

# Designing Adaptive Explainable Interfaces for Agentic Health Assistants



**Master Thesis**  
Strategic Product Design,  
Delft University of Technology

**Ole Kaars Sijpesteijn**

Author

**Ole H.E. Kaars Sijpesteijn**

Student Number: 5089689

Submitted in partial fulfilment of the requirements for the degree of:

Master of Science

in Strategic Product Design

Faculty of Industrial Design Engineering

At Delft University of Technology, to be defended publicly on  
20th of February, 2026

Thesis Committee:

Chair: **Dr. Ir. R.G.H. (Bart) Bluemink**

Mentor: **Dr. P. (Pan) Wang**

Company Accenture Song:

First mentor: Ir. Féline Mollerus

Second mentor: Loys Starink

# Abstract

The global healthcare sector confronts a critical disparity between rising demand and capacity, driving a transition toward data-driven, preventive care models enabled by Agentic Artificial Intelligence (AI). However, the integration of these autonomous systems encounters a significant “trust paradox,” where patient hesitation persists despite the technical potential of AI to enhance clinical outcomes. While Explainable AI (XAI) offers a potential bridge, existing approaches often prioritise technical interpretability over the cognitive and emotional requirements of non-expert patients. This thesis investigates how user-centred interfaces can be designed to foster patients’ trust and understanding in agentic AI health assistants.

Initial exploratory research identified that trust in healthcare is fundamentally relational, constructed through empathy, clarity, and “responsible transparency” rather than exhaustive data exposure. These insights guided the development of three distinct interface concepts—ranging from text-dominant summaries to complex, interactive data visualisations—which were evaluated in a within-subject user study (N=11).

The evaluation demonstrated that while complex visualisations may signal competence, trust functions as the primary gatekeeper for adoption; perceived trust and immediate reassurance (outcome-first guidance) were the strongest predictors of behavioural intention, whereas perceived ease of use played a secondary role and information density did not reliably increase intention. Although participants sometimes described high-complexity interfaces as theoretically “most trustworthy,” they preferred designs that build trust without cognitive overload: actionable steps by default, a brief plain-

language “why” for interpretation, and deep evidence reserved for on-demand inspection.

Synthesising these findings, the thesis presents the Explanatory Depth Framework, a staged interaction model that organises AI communication into four layers: (1) Outcome & Next Steps, (2) Brief Explanation, (3) Traceability, and (4) Deep Evidence. A core component of this framework is the “Question Moderator,” a mechanism designed to calibrate the default depth of explanation based on the user’s emotional state (e.g., anxiety) and informational intent. The project concludes with a strategic roadmap for Accenture Song, positioning explainability not as a static visual feature, but as a governed interaction capability essential for the responsible deployment of future agentic healthcare systems.

# Preface

This thesis, *Designing Adaptive Explainable Interfaces for Agentic Health Assistants*, marks the completion of my Master's programme in Strategic Product Design at Delft University of Technology. From September 2025 to February 2026, I worked on this graduation project in collaboration with Accenture Song, focusing on UI/UX design, agentic AI, healthcare, and explainable AI.

This project has taught me a lot. When I started at the Faculty of Industrial Design Engineering, I did not expect to end up designing interfaces for agentic experiences—nor to explore how AI communication can influence whether people feel supported, overwhelmed, or confident about the guidance they receive.

One belief became stronger throughout this work: improving AI adoption is not only about building more AI solutions, but about making AI feel more human and empathic. Today, the focus is often on creating new AI capabilities as quickly as possible. I believe the next step is to improve the quality of the interaction—making it more personal, more caring, and more aware of what someone needs in that moment. This thesis contributes to that goal by showing how designers can shape and adjust explanations to different user needs, so AI guidance becomes easier to understand and easier to trust.

I am grateful to everyone at Accenture Song who helped me throughout this project. I want to start by thanking **Féline** for bringing me into this interesting topic. She challenged me with sharp questions that went beyond the scope of the thesis itself. I learned a lot from her way of thinking about healthcare, design and AI. Her knowledge—and especially her energy and passion for AI tools—was a real source of inspiration during the process.

I want to thank **Loys** for being a true buddy during my time at Accenture Song. She helped me find my way as an intern and made this period not only educational, but genuinely enjoyable. I wish her all the best with her new chapter in Sydney.

I am thankful to my TU Delft supervisors for their guidance and feedback. **Bart** supported me throughout the project with calm direction and clear design guidance. I have a lot of respect for how he managed to be chair for so many students while finishing his PhD at the same time. He truly felt like a mentor, and he helped me think ahead to what comes after graduation.

I want to thank **Pan** for her expertise in human-computer interaction and AI. Her feedback was extremely valuable, and I am especially grateful that she consistently made time for our weekly calls—often staying up late in China—so I could keep making progress.

Lastly, I want to thank my friends and family for their encouragement, and support throughout this period.

I hope you find this thesis as interesting to read as it was for me to write!

