who will serve as the main moderator. Other researchers may remain in the room and ask follow-up questions, but the main moderator is primarily responsible for leading the focus group.

### A4.3.1 First Phase

Familiarise yourself with the survey: [link redacted]

#### Welcoming Participants, Consent Form, and Pseudonym Choice

Welcome all participants: Hello! Welcome to the study on scrutable intelligent personal assistants. Please read this informed consent form if you haven't already. Take the time you need and feel free to ask any questions:

[a link to the User Study Informed Consent Form, as reproduced here in Subsection A4.1]

Wait for the participant reads and signs the consent form as well as to then answer all the questions.

**Pseudonym assignment:**

- Ask the participant to choose a pseudonym.
- **Before confirming** check that the pseudonym hasn't already been taken (link below).
- Explain that the pseudonym is used for anonymisation, and that the participant will be asked to introduce themselves using this pseudonym during the group discussion.

**On the consent form, add the participant's pseudonym at the bottom.** This is to ensure that we can delete the data post-study if the participant asks us to do so later.

Also **add the date and the pseudonym to the list in Google Sheets**. Do NOT put the participant's real name on Google Sheets or anywhere else.

[link to Google Sheet Drive redacted]

#### Survey

Open the survey. Make sure the Pseudonym entered for the first question is the same as the one written on the Consent Form.

[link redacted]

Let the user fill in survey. It's okay to answer any queries they may have.

### A4.3.2 Second Phase

Familiarise yourself with the *Week at a Glance*, *User Study Tasks*, as well as the *Locations, Priorities, and Commute Times beforehand* [all in Section A4.2].

#### Show the User the Structure of the Week

Tell the user that about the next part of the study.

**Introduce the next part of the study:**

*Now we'll move on to the next part of the study.*

*You'll be asked to make some changes to an agenda for the upcoming month.*

*You'll do this using a simple program that includes a text-based virtual assistant.*

**Show the calendar and set expectations:**

*Before we begin, please take a moment to look over the following weekly schedule, which you'll be starting from.*

*You don't need to memorise anything—just get a general sense of what the week looks like.*

Inform them that at that point you will **turn on the recording**. Ask them to agree to this, after which you start the recording.

*From this point, I will now turn on the audio recording. We will keep the recording going while you look at the agenda and while you are solving the tasks. So please, while you look at the agenda and complete the tasks, please think out loud. That means **verbalise what you're thinking**— what you notice, what you're unsure about, what you're planning to do. This helps us understand how you interpret and interact with the system.*

You can also phrase the introduction slightly differently. Ask them questions about what they see. Our goal here is not to note, write down their answers, just to encourage them to think about the information they see. For instance, you may ask:

- *Do you think this is a reasonably organised week?*
- *Would you say this is a typical week for someone living in Dublin?*
- *What kind of person do you imagine might have a calendar like this?*
- *Does this seem like a healthy work–life balance to you?*

Once they have answered these questions, you can ask **if they are happy to start using the calendaring assistant**.

If they ask for more time with the calendar, allow them this time, but reiterate that they do not need to memorise the agenda, but only to familiarise themselves with it. Reiterate that they will be able to see the agenda on their screens during the tasks as well.

When the participant is happy to continue, lead them to a computer with the prototype and start the task.

### Calendaring Assistant Interaction and Onboarding Protocol

For the next phase, open and run your local instance of the calendaring assistant.

Introduce the participant to the interface **and do a short onboarding**:

*You can take some time to familiarise yourself with the interface of the assistant. I'll give you* ***the task sheet*** *whenever they are ready.* [The User Study Tasks can be seen in Subsection A4.2.2]

*As you can see this is a simple programme with three main windows:*

- *The calendar, with short overviews of events for the upcoming week — this is the week you are asked to modify from the perspective of the calendar owner.*
  - *You can click on an event or empty time in a day. Doing this will give you more details about the event in the context window, at the right hand side below.*
  - *It will also give you the calendaring assistant's advice, in the window at the right hand side above.*
- *The context window to the right below is where you will be doing most of the event manipulation, as this programme is quite simple and has not textual or voice interface, but just buttons and fields.*
  - *You will also not be able to do event manipulation through the calendar view to the left, e.g. you cannot drag the events to reschedule them, you have to click the event and then enter the new times in the context window.*
- *The widow to the right and above is the main element of the calendaring assistant, it is where the assistant's suggestions will appear.*
  - *The assistant know some things about the user, depending on the so-called permissions stage.*
  - *The higher the permissions stage, the more information the assistant stores in its user model. You can think of the user model essentially like a memory o the information the assistant has on the user.*
  - *As you go through the tasks, you may or may not find the suggestions helpful and you are free to choose if you want to follow them or not.*
  - *You can also click on the Explanation button for each suggestion, if you want to understand more on how and why it was created, and what information the assistant used for it.*

- *The assistant offers suggestions for each day of the week.*
- *You can scroll the suggestions using the arrows. If the permissions stage permits it, there might also be weekly suggestions available for you. These are always listed at the end of the suggestions for the day.*

Encourage the participant to **think aloud** while exploring the UI, in line with the think-aloud protocol:

*As before, please continue verbalising your thoughts as you interact with the system—what you're noticing, what you're unsure about, what you're trying to do.*

**Do not interfere.**

If asked about how to do something, say that it would be ideal if they could try to resolve the issue on their own.