Amsterdam, February 2026  
Ole Kaars Sijpesteijn

# How to Read This Thesis

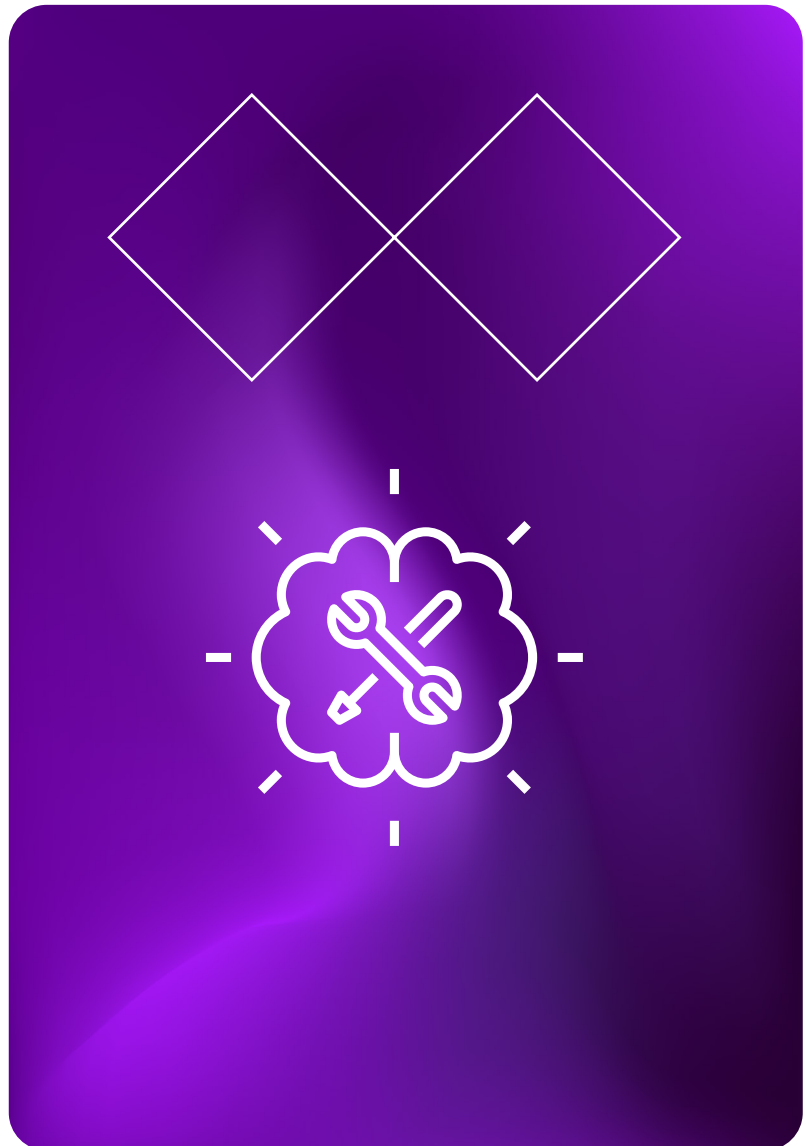
This guide supports readers in navigating the thesis and understanding how each chapter contributes to the overall design process and research logic.

Each chapter begins with a short introduction that states the chapter's objective and scope and briefly summarizes the methods or activities conducted.

The Double Diamond figure indicates the project phase addressed in the current chapter and can be seen on the right.

Where relevant, chapters include an AI-tool icon that lists the AI tools used and their role in the work during that chapter

*Important Quotes and Statements are written in italic and purple just like this text.*



# Glossary of Abbreviations

<b>Abbreviations</b>	<b>In Full</b>
<b>AI</b>	Artificial Intelligence
<b>API</b>	Application Programming Interface
<b>BHLS</b>	Brief Health Literacy Screen
<b>CTA</b>	Call to Action
<b>D&amp;DP</b>	Design & Digital Products (Accenture Song team)
<b>DALL·E</b>	OpenAI text-to-image model (used for visual ideation)
<b>DID / DIDs</b>	Decentralized Identifier(s)
<b>DL</b>	Deep Learning
<b>EHDS</b>	European Health Data Space
<b>EHR</b>	Electronic Health Record
<b>GDPR</b>	General Data Protection Regulation
<b>GP</b>	General Practitioner
<b>HCI</b>	Human-Computer Interaction
<b>HIPAA</b>	Health Insurance Portability and Accountability Act
<b>HMW</b>	“How Might We” (design question format)
<b>JSON</b>	JavaScript Object Notation
<b>LLM</b>	Large Language Model
<b>MAS</b>	Multi-Agent Systems
<b>ML</b>	Machine Learning
<b>PGHD</b>	Patient Generated Health Data
<b>PHA</b>	Personal Health Agent
<b>PHR</b>	Personal Health Record
<b>S-TIAS</b>	Short Trust in Automation Scale
<b>SSI</b>	Self-Sovereign Identity
<b>TAM</b>	Technology Acceptance Model
<b>UI</b>	User Interface
<b>UX</b>	User Experience
<b>XAI</b>	Explainable Artificial Intelligence

# Table of Content

<b>Abstract</b>	<b>3</b>
<b>Preface</b>	<b>4</b>
<b>How to Read This Thesis</b>	<b>5</b>
<b>Glossary of Abbreviations</b>	<b>6</b>
<b>Project Introduction</b>	<b>10</b>
1.1. Context	10
1.2. Accenture Song	9
1.3. Research Intro + RQ's and SQ's	10
1.4 Project Approach	11
1.4.1 AI Declaration	11
1.4.2 Mixed Method Approach	12
1.4.3 Double Diamond and Stingray	12
<b>Literature Review</b>	<b>14</b>
2.1. The Evolution of AI in Healthcare	15
2.2. Agentic AI Systems in Healthcare Context	16
2.3. Healthcare Data Landscape and Patient Ownership	18
2.4 The Trust Paradox in Healthcare AI	20
2.5 Explainable AI as a Human-Facing Design Challenge	23
2.6 Key Takeaways	26
<b>Exploratory Research</b>	<b>27</b>
3.1 Research Approach	28
3.2 Sampling of Participants	28
3.3 Data Analysis Process	28
3.4 Design Criteria Derived from the Exploratory Findings	29

<b>Framing the Design Direction</b>	<b>31</b>
4.1 Insights	32
4.2 Focus Area	33
4.3 Design Brief	34
4.4.Integration and Refinement	36
<b>Concept Development and Evaluation</b>	<b>37</b>
5.1 Exploring and Ideating a Healthcare assistant	38
5.2 From Idea to Final Concepts	42
5.3 Synthetic User Testing	43
5.3.1 Rationale and Approach	43
5.3.2 Evaluation Setup	43
5.4 Final Concepts	45
5.5 Rationale for Concept Diversity	48
5.6 Testing Approach and Study Design	50
5.7 Quantitative Results and Comparative Analysis	53
5.8 Qualitative Results	57
5.9 Key Insights	58
<b>Framework, Roadmap and Prototype</b>	<b>60</b>
6.1 From Insights to Deliverables (Development Process)	61
6.2 The Explanatory Depth Framework: Components and Structure	65
<b>Explanatory Depth Framework</b>	<b>67</b>
6.3 Applying the Framework	68
6.4 Final Prototype (Demonstration of the Framework)	70
6.5 Roadmap for Implementation and Future Development	75
6.6 Validating the Framework (Expert Review)	79
6.7 Reflection on the Framework and Deliverables	81

<b>7. Conclusion</b>	<b>82</b>
<b>8. Discussion</b>	<b>83</b>
<b>9. Limitations</b>	<b>84</b>
<b>10. Recommendations</b>	<b>85</b>
10.1 Recommendations for Accenture Song	85
10.2 Recommendations for Design Researchers and Design Teams	86
<b>Reflection</b>	<b>87</b>
<b>Reference List</b>	<b>88</b>
<b>Appendix</b>	<b>92</b>

# 1. Project Introduction

## 1.1. Context

Global healthcare systems are facing increasing pressure due to an accelerating mismatch between demand and capacity. By 2030, the world is expected to face a shortfall of nearly 10 million physicians, nurses, and midwives (McKinsey, 2020). The combination of an ageing population, rising chronic illnesses, and growing expectations for accessible and high-quality care continues to strain national health services. In many countries, patients wait an average of around 70 days for specialist appointments, and one in three report that their health worsened while waiting for care (Philips, 2025). These figures illustrate a structural imbalance that cannot be solved by expanding the workforce alone; it requires systemic innovation that redefines how care is delivered.