If the participant insist on needing help, becomes frustrated, or asks to skip a task, you can help them on the specific task and then encourage them to move on their own.

If the participant wants to leave the study, they can do so at any time, even in the beginning or in the middle of performing tasks with the calendaring assistant. Please remind them that they would not receive the participation gif voucher in that case.

**While the participant works:**

- Your role is simply to **observe and wait**.
- Occasionally remind them to keep talking, if needed.

**Once the participant is done interacting with the calendaring assistant:**

- **Copy the log file** from *<local_folder>\prototype\backend\logs\logger_logs\<current-_date>\<latest_log_file>* (e.g. if using Jovan's Trinity laptop: *C:\Users\Jovan-\OneDrive - Trinity College Dublin\UniLife\PhD\Coding-WorkLaptop\SPIPA4TM\prototype-\backend\logs\logger_logs\2025-04-16.log*), or, in the case that multiple logs were created, copy them all. Then go to the dedicated Google Drive folder below, create a new subfolder, name it after the pseudonym of the participant, and copy the logs. [link redacted]
- Thank the participant for their effort in this phase of the study;
- **End the audio recording and make sure it is saved with the participant's pseudonym as the name;**
- Coordinate with other moderator(s) to see if the other participant(s) is done.

In the case that **you have more than one participant for the slot**: Repeat **Parts 1 and 2** (survey and calendaring assistant tasks) with the next participant(s) before proceeding to **Part 3** (the focus group discussion).

### A4.3.3 Third Phase

Once all participants who have come for the slot (at most 6, at least 2) have completed their parts 1 and 2, ask them all to come together in a room for the final part — the user study.

#### Introduction to the Discussion

Begin by again thanking everyone and then introduce phase:

*Now that everyone is here, we're ready to begin the third and final part of the user study.*

*This final part is a group discussion. The goal is to hear your honest thoughts and reflections on what you just experienced.*

*Please don't worry about saying the "right" thing or how your responses might affect the study results. There are no right or wrong answers—we're simply interested in your perspectives. You can speak naturally and openly.*

*The only thing we ask is that you remain civil and respectful toward each other throughout the discussion.*

#### Pseudonym paper:

*Please take a moment to write your pseudonym on a piece of paper and place it in front of you.*

*For the rest of the session, please refer to one another using these pseudonyms only—even if you introduced yourselves earlier using real names.*

Ask if there are any questions about the last part of the study itself and answer them if necessary.

#### Inform the participants that you are about to turn the audio recording again, and do so.

Then begin the moderation.

### Moderation of the Discussion and Questions

This is an exploratory semi-structured study, whose output is to be qualitatively analysed using thematic analysis. That means we want to allow the participants to express their views freely, with questions being there as a guidance, not a prescription.

In your role as a moderator, you should ask questions in order to keep the conversation going, or to include participants who have spoken less or seem afraid to raise their own opinion.

You may deviate from the potential questions below, especially if the participants start discussing potentially interesting insights. Always try to phrase questions in an open-ended manner (e.g. "How did you like. . ." rather that "Did you like. . .").

*Potential Questions* are below.

#### Prototype-Related Questions

*These can be used as opening questions, to help the participant situate their opinions in the context of the tasks they had just solved.*

1. How did you like the calendaring assistant?
2. What types of suggestions did you find helpful?
3. Was there anything about the assistant that felt inappropriate or out of place?
4. To what extent did the presence of the assistant make the tasks easier to solve?
5. What would determine if you would use such a calendaring assistant in your daily life?

#### Scrutability

*Here we go into the crux of the study, try to ask most questions from this category*

1. How important is it to you to be able to check what information about you has been collected?
2. How interested are you in understanding how the assistant models or interprets that information?
3. How much time or effort would you realistically be willing to spend understanding how an assistant like this works?
4. Does it matter to you whether the assistant's advice is personalised specifically to you, or is it enough to understand how it works in general?
5. What kinds of assistant behaviour might prompt you to ask *why* it did something?

#### Proactivity, Delegation, and Staging

*With these questions we want to test whether proactivity and delegation cause new challenges for scrutability and whether staging is as a desired principle of Scrutability.*

1. Imagine that the assistant was capable of proactively changing events in your calendar. What would need to happen for you to be comfortable with that?
2. What would you want to know about the assistant in order to trust it to contact others on your behalf?
3. Would you want to be notified before the assistant takes any proactive action, or can you imagine knowing enough about how it works to simply trust it?
4. Would you want to be notified each time the assistant contacts someone—to schedule or reschedule a meeting? How much time and effort would you be willing to invest in reviewing what it wants to say or do on your behalf?
5. In your interaction with the prototype, you were shown "stages" where you choose what information to give in exchange for certain services. Did you find that useful? Would you like the assistant to offer more such options if they came with tangible benefits?

**Usage Willingness Questions**

*These questions are aimed to incite participants to vocalise what would decide if they would be comfortable using such an assistant.*

1. What would need to change or improve in this assistant for you to be more likely to use it regularly?
2. Would you feel more or less comfortable using this assistant compared to virtual assistants like Siri, Google Assistant, or Alexa?
3. Would you feel more or less comfortable using this assistant compared to chat-based AI tools like ChatGPT or Gemini?
4. Would you be more comfortable using a calendaring assistant like this, or a smart speaker?
5. Would you prefer using this over a traditional calendaring app like Google Calendar?
6. If you could bring together a virtual assistant/chat-based AI with a calendar, what would matter to? (talk to the calendar/explanations, mode of explanation vs mode of interaction for calendar specifically)

The questions might be modified, removed, or added as we go.

**Completing the Discussion**

Once the study is over, thank everyone for the participation yet again, and remind them of the consent form and its details, as well as the gift vouchers.

Before they leave, ask them to fill in the form for the gift card distribution: [link redacted]

*Before you leave, please fill in the form with the information the SCSS Accounting Department needs for the one4all 45 Euro digital gift card to be sent to you.*

*Thank you all once again for taking part in the study—we really appreciate your time and insights.*

*Before you go, please make sure to take your copy of the consent form with you. If you'd like, you can write your pseudonym on it. This will allow you to reach out later if you have any questions or would like your data to be withdrawn from the study.*

*If you decide to do that, you can contact the principal investigator, Jovan Jeromela, at jeromelj@tcd.ie.*

*Please note that data can be withdrawn* ***within the next seven days****.*

*Finally, your gift voucher will be sent to you within a week by the SCSS TCD accounting office to the email address you provided in the final survey.*

*Thanks again, and have a great day!*

### A4.3.4 Additional Information

This includes additional information you should familiarise yourself before the study starts. In particular, you will see what the weekly agenda, that the user is to interact with, looks like. There is also a list of (non-finalised) user tasks, and a page with current TO-DOs on the technical side (which should be empty by the time the study begins).

[The linked Additional Information content is reproduced in Subsection A4.2]

## A4.4 Thinking-Aloud Themes and Codes

Table A4.2 provides an overview of all thinking-aloud statements of interest and the themes and subthemes they were attributed to.

Table A4.2: Interaction Findings Codes: themes, subthemes, statements of interest, statement count (SC), and participant count (PC)

| Theme | Subtheme | Code | SC | PC |
|---|---|---|---|---|
| Active Scrutiny | Scrutability | Approach: Figuring out by exploring before starting | 3 | 3 |
| | | Features: Graph is intuitive | 1 | 1 |
| | | Features: Tried to interpret the graph | 4 | 4 |
| | | Finding: Assistant has a user model that maps both event time and travel time | 19 | 8 |
| | | Finding: Considers going into explanations to understand the possibilities/features of the IPA | 2 | 2 |
| | | Finding: Interest in the assistant's learning mechanism | 23 | 11 |
| | | Task design: Used the user model to find out the user's preferences (task design limitation) | 3 | 1 |
| | User Control (Desired) | Approach: Changes permission stages to fix 'bugs' | 2 | 2 |
| | | Approach: Have the highest level of assistant setting to 'help out as much as it can' | 2 | 2 |
| | | Considerations: Expected increase in the stage would make the assistant actually do the task | 2 | 1 |
| | User Control (Implemented) | Approach: Starts with a lowest level of permissions and builds up | 1 | 1 |
| | | Features: Appreciates multiple permission stages | 2 | 2 |
| | | Features: Wants to change the user model | 2 | 2 |
| | | Finding: Wants to customise the model information | 2 | 2 |
| Interface Friction | Control | Features: Right-click on time to create event | 2 | 2 |
| | | Features: Surprise by timing being broken down in minutes | 1 | 1 |

Continued on next page

**Table A4.2 – continued from previous page**

| Theme | Subtheme | Code | SC | PC |
|---|---|---|---|---|
| | | Features: UI change to allow dragging of events on calendar | 1 | 1 |
| | | Features: Wanted links or resources in project descriptions of the events | 1 | 1 |
| | Interruption | Features: Frustrated with pop-ups about event changes | 6 | 2 |
| | | Finding: Confused whether the duration change suggestion has also changed the duration as well | 3 | 1 |
| | | Finding: Did not see that description is in fact not mandatory because of the suggestion | 1 | 1 |
| | | Finding: Gives up with the change after a pop-up suggestion | 2 | 1 |
| | Visualisation | Features: Better visualization between event categories | 1 | 1 |
| | | Features: Wanted asks of different flexibilities to be coded or coloured differently | 2 | 1 |
| Scheduling Considerations | | Approach: Considering mutable vs. immutable events | 4 | 4 |
| | | Approach: Wants to have a balanced work week | 2 | 2 |
| | | Considerations: Assistant requires perfect information | 11 | 5 |
| | | Considerations: Assuming non-calendared events | 6 | 3 |
| | | Considerations: One-time vs. recurring events | 2 | 2 |
| | | Considerations: Overlapping events may be intentional multitasking | 5 | 5 |
| | | Considerations: Social considerations made explicit | 21 | 9 |
| | | Finding: Calendars in real life are more than just time and place: it is event information (tickets, notes, etc.) | 4 | 3 |