At the same time, healthcare is becoming increasingly data-driven. The rapid growth of digital records, mobile health apps, and wearable sensors enables the continuous collection of physiological and behavioural data (Huhn et al., 2022). These technologies generate valuable real-world evidence that could help shift care from reactive to preventive models. By capturing daily fluctuations in activity, sleep, or vital signs, they allow a more detailed and timely understanding of health than periodic consultations ever could. Yet the data infrastructure that is necessary for this transformation remains far from perfect. Health information is still fragmented across incompatible systems, data quality varies, and interoperability barriers limit the integration of insights from different sources (Dinh-Le et al., 2019).

Within this emerging ecosystem, artificial intelligence (AI) is widely regarded as a key driver of healthcare transformation. AI systems can analyse vast quantities of data, automate repetitive administrative work, and support clinical decision-making. Studies predict that AI could automate up to half of all documentation tasks, enhance diagnostic accuracy, and enable earlier interventions (McKinsey, 2020; Philips, 2025). In doing so, AI offers a route to alleviate workforce shortages and extend healthcare capacity. Yet these opportunities are accompanied by new forms of uncertainty. Patients and professionals alike express hesitation toward relying on systems whose reasoning processes remain opaque. Concerns about bias, data security, and accountability highlight that technological progress alone is insufficient; public confidence depends on the ability to make AI understandable, transparent, and trustworthy.

As AI becomes integrated into patient-facing services - from digital symptom checkers to virtual health assistants - design assumes a crucial mediating role. Clear communication of how recommendations are generated, and why certain data influence them, helps users feel informed and in control. Explainability and user experience (UX) design are therefore essential to ensure that intelligent systems remain aligned with human expectations and values. They bridge the gap between complex computation and everyday understanding, shaping whether technology strengthens or undermines trust in healthcare.

## 1.2. Accenture Song

Accenture Song is the design and innovation arm of Accenture, one of the world's largest professional services firms. Operating in more than 120 countries, Accenture Song combines strategy, design, marketing, and technology to help organisations create meaningful experiences and strengthen brand relevance (Accenture, 2026). The company blends creativity with data-driven insight to translate technological progress into human-centred solutions. By integrating design thinking with digital transformation, Accenture Song enables businesses to remain adaptive and trusted in rapidly changing environments.

Within Accenture Song, the Design & Digital Product (D&DP) team focuses on shaping end-to-end digital experiences. The team combines user research, interaction design, and technology development to translate business challenges into validated solutions. Its work spans early-stage concept development and prototyping through to large-scale implementation, serving clients across multiple sectors.

Accenture has made significant global investments in artificial intelligence, recognising it as a defining technology for the next decade. In 2023, the company announced a USD 3 billion investment to expand its AI capabilities and help clients across industries reinvent their operations through scalable AI solutions (Accenture, 2023). This strategic commitment enhances Accenture's capacity to embed AI into design and digital product development, ensuring that creative teams such as those at Accenture Song can apply advanced technologies effectively in client projects.

This research is therefore directly relevant to Accenture Song's design practice. While it focuses on patient-facing AI in healthcare, its insights extend to other domains in which the company operates. Accenture Song

supports clients in sectors such as banking, insurance, and pension management, where trust likewise determines whether people are willing to engage with intelligent, data-driven systems when their own personal data is used. Findings from this study can help the company better understand how user experience and explainability influence confidence in AI, and how design can make advanced technology feel reliable, comprehensible, and human-centred. By examining these dynamics within healthcare, this project contributes to the development of a framework and design principles that are transferable across industries, reinforcing Accenture Song's broader mission to create digital experiences that people can genuinely trust.

The logo for Accenture Song, featuring the word "Accenture" in a bold, black, sans-serif font, followed by the word "Song" in a lighter, blue, sans-serif font. The "S" in "Song" is notably larger and more stylized than the other letters.

Figure 1.1: Accenture Song logo

### 1.3. Research Intro + RQ's and SQ's

Given the current challenges in the healthcare market and the gap in designing for trust in patient-facing AI, combined with the technological opportunities outlined above, this thesis focuses on designing user-centred visualizations to foster trust and understanding in patient-facing agentic AI healthcare systems. The project responds to the growing need for transparent, comprehensible, and empathetic user experiences as AI becomes an active participant in healthcare interactions.

The main outcome of this graduation project will be a design framework that provides actionable guidelines and strategies for creating transparent and trustworthy user interactions with AI agents in healthcare. This framework will synthesise design principles, visualization methods, and interaction flows, offering practical support for designers and healthcare stakeholders seeking to implement explainable AI that patients can understand and trust. The central research question guiding this thesis is:

***How to design user-centred interfaces to enhance patients' trust and understanding in an agentic AI health assistant?***

The focus of this research lies not in backend data management or interoperability, but in the front-end experience—empowering patients to understand how AI systems utilise their health data and arrive at their recommendations. By addressing design principles, visualization strategies, and interaction flows, the research aims to develop actionable methods that help users comprehend and trust AI-driven healthcare insights while feeling confident about sharing their data.

To achieve this, the project is structured around the following sub questions:

1. What information and interaction flows enable patients to understand and feel confident about how their data are used within agentic AI systems?
2. Which design principles and visualization strategies best communicate AI reasoning and outcomes to patients in a clear, trustworthy, and human-centred way?
3. How can patients' trust and understanding be assessed and validated through interaction with patient-facing AI prototypes?

The answers to the research question and sub-questions are synthesised and discussed in the Conclusion, drawing on the study results and design outcomes. Together, these sub questions guide the project across the Double Diamond process. The first question informs the Discover and Define phases, focusing on understanding patient needs, data flows, and trust barriers. The second directs the Develop phase, translating insights into design principles and visual prototypes. The third shapes the Deliver phase, where prototypes are tested and validated to evaluate how visualizations influence patient trust and understanding.