Continued on next page

**Table A4.2 – continued from previous page**

| Theme | Subtheme | Code | SC | PC |
|---|---|---|---|---|
| | | Finding: Explicitly acknowledges the cognitive load of using calendar | 2 | 2 |
| Scrutability Barriers | Implicit Scrutiny (Explanation Extraction) | Approach: Answering questions about the assistant and its data use by guessing, rather than looking at the model/explanations | 3 | 1 |
| | | Approach: Anticipating reactive suggestions to reappear | 5 | 4 |
| | | Approach: Assumptions on AI knowledge based on its actions | 16 | 7 |
| | | Approach: Doing mental calculations | 20 | 13 |
| | | Approach: Explores the connection between stage and clash detection | 3 | 3 |
| | | Approach: Figuring out by doing | 23 | 15 |
| | | Approach: Forgets to/does not turn to assistant (needs to be reminded) | 1 | 1 |
| | | Approach: Modeling assistant's model (2nd order ToM) | 2 | 2 |
| | | Features: Commute time implied | 11 | 8 |
| | Interpretability | Features: Confusion about a feature or the assistant | 33 | 8 |
| | | Finding: Assistant features and functions hard to parse | 22 | 7 |
| | | Finding: Confused by there being deeper explanations | 2 | 2 |
| | | Finding: Explanations self-explanatory | 4 | 1 |
| | | Finding: Feeling of incompetency to understand explanations, graphs, etc. | 1 | 1 |
| | Interpretability (Stages) | Approach: Forgets to change stages | 1 | 1 |
| | | Finding: Expects more obvious differences between stages | 2 | 2 |
| | | Finding: Not understanding what information is needed for each stage | 4 | 3 |

Continued on next page

**Table A4.2 – continued from previous page**

| Theme | Subtheme | Code | SC | PC |
|---|---|---|---|---|
| | | Task design: Confusion about changing permissions | 3 | 3 |
| | Purpose-Driven Scrutiny | Considerations: Doesn't care about the explanation | 3 | 3 |
| | | Finding: Acknowledges the graph, but does not spend more than 3s looking at it, decides to act thereafter | 4 | 4 |
| | | Finding: Does not even look at suggestion text until 25 mins, instead just focuses on colours and thinks that is the assistant | 2 | 1 |
| | | Finding: Likes more concrete suggestions | 2 | 2 |
| | | Finding: Not interested in more details unless task-relevant | 5 | 3 |
| | | Finding: Not understanding (being surprised by) a suggestion leads to looking for the explanation | 7 | 6 |
| | | Finding: Suggestions are not interesting on their own | 9 | 5 |
| | | Finding: Turns to assistant for help but finds it not useful for the expected goal/task | 20 | 6 |
| Shortcomings of Assistance (Insufficient Autonomy) | | Considerations: Assistant(s) can be used for multiple calendars and people auditing | 1 | 1 |
| | | Considerations: Assumption of the assistant learning | 1 | 1 |
| | | Features: Expecting more from the agent | 19 | 10 |
| | | Features: Expects automatic event duration change | 3 | 2 |
| | | Features: Wonders if there is an accept button | 3 | 2 |
| | | Features: Would like it if the assistant prioritised events | 1 | 1 |

Continued on next page

**Table A4.2 – continued from previous page**

| Theme | Subtheme | Code | SC | PC |
|---|---|---|---|---|
| | | Features: Would like the assistant to track the number of specific event types scheduled over the week | 1 | 1 |
| | | Finding: Assistant should be able to balance the workload across the week | 2 | 2 |
| | | Finding: Wants assistant to makea change by itself | 4 | 2 |
| Study Artefacts | | Considerations: Says that they changed their approach to fit the perceived experiment expectations | 1 | 1 |
| | | Task design: Confusion about task design | 10 | 5 |
| | | Task design: Struggling with the constraints of the task | 10 | 6 |
| Trust & Reliability | | Finding: Anthropomorphised AI | 1 | 1 |
| | | Finding: Does not prefer calendars with shared or external input | 1 | 1 |
| | | Finding: Does not trust AI for being good at maintaining, updating, or reminding about tasks | 1 | 1 |
| | | Finding: Does not trust AI to communicate change in calendar effectively | 2 | 1 |
| | | Finding: Prefers lowest permission | 1 | 1 |
| | | Finding: Trusts assistant understands what is important to the user | 1 | 1 |
| Utility of Assistance | Attention Steering | Approach: Exploring the assistant to see if they had missed something | 12 | 7 |
| | | Features: LIked clash reporting | 2 | 2 |
| | | Features: Suggestions helped to find the clashes | 24 | 7 |
| | Contextual Information | Features: Likes duration-changed suggestions | 2 | 2 |
| | | Features: Likes the commute time suggestions | 1 | 1 |

Continued on next page

**Table A4.2 – continued from previous page**

| Theme | Subtheme | Code | SC | PC |
|---|---|---|---|---|
| | | Finding: Indication that the suggestion has helped | 12 | 8 |
| | Ideation | Approach: Looking at the assistant for the open-ended task, but only after looking at the calendar themselves first | 3 | 3 |
| | | Approach: Turning to an assistant when there is a lack of information | 2 | 2 |
| | | Features: Suggestions for ideation | 7 | 6 |
| | | Finding: Excited about assistant's recommendations based on the user model | 2 | 1 |

# A5 Publications

From the period of the beginning of research work on this dissertation to its submission date, ten submissions were published, listed below. Additionally, the last paper in the list, with the evaluation of our main user study, has just been accepted for publication at the *34th Conference on User Modeling Adaptation and Personalization (UMAP '26)*. In the list, we underline titles of research papers whose content directly relates to this dissertation and **bold** the author's name to emphasise the co-author rank.

1. **Jovan Jeromela**. "Scrutability of intelligent personal assistants". *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22)*. 2022.

   "Intelligent personal assistants (IPAs) have become widely available, yet they remain primarily used for discrete, straightforward tasks. By contrast, both user studies and literature reviews indicate that IPAs of the future are to be personalised, proactive, and capable of performing elaborate undertakings. Such systems would have to be based on complex and dynamic user and context models. We believe that scrutability – i.e. the ability of the user to actively study and modify the models towards tuning personalisation – could emerge as an essential element of such a human-assistant interaction paradigm. Yet, to the best of our knowledge, no work so far has investigated how the principles of scrutability, as presented in Kay and Kummerfeld [2013], relate to the context and novel challenges raised by the proactive IPAs and how scrutability could facilitate effort-efficient control of the assistants. This paper introduces our vision of the confluence of the research fields of IPAs and scrutability, presents a diagram of the proposed interaction structure, and reanalyses data from user studies originally presented in [Cowan et al., 2017, Völkel et al., 2021] to better understand user expectations regarding scrutability and proactivity of IPAs." [Jeromela, 2022]

2. Jovan Jeromela. "Context-Based Tweet Engagement Prediction" *Diploma Thesis, TU Wien*. 2023.

   *In this diploma thesis, we used the RecSys 2020 Challenge dataset and evaluation pro-*

*cedure to investigate how well context alone may be used to predict tweet engagement likelihood. In doing so, we employed the Spark engine on TU Wien's Little Big Data Cluster to create scalable data preprocessing, feature engineering, feature selection, and machine learning pipelines. We manually created just under 200 additional features to describe tweet context.*

"The results indicate that features describing users' prior engagement history and the popularity of hashtags and links in the tweet were the most informative. We also found that factors such as the prediction algorithm, training dataset size, training dataset sampling method, and feature selection significantly affect the results. After comparing the best results of our context-only prediction models with content-only models and with models developed by the Challenge winners, we identified that the context-based models underperformed in terms of the RCE score. This work thus concludes by situating this discrepancy and proposing potential improvements to our implementation, which is shared in a public git repository." [Jeromela, 2023]

3. **Jovan Jeromela**, Dipto Barman, Hassan Zaal, Alok Debnath, Awais Akbar, Judy Kay, and Owen Conlan. "1st Workshop on Context Representation in User Modelling". *Adjunct Proceedings of the 31st ACM Conference on User Modeling Adaptation and Personalization (UMAP '23 Adjunct)*. 2023.

   "Context, a critical aspect of personalisation and human-computer interaction, is becoming increasingly significant with the proliferation of user-centric applications. The First ACM Workshop on Context Representation in User Modelling (CRUM 2023), organised in conjunction with the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2023), provided a venue for research in the representation and utilisation of contextual information in intelligent systems. The workshop aimed to bring together researchers from various disciplines, including natural language processing, user modelling, explainable systems, and human-robot interaction and to facilitate conversation about the role of context in adaptive applications. The prospective authors were invited to submit papers of up to seven pages, four of which were accepted for publication and are introduced in this summary." [Jeromela et al., 2023]

4. **Jovan Jeromela** and Owen Conlan. "Voicing Suggestions and Enabling Reflection: Results of an Expert Discussion on Proactive Assistants for Time Management". *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. 2023.