## 1.4 Project Approach

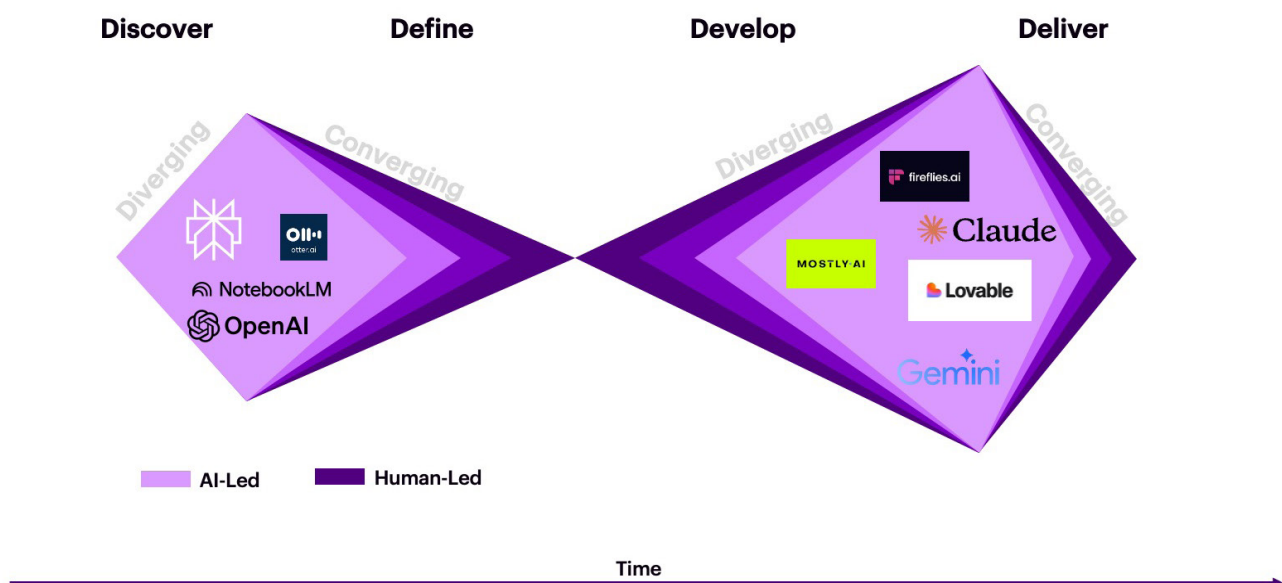
This graduation project follows a design-led research approach structured around the Double Diamond model. The Double Diamond provides an explicit progression from broad problem exploration (Discover) to focused framing (Define), followed by divergent solution generation (Develop) and convergent consolidation into final outputs (Deliver). This structure is appropriate for a design-focused thesis because it makes the relationship between research insights, design decisions, and resulting artefacts explicit, while still allowing iterative refinement across phases.

Across the project, the central focus remains on the front-end experience of explainability in patient-facing agentic AI: how an AI system's reasoning is communicated through interaction, visualisation, and tone to support trust and understanding. The project deliberately does not aim to improve back-end clinical performance or data interoperability. Instead, it investigates how interface design can mediate user confidence in AI-driven health guidance and how explanation depth can be staged to match users' informational needs.

### 1.4.1 AI Declaration

Because this thesis investigates how AI systems shape patient-facing interfaces—and how explainability can be designed for agentic assistants—the project itself was developed with the support of AI tools. AI was used as an enabler for exploration and iteration, helping to accelerate concept generation, synthesis, and prototyping across the process. This allowed the project not only to study AI as the subject of the interface, but also to reflect on how AI can influence the design process.

Figure 1.2 visualises the AI tools used across the Double Diamond and positions them in relation to the Stingray Model discussed in Section 1.4.3. The figure should be read as a project-specific interpretation of how AI-supported activities and human-led decision-making interacted over time: AI contributed to speed and breadth during divergence (generating alternatives and directions), while convergence remained guided by human judgement through selection, synthesis, and design justification.



**Figure 1.2:** Project-specific interpretation of the Stingray Model, indicating the AI tools used across the Double Diamond

## 1.4.2 Mixed Method Approach

To address the research question and sub-questions, the thesis applies a sequential mixed method approach, in which different methods serve distinct roles across the design trajectory.

Qualitative methods are used primarily to build understanding of the problem and to define design requirements. These inputs include the synthesis of prior research on trust and explainability and exploratory expert interviews that clarify how trust is established and disrupted in patient-facing communication, and what risks arise when explanations are overly technical or emotionally misaligned. The qualitative phase is used to derive and refine design principles that can be operationalised in interface prototypes.

Quantitative and structured evaluation methods are used in the Develop phase to compare interface alternatives. Three distinct prototype concepts are tested through an in-person within-subject (A/B/C) study design, in which each participant interacts with all concepts and completes the same measurement blocks after each condition. This structure supports controlled comparison between concept directions while limiting the influence of individual differences, since each participant serves as their own reference point.