   "While voice-controllable Intelligent Personal Assistants (IPAs) have become widespread in recent years, they remain primarily reactive with rather constrained calendaring capa-

bilities. Anticipating more adaptive and complex assistants in the future, we organised a multidisciplinary expert discussion investigating potential use cases, interaction principles, and user modelling challenges within proactive IPAs for time management. This paper presents the identified themes and deliberations on enticing self-reflection, longitudinal task assistance, interaction modality, dialogue design, perception of the system, usage willingness, onboarding, and explainability. These findings outline a framework of advanced IPAs for time management." [Jeromela and Conlan, 2023a]

5. **Jovan Jeromela** and Owen Conlan. "Onboarding stages and scrutable interaction: How experts envisioned explainability in proactive time management assistants". *Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23)*. 2023.

   "Intelligent Personal Assistants (IPAs) have become increasingly ubiquitous, yet they remain primarily reactive, non-personalised, and inscrutable. Moreover, concerns regarding user control, data stewardship, and communication design persist in the literature. Aiming to shape an appropriate human-assistant interaction framework, we organised a multidisciplinary expert discussion focusing on proactive IPAs for time management – a bounded yet complex domain that may catalyse identifying and tackling paradigmatic challenges. We invited the experts to propose, debate, and chart interaction scenarios, desired characteristics and constraints, and user modelling for proactive IPAs. This paper presents the thematic analysis of the discussion and the resulting interaction diagram. The ability of the user to scrutinise the assistant's models that underpin personalisation and to receive adequate explanations were identified to be of paramount significance. Moreover, a proposed onboarding and control mechanism that may help align the user's perception of the system and the system's actual capabilities is discussed." [Jeromela and Conlan, 2023b]

6. Dipto Barman, **Jovan Jeromela**, Alok Debnath, Marloes Vredenborg, Anouk Van Kasteren, Judy Kay, and Owen Conlan. "Second Workshop on Context Representation in User Modelling". *Adjunct Proceedings of the 32nd ACM Conference on User Modeling Adaptation and Personalization (UMAP '24 Adjunct)*. 2024.

   "The second edition of the Context Representation in User Modelling (CRUM) workshop, themed 'Human-Centred Context', aimed to provide a venue for research on contextual information within the evolving landscape of UMAP. This workshop explored context through a human-centred lens, emphasising the nuanced interplay between subjective and objective contexts in user experience rather than traditional domain-centred approaches. This theme encouraged accounting for the interaction between the user model and the contextual environment. The workshop also fostered discussion on the landscape of contextual representation – ranging from physical settings and social dy-

namics to technological interfaces – as well as on interaction design, user experiences, and influencing user behaviour. Prospective authors were invited to submit two types of papers: full papers of up to seven pages and two-page provocation papers. Accepted submissions are summarised in this workshop report." [Barman et al., 2024a]

7. **Jovan Jeromela** and Owen Conlan. "Devising scrutable user models for time management assistants". *Adjunct Proceedings of the 32nd ACM Conference on User Modeling Adaptation and Personalization (UMAP '24 Adjunct)*. 2024.

   "Intelligent Personal Assistants (IPAs) have become ubiquitous through integration into smartphones, smart speakers, and standalone devices. However, prior studies raised noteworthy usability concerns and determined that IPAs remain primarily used for simple tasks. Such findings contrast with the reported user aspirations for more proactive and truly personalised IPAs. By focusing on the use case of time management, in this paper, we contemplate how scrutability – i.e. the ability of the user to study their assistant and its underlying user model – fits within the vision of more complex IPAs. Furthermore, we describe our ongoing project investigating user interest in and expectations of the scrutability of a proactive calendaring assistant. Lastly, by deliberating on the challenges and benefits of making IPAs scrutable, this paper outlines potential avenues for further research." [Jeromela and Conlan, 2024]

8. Jasmina Gajcin, **Jovan Jeromela**, and Ivana Dusparic. "Semifactual explanations for reinforcement learning". *Proceedings of the 12th International Conference on Human-Agent Interaction (HAI '24)*. 2024.

   "Reinforcement Learning (RL) is a learning paradigm in which the agent learns from its environment through trial and error. Deep reinforcement learning (DRL) algorithms represent the agent's policies using neural networks, making their decisions difficult to interpret. Explaining the behaviour of DRL agents is necessary to advance user trust, increase engagement, and facilitate integration with real-life tasks. Semifactual explanations aim to explain an outcome by providing 'even if' scenarios, such as 'even if the car were moving twice as slowly, it would still have to swerve to avoid crashing'. Semifactuals help users understand the effects of different factors on the outcome and support the optimisation of resources. While extensively studied in psychology and even utilised in supervised learning, semifactuals have not been used to explain the decisions of RL systems. In this work, we develop a first approach to generating semifactual explanations for RL agents. We start by defining five properties of desirable semifactual explanations in RL and then introducing SGRL-Rewind and SGRL-Advance, the first algorithms for generating semifactual explanations in RL. We evaluate the algorithms in two standard RL environments and find that they generate semifactuals that are easier to

reach, represent the agent's policy better, and are more diverse compared to baselines. Lastly, we conduct and analyse a user study to assess the participant's perception of semifactual explanations of the agent's actions." [Gajcin et al., 2024]

9. Jasmina Gajcin, **Jovan Jeromela**, and Ivana Dusparic. "Towards Personalised and User-Friendly Counterfactual Sequences for Failure Correction". *Adjunct Proceedings of the 33rd ACM Conference on User Modeling Adaptation and Personalization (UMAP '25 Adjunct)*. 2025.