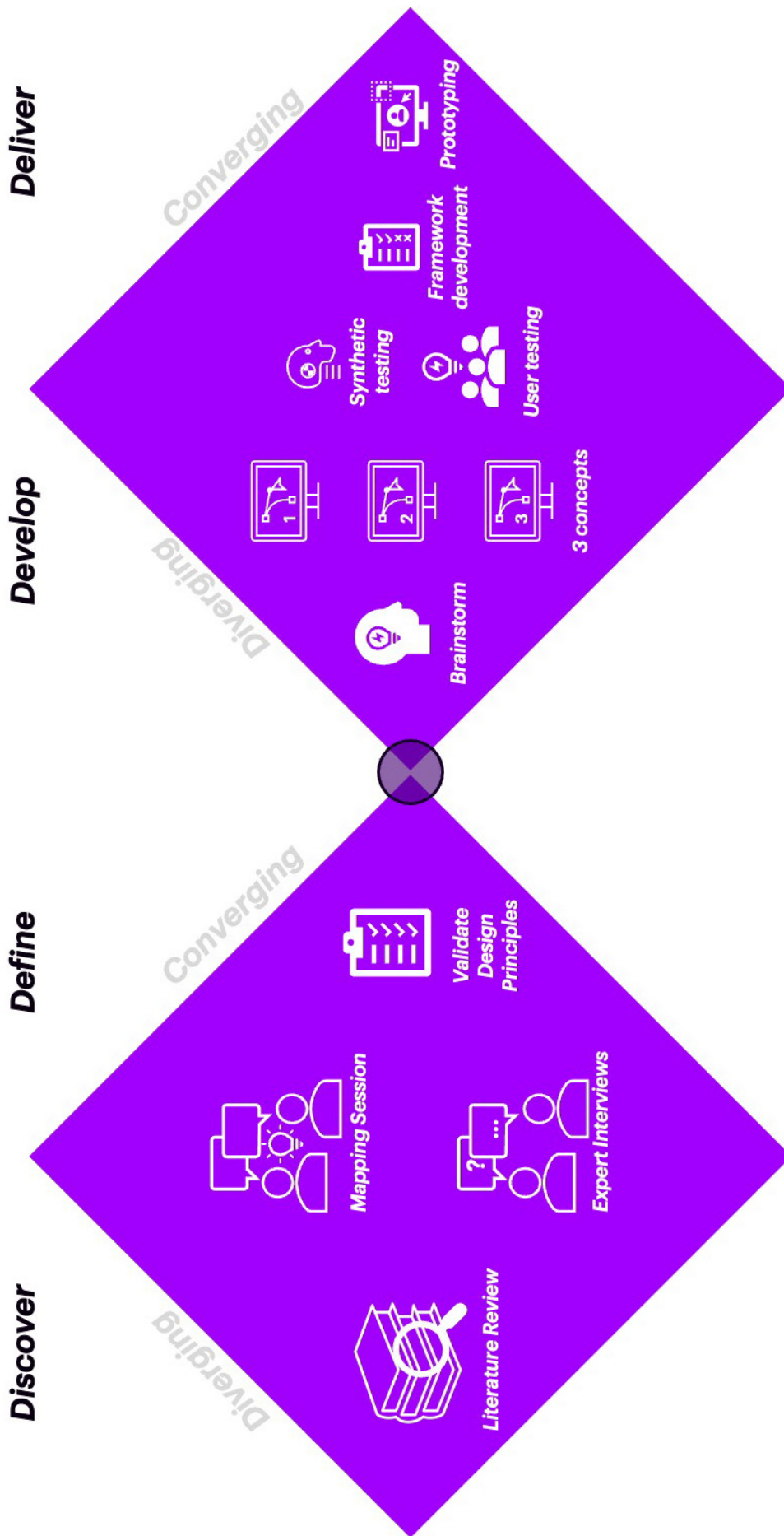
Finally, the project includes a demonstrator prototype as an implementation-oriented component. This prototype is used to illustrate how the framework can be instantiated in an end-to-end interaction using a synthetic patient dataset and an orchestrated workflow. Its purpose is to demonstrate the feasibility of staged explanation behaviour, rather than to validate clinical correctness.

## 1.4.3 Double Diamond and Stingray

The activities of the thesis map onto the Double Diamond phases as follows:

- **Discover:** the project establishes theoretical and domain understanding through a focused literature review on trustworthy AI and explainability, complemented by exploratory inputs that clarify what users need in order to feel informed and safe when interacting with AI health guidance.
- **Define:** insights are synthesised into a clear design direction, including design principles and a testable scenario that reflects a realistic, non-acute patient-facing interaction. This phase frames what will be evaluated and why.
- **Develop:** multiple concept directions are generated using the design principles as constraints (e.g., differences in interaction depth, visual support, and transparency strategy). These concepts are prototyped and prepared for comparative evaluation.
- **Deliver:** the final phase consolidates the research and evaluation outputs into a set of design artefacts: the framework and supporting materials (playbook, roadmap, and prototype) that translate research insights into actionable guidance for design practice.

In addition to the Double Diamond, the project draws selectively on the Stingray Model (Board of Innovation, 2025) as an execution scaffold for fast, structured cycles of exploration and validation when working with AI tools. Here, Stingray clarifies where AI accelerates divergence (idea generation, variation, rapid drafting) and where human-led convergence remains essential (framing, synthesis, and justification of the final artefacts). Figure 1.2 complements the Double Diamond by visualising this interaction across the project timeline.



# Literature Review

Chapter 2 situates this thesis within the broader transformation of healthcare through AI, with a specific focus on the emergence of agentic AI systems and the conditions required for them to be adopted in practice. It begins by outlining the promise of AI in healthcare—a shift toward more preventive, personalised, and continuous forms of care—while acknowledging that real-world implementation remains constrained by systemic and human factors rather than technical capability alone.

The chapter then progressively narrows toward the specific problem space this thesis addresses. It moves from (1) the evolution of AI in healthcare and its drivers, to (2) what distinguishes agentic AI from earlier predictive or task-based systems, including the move toward multi-agent architectures and end-to-end orchestration across care pathways. Next, it examines (3) the healthcare data landscape and the shift from institution-controlled records to patient-owned and interoperable ecosystems, framing why data access, control, and transparency become foundational in an AI-mediated care model. Building on this, the chapter addresses (4) the trust paradox: despite optimism about AI's potential, adoption is limited when stakeholders lack confidence in how AI is used and governed, and when trust research focuses on measuring attitudes instead of enabling trustworthy experiences. Finally, it positions (5) explainability as a human-facing challenge—relevant not as a purely technical concept, but as part of how AI reasoning is communicated through interfaces to support understanding and confidence.

Each subsection concludes with a short Synthesis that consolidates the key takeaways and explicitly connects them to the next section, ensuring a clear line from the macro-level promise of agentic AI in healthcare to the specific design-relevant gap this thesis targets.

# 2



## AI used in this chapter:

- Perplexity for finding additional sources
- NotebookLM for analysing sources and extracting information
- ChatGPT for rewriting text

## 2.1. The Evolution of AI in Healthcare

Over the past several decades, healthcare in general has undergone profound digital transformation, with artificial intelligence (AI) and machine learning (ML) emerging as pivotal technologies reshaping clinical practice, operations, and patient engagement. While early neural network research in the 1940s laid the conceptual groundwork, AI's potential in healthcare really started in the twenty-first century with the rise of electronic health records (EHRs), wearable devices, and large-scale data availability (LaBoone & Marques, 2024).

### 2.1.1 Contemporary Applications and Drivers

Modern AI now extends far beyond rule-based expert systems. Advances in machine learning and deep learning (DL) have enabled applications capable of complex tasks such as diagnostic imaging, risk classification, natural language processing, and clinical decision support (McKinsey, 2020; Sadeghi et al., 2024). Convolutional neural networks have begun to rival or even surpass clinicians in image analysis, while ML models extract predictive insights from longitudinal EHR data to inform clinical care and resource management (Taleb et al., 2025). Together, these developments illustrate the field's shift from knowledge-based reasoning toward adaptive, data-driven decision-making.

Recent surveys highlight the systemic pressures driving AI adoption. Patients now face average global wait times of around seventy days for specialist care, with more than one-third reporting worsened health due to delays and over one in four requiring hospitalization as a result (Philips, 2025). Simultaneously, clinicians lose valuable time to administrative inefficiencies: 77 percent report losing clinical time because of incomplete or inaccessible patient data—equivalent to roughly twenty-

three working days per professional each year (Philips, 2025). Such inefficiencies reveal the pressing need for digital solutions capable of improving accessibility, continuity, and decision-making across healthcare systems.

Healthcare professionals – such as doctors, dentist and general practitioners - also express confidence in AI's transformative potential. Surveys indicate that 84 percent believe AI can automate repetitive tasks, 77 percent expect it to improve patient throughput, 78 percent anticipate expanded capacity to serve more patients, and 76 percent believe it can reduce wait times (Philips, 2025). Beyond operational improvements, 82 percent of clinicians believe AI and predictive analytics could save lives by enabling earlier interventions, while 75 percent expect digital health technologies to reduce hospital admissions. Economically, widespread AI adoption could reduce healthcare spending by five to ten percent in the United States alone, where cardiovascular diseases account for more than \$400 billion annually in direct and indirect costs. These figures collectively signal that the motivation for AI integration is not solely technological enthusiasm but also a response to systemic inefficiency and rising care demand (Philips, 2025).

### 2.1.2 Transition Toward Preventive and Personalized Medicine

AI technologies are already moving from research into practice. Systems now support oncology diagnostics, virtual health coaching, and remote monitoring across real-world contexts. Comparative studies show that AI-based symptom checkers such as Ada Health and ChatGPT achieve diagnostic accuracies approaching those of physicians — 30-40 percent for top-1 and 50-63 percent for top-3 diagnoses, compared with 47 and 69 percent for doctors, respectively (Fraser et al., 2023). These findings indicate a growth between algorithmic intelligence and clinical expertise in narrowly defined domains and exemplify the field's shift toward collaborative human-AI decision-making.

Recent analyses describe this transition as 'Medicine 3.0,' emphasizing the shift from reactive treatment toward preventive, data-driven care (Boston Consulting Group, 2025). AI-driven personalization is now able to combine data from genetics, daily habits, and wearable sensors to create tailored insights for each individual. At the same time, emerging tools such as smart implants and voice-based biomarkers enable earlier and less invasive detection of diseases, as well as continuous adjustment of treatments. This shift from treating illness to preventing it changes how healthcare is designed—from one-time appointments to ongoing, data-informed relationships between patients and providers.

### 2.1.3. Democratization and Ongoing Challenges

The implementation of AI in healthcare is already visible. One in six adults now use AI-enabled chatbots for health information, rising to 25 percent among those under 30 (Boston Consulting Group, 2025). Remote diagnostics and continuous monitoring have the potential to expand access for underserved populations, while AI-driven decision-support tools provide clinicians with instant, evidence-based

insights that reduce diagnostic uncertainty and improve treatment precision.

### 2.1.4 Synthesis

Despite these advances, challenges surrounding data fragmentation, privacy, and regulatory complexity continue to limit full integration. Only 38 percent of healthcare professionals feel that new technologies are designed around their needs, even though 69 percent are actively involved in their development (Philips, 2025). This tension between rapid technological progress and limited user-centred design reveals that the transformation of healthcare is as much a design challenge as a technical one. Overall, AI is demonstrably improving diagnostic accuracy, streamlining workflows, and enabling proactive, patient-centred models of care (LaBoone & Marques, 2024; Taleb et al., 2025). Driven by interconnected data infrastructures and interoperable health platforms, the field is shifting toward a more personalized, predictive, and equitable system—laying the foundation for the emergence of more autonomous and agentic AI, discussed in the next section.

## 2.2. Agentic AI Systems in Healthcare Context

### 2.2.1 From Predictive to Agentic AI

Building on the evolution of AI in healthcare, recent advances in foundation models and large language models (LLMs) have introduced a new paradigm: agentic artificial intelligence (AI). These systems differ from earlier predictive or rule-based approaches by exhibiting autonomous reasoning, goal-directed behaviour, and end-to-end workflow execution. In contrast to conventional automation, agentic AI enables systems to interpret intent, plan and coordinate actions, and learn continuously within dynamic environments.

The growing interest in agentic AI reflects both the rising complexity of healthcare delivery and the expanding capabilities of LLMs (McKinsey & Company, 2024b). Early AI agents—autonomous software entities engineered for goal-directed task execution—perform within bounded digital environments, perceiving inputs, reasoning over context, and acting to accomplish specific objectives (Sapkota et al., 2026). Their foundational traits of autonomy, task-specificity, and adaptive reactivity enable modular and efficient behaviour but keep them constrained to narrowly defined goals (Sapkota et al., 2026).

### 2.2.2 From Single to Multi-Agent Systems

Agentic AI extends these capacities beyond single-agent autonomy. It coordinates multiple specialised agents that communicate, share memory, and align their actions toward higher-level objectives (Sapkota et al., 2026). These systems display broader autonomy, meta-learning, and inter-agent communication, allowing them to plan, adapt, and self-correct across changing contexts. This architectural shift—from task-bounded automation to distributed, cooperative intelligence—

introduces new potential for complex, safety-critical environments such as healthcare.

Agentic architectures are commonly realised through multi-agent systems (MAS), in which specialised agents collaborate under a central orchestrator (Heydari et al., 2025). Inspired by modular cognition and multidisciplinary teamwork, MAS frameworks coordinate agents such as those in the Personal Health Agent (PHA) model as can be seen in Figure 1.1, combining Data Science, Domain Expert, and Health Coach agents (Heydari et al., 2025). Agents employ reasoning strategies such as hierarchical planning and contextual inference to manage dependencies and execute multi-step action (Heydari et al., 2025; Wang et al., 2025).