   "Understanding how failures occur and how they can be corrected is essential for debugging, maintaining user trust, and developing personalised policies. Counterfactual sequences, which provide an alternative sequence of actions that delivers an improved outcome, have been used to correct failure in sequential decision-making tasks. However, prior work on counterfactual sequences has focused primarily on the algorithmic side of sequence generation, mostly overlooking the potential of counterfactuals as a user-friendly explanation method. In this work, we lay the groundwork for human-centred and personalised counterfactual sequence generation. Informed by insights from psychology and cognitive science, we propose a set of desiderata for understandable and useful counterfactual sequences. We then introduce an algorithm based on these desiderata that generates diverse counterfactual sequences, enabling the user to correct the failure in line with their preferences." [Gajcin et al., 2025]

10. Jasmina Gajcin, **Jovan Jeromela**, Joel Wester, Sarah Schömbs, Styliani Kleanthous, Karthikeyan Natesan Ramamurthy, Hanna Hauptmann, and Rifat Mehreen Amin. "CURE-@IUI: Workshop on Communicating Uncertainty to foster Realistic Expectations via Human-Centered Design". *Companion Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26 Companion)*. 2026.

    "CURE2026 workshop provides a venue for research on how interactive and intelligent interfaces can be leveraged to ensure effective uncertainty communication. With the increased capabilities of black-box AI models, particularly large language models (LLMs), there is a growing urgency to understand exactly what these models can and cannot do. Effective uncertainty communication can help users calibrate their trust and reliance on the AI system. However, effectively conveying uncertainty, be it through conversation or by visualizing their confidence levels, remains a major challenge, as users often get confused or misinterpret uncertainty information due to poorly designed interfaces. CURE2026 workshop brings together researchers from HCI, AI, design, and psychology to investigate how we can design, develop, and evaluate intelligent and interactive interfaces for uncertainty communication. The workshop includes a mini-conference section aimed at inspiring discussion and investigating the role of intelligent and interactive inter-

faces in uncertainty communication. The workshop also features keynotes by Professors Jessica Hullman and Katrien Verbert. Finally, a part of the workshop is dedicated to an interactive activity to discuss and develop guidelines for communicating uncertainty across different domains and tasks." [Gajcin et al., 2026]

11. **Jovan Jeromela**, Matt Murtagh White, Jasmina Gajcin, Alok Debnath, Dipto Barman, and Owen Conlan. "'As Long as It Does What I Want, I'd Be Happy to Trust It': Exploring User Perspectives on a Scrutable Assistant for Time Management" *Unpublished work currently in review for UMAP '26*. 2026.

    "The integration of large language models (LLMs) into digital libraries introduces new possibilities for exploratory search while reshaping how users experience confidence, workload, and synthesis during discovery. This study examines how interface modality, including keyword, semantic, conversational, and conversational-graph search, affects user performance and perception in literary exploration tasks on Project Gutenberg. In addition to the original keyword-based search, we developed three LLM-based interfaces that share a retrieval-augmented generation (RAG) backend, producing grounded and auditable book recommendations. In a within-subjects laboratory study combining quantitative measures with qualitative interviews, conversational interaction improved synthesis quality and reduced perceived workload by supporting iterative refinement and conversational momentum. However, conversational interaction also elevated self-reported confidence without corresponding accuracy gains, indicating potential risks of premature closure. Adding a relational graph visualization enhanced cross-item comparison but increased cognitive effort, highlighting trade-offs between exploratory power and usability. We conclude with design implications for building LLM-powered exploratory systems that balance fluency, reflection, and cognitive demand." [Jeromela et al., 2026]

In addition, one collaborative work submission was rejected at the 31st International Conference on Intelligent User Interfaces (IUI '26) and is currently being prepared for a resubmission.

1. Matt Murtagh White, Yunkai Xu, Nicole León, **Jovan Jeromela**, Felix Vanden Borre, Hiba Al-Nabhani. "From Queries to Conversations: How Interface Modality Shapes Exploratory Search Behavior". *Unpublished work*. 2026.