(b) Architecture of the Personal Health Agent (PHA)

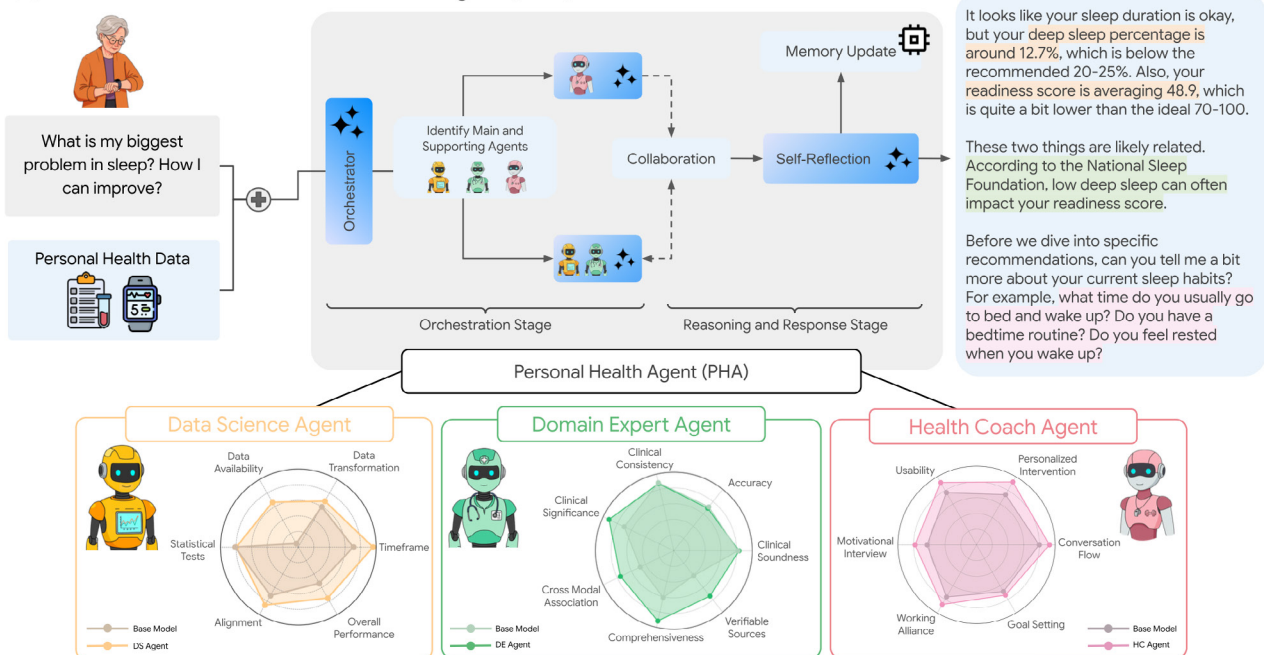


Figure 2.1: Architecture of the Personal Health Agent (Heydari et al., 2025)

### 2.2.3 Applications in Healthcare Practice

In healthcare, agentic AI enables autonomous orchestration across complete care pathways, automating patient episodes rather than isolated tasks (Gartner, 2025a; Taleb et al., 2025). In addition, by integrating data from connected medical devices and wearables, these systems could support continuous monitoring and real-time intervention. (LaBoone & Marques, 2024a; Taleb et al., 2025). Combining personal health records with domain expertise allows generation of individualised treatment plans attuned to genetic, behavioural, and lifestyle factors (Heydari et al., 2025). For instance, if a patient moves less because they are injured, the agent does not push more activity but adjusts recommendations to support recovery. Multi-agent coordination also mirrors the collaborative dynamics of human clinical teams, enhancing coherence in decision-making.

Agentic AI demonstrates potential value across operational and clinical domains. In hospital management, agents can forecast staffing needs, adjust schedules, and streamline patient flow through automated discharge planning and workflow tracking (Gartner, 2025a; McKinsey & Company, 2020a). In direct patient care, they can coordinate triage, manage appointments, and deliver adaptive treatment guidance, including virtual nursing support (LaBoone & Marques, 2024a; Taleb et al., 2025). Beyond these immediate efficiencies, Sapkota et al. (2026) describe how agentic AI underpins collaborative medical-decision support: specialised agents handle diagnostics, vital-sign monitoring, and treatment optimisation through shared memory to produce coherent, safe recommendations.

### 2.2.4 Synthesis

In summary, agentic AI marks a transition from static, rule-based automation to adaptive, self-improving systems capable of orchestrating

complex clinical and operational processes. By combining autonomy, adaptability, continuous learning, and contextual awareness, these systems establish the technical foundation for the next generation of healthcare intelligence. This evolution also signals a broader paradigm shift: healthcare AI is no longer a passive decision-support tool but an active collaborator capable of shaping and mediating interactions among patients, professionals, and data systems. For design research, this shift opens critical questions about how such autonomy should be experienced, governed, and made transparent within clinical and patient-facing contexts.

## 2.3. Healthcare Data Landscape and Patient Ownership

The transition from institution-controlled health data to patient-owned, interoperable ecosystems represents a fundamental reconfiguration of healthcare infrastructure. Interoperability is the ability of different information systems to communicate, exchange data, and interpret that data so it can be used in an AI system for instance. While current systems remain fragmented and provider-centric, literature increasingly envisions a future in which individuals hold and manage their own health information. This shift is structural, redistributing control, agency, and accountability in ways that make AI's data requirements practically feasible.

### 2.3.1 The As-Is: Fragmented and Institution-Controlled Data Systems

Contemporary healthcare data environments are dominated by institutional ownership and siloed information systems. Health data are stored across heterogeneous Electronic Health Record (EHR) platforms that are often incompatible and inaccessible outside the originating institution (Haddad et al., 2023; Mollerus et al., 2025). Despite digitization, interoperability remains limited: differences in standards, terminology, and governance

prevent seamless data exchange across organizations (David, 2024). Consequently, patients cannot easily access, consolidate, or transfer their information, and clinicians struggle to obtain a complete overview of a person's medical history.

Beyond technical fragmentation, the issue is one of control. EHRs are typically maintained by providers rather than by the individuals whose data they contain (Haddad et al., 2023). Institutional priorities, commercial interests, and privacy regulations converge to restrict data mobility and reinforce centralized authority (Kruse et al., 2016, as cited in Mirchev et al., 2020). Although regulatory frameworks such as the General Data Protection Regulation (GDPR) and HIPAA give patients certain access and consent rights, they stop short of recognizing data as a form of personal property (Mohammed & Malhotra, 2024). As Mirchev et al. (2020) note, privacy is protected, but ownership remains undefined—leaving individuals with limited influence over how their information is used.