    "The integration of large language models (LLMs) into digital libraries introduces new possibilities for exploratory search while reshaping how users experience confidence, workload, and synthesis during discovery. This study examines how interface modality, including keyword, semantic, conversational, and conversational-graph search, affects user performance and perception in literary exploration tasks on Project Gutenberg. In addition to the original keyword-based search, we developed three LLM-based inter-

faces that share a retrieval-augmented generation (RAG) backend, producing grounded and auditable book recommendations. In a within-subjects laboratory study combining quantitative measures with qualitative interviews, conversational interaction improved synthesis quality and reduced perceived workload by supporting iterative refinement and conversational momentum. However, conversational interaction also elevated self-reported confidence without corresponding accuracy gains, indicating potential risks of premature closure. Adding a relational graph visualization enhanced cross-item comparison but increased cognitive effort, highlighting trade-offs between exploratory power and usability. We conclude with design implications for building LLM-powered exploratory systems that balance fluency, reflection, and cognitive demand."

# A6 Other Academic Activities

Together with fellow PhD candidates under the supervision of Professor Owen Conlan: Dipto Barman, Hassan Zaal, Alok Debnath, and Awais Abkar, I co-organised the first edition of the *Context Representation in User Modelling* workshop (CRUM '23), held with the *31st ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '23). The workshop was co-chaired by Professor Owen Conlan and Professor Judy Kay, the originator of the concept of Scrutability [Kay and Kummerfeld, 2013] and a professor at the University of Sydney. As the workshop's leading co-organiser, my roles included communication with the UMAP 2023 workshop chairs, invitation of most of the programme committee members, and acting as the corresponding co-organiser author. The workshop accepted four papers for conference-style presentations and it included a panel discussion.

The second edition of the workshop was organised the following year (CRUM '24), with the organising committee including two new co-organisers from Utrecht University, Anouk van Kasteren and Marloes Vredenborg, in place of Hassan Zaal and Awais Abkar. This edition accepted two papers for publication and was realised alongside the workshop in Human Aspects in Adaptive and Personalized Interactive Environments (HAAPIE '24) during the *32nd ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '24). This edition featured a keynote by Christine Bauer, a professor at the Paris Lodron University Salzburg, as well as a closing interactive activity on human aspects in context modelling. The abstracts of the two accepted papers can be seen at the workshop website[1]

Alongside Jasmina Gajcin, a Trinity College Dublin (TCD) doctoral graduate and a science researcher at IBM Research Ireland, I co-founded CURE, a workshop on how intelligent user interfaces can effectively communicate uncertainty in AI systems to help users form realistic expectations and calibrate their trust and reliance on these systems. The half-day workshop, held at the *31st ACM Conference on Intelligent User Interfaces* (IUI '26) accepted four full papers for publication and presentation. As can be seen on the workshop website[2], the final organising team comprised of eight researchers from seven institutions: Jasmina Gajcin

[1] CRUM '23: `https://crum-workshop.github.io/2023/`; CRUM '24: `crum-workshop.github.io/`
[2] Available at `https://cureworkshop.github.io/cure-2026/`

(IBM Research Ireland), Jovan Jeromela (TCD), Joel Wester (University of Copenhagen), Sarah Schömbs (University of Melbourne), Styliani Kleanthous (Open University of Cyprus), Karthikeyan Natesan Ramamurthy (IBM Research USA), Hanna Hauptmann (Utrecht University), and Rifat Mehreen Amin (LMU Munich). The programme included two invited keynote presentations by Professor Jessica Hullman (Northwestern University) and Professor Katrien Verbert (KU Leuven), as well as an interactive session moderated.

In addition, I helped facilitate three citizen think-ins organised by the ADAPT Centre's Education and Public Engagement group. The purpose of the events was to inform the general public about the current and emerging applications of artificial intelligence in public life and to empower participants to form and voice their views. In particular, the first citizen think-in was co-located across multiple higher education institutions in Ireland and focused on video surveillance, AI, and privacy. The second citizen think-in was part of the European Researchers' Night 2022, and was co-organised by TCD and the University of Malta. The third citizens' think-in took place in Dundalk Public Library in October 2025, as part of a broader Age-Friendly AI project, inviting senior citizens to discuss their concerns and opinions regarding emerging AI-based technologies and the ethical and practical challenges they raise.

Furthermore, I was a teaching assistant for courses in Computer Engineering, Algorithms and Data Structures I, Artificial Intelligence, Information Management II, and Authoring for Digital Media. I also organised and held a guest lecture on the Scrutability of Intelligent Personal Assistants for students partaking in the Adaptive Systems course. Finally, I assisted Professor Owen Conlan in supervising a master's student's work towards his MAI thesis on the scrutability of web-surfing behaviour. I was the elected postgraduate representative at TCD governing bodies, serving at the Student Life Committee in ('23–'24 term) and the Board of Trinity College ('24–'25)

Lastly, I was selected for a four-month full-time paid research internship with the Trustworthy AI Group at IBM Research Ireland. The internship took place in Dublin during the summer of 2024, during which I paused my dissertation research. I was supervised by Elizabeth Daly and mentored by Seshu Tirupathi and Radu Marinescu. My research focused on applying large language models to decision-support systems in AI risk assessment and compliance. My internship was closely aligned with IBM's AI Risk Atlas[3].

[3]More details available at `https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas` (last access for all links in this appendix: 2026-03-10)