In contrast and parallel to provider-held EHR data, a growing share of health-relevant information is generated outside clinical settings through wearables, smartphones, and self-tracking apps. This “patient-generated health data” (PGHD) includes device-captured biometrics (e.g., activity, heart rate, sleep), symptom diaries and patient-reported outcomes, and lifestyle logs such as diet, hydration, and medication adherence—often recorded by patients or family caregivers. Demiris et al. (2019) and Shapiro et al. (2012) emphasise that PGHD can complement episodic clinical records by capturing day-to-day behaviours and context that are typically invisible in routine care.

For the integration of AI in healthcare, AI systems depend on large, longitudinal, and interoperable datasets to generate reliable and

equitable insights. When data are fragmented or inaccessible, algorithmic performance declines, and the possibility of personalized, adaptive care remains unrealized. The as-is structure thus constrains both clinical efficiency and the ethical deployment of AI technologies.

### 2.3.2 The To-Be: Patient-Owned and Interoperable Ecosystems

In contrast, recent scholarship outlines a to-be scenario in which patients become the primary stewards of their health data. The concept of Personal Health Records (PHRs) illustrates this transition: systems that allow individuals to access, manage, and selectively share their medical information (Fang et al., 2021). PHRs invert the traditional hierarchy by placing data control in the hands of the patient rather than the provider. Such frameworks represent the conceptual foundation for patient data ownership, in which health information is treated as a personal asset governed by consent and autonomy rather than institutional custodianship. A real-life example of this, is Sweden's 1177 health portal and app. Through secure login, individuals can access and manage their own medical records, view prescriptions and test results, book appointments, and interact with care services directly (1177.se, 2026).

Emerging technological and policy developments support this paradigm. Decentralized architectures such as blockchain-based management systems embed ownership at the infrastructural level by allowing individuals to retain encrypted records while maintaining verifiable audit trails (Haddad et al., 2023). Similarly, emerging identity technologies such as Self-Sovereign Identity (SSI) and Decentralized Identifiers (DIDs) allow individuals to verify their identity and manage access to their health information through secure digital wallets (Welzel et al., 2025). These systems decentralize authentication, giving users

control over who can view or use their data. In parallel, the European Health Data Space (EHDS) proposes a framework for transparent, cross-border data use in which patients grant consent for national and international health research (Baumgart, 2024).

These developments do not merely address technical inefficiencies; they redefine the ethical and operational foundations of healthcare. Patient ownership aligns with broader trends toward autonomy and participatory governance, positioning individuals as active participants in their data ecosystems. As Kish and Topol (2015, cited in Mirchev et al., 2020) argue, “for the benefits of digital medicine to be fully realized, we not only need to find a shared home for personal health data but also give individuals the right to own them.” By enabling people to manage and share their information directly, ownership becomes both a right and an enabler of system-wide intelligence.

### **2.3.3 Why Patient Ownership in the Age of AI**

The envisioned shift toward patient-controlled data ecosystems is widely regarded as essential for realizing the full potential of AI in healthcare. Machine learning models require large, representative, and ethically sourced datasets that reflect the diversity of real-world populations. When individuals can combine and share their health information across providers and devices, these datasets become richer and more inclusive—supporting the development of fairer and more reliable algorithms. Fang et al. (2021) associate PHR adoption with improved outcomes and reduced costs, suggesting that empowered patients not only contribute to better data but also benefit from more personalized, data-driven care.

Moreover, ownership introduces transparency and accountability into data flows, allowing individuals to understand and govern how

their information fuels digital systems. This rebalancing of power supports the broader societal shift from data extraction to data collaboration. As Baumgart (2024) notes, collaborative data ecosystems depend on a “cultural change with the establishment of trust in a precompetitive space.” By embedding control at the individual level, patient ownership provides the preconditions for that cultural shift, enabling AI systems to operate within ethical and participatory boundaries.

### **2.3.4 Synthesis**

The evolution from fragmented, provider-controlled systems to patient-owned and interoperable data ecosystems reflects a structural realignment of healthcare around autonomy, transparency, and shared accountability. While the as-is landscape restricts data mobility and limits AI’s potential, the to-be paradigm reimagines patients as data custodians whose control and consent are integral to the functioning of intelligent health systems. Literature across technological, ethical, and policy domains converges on the view that such ownership is not only desirable but necessary for the sustainable integration of AI into care.

## **2.4 The Trust Paradox in Healthcare AI**

Trust has become one of the defining requirements for the responsible adoption of Artificial Intelligence (AI) in healthcare. As intelligent systems increasingly support or even automate clinical decisions, their success depends less on technical capability than on whether people—patients, clinicians, and institutions—are willing to rely on them. Trust therefore functions as the social and ethical infrastructure upon which the digital transformation of healthcare must rest.

### 2.4.1 Why Trust Matters

Trust is foundational to healthcare because the clinical relationship itself is built on vulnerability, uncertainty, and dependency. Patients routinely rely on healthcare professionals to act in their best interest; integrating AI adds a further layer of abstraction and risk (Goisauf et al., 2025; Starke et al., 2025). In this context, trust operates as both a psychological enabler—allowing users to engage with systems they cannot fully understand—and a practical condition for adoption (Steerling et al., 2023). When trust is present, users are more willing to rely on AI-assisted decisions; when it is absent, adoption stalls even when accuracy is demonstrated (Starke et al., 2025).

Trust also mediates how patients interpret AI's potential benefits. Nong and Ji (2025) found that individuals who place greater trust in their healthcare systems and providers are significantly more likely to expect that AI will improve affordability, access, and patient-provider relationships. This indicates that expectations of AI are not determined solely by perceived technical quality but are embedded within broader experiences of fairness, inclusion, and institutional integrity.

### 2.4.2 The Trust Gap

Empirical research consistently identifies a trust gap between healthcare professionals, institutions, and patients (Philips, 2025). Academic research on clinician acceptance of AI-driven clinical decision support systems found that trust and performance expectancy are the strongest predictors of clinician intention to use AI systems (Dingel et al., 2024). In a meta-analysis of 17 studies comprising 3,871 healthcare practitioners, Dingel and colleagues (2024) examined the Unified Theory of Acceptance and Use of Technology (UTAUT) framework and identified trust as emerging as the most influential predictor overall, accounting for 29–35% of the variance explained in use intention, followed

by performance expectancy at 19–24% of the variance. This provides important context for understanding why 82% of clinicians express belief in AI's potential (Philips, 2025)—the underlying mechanisms relate to demonstrable clinical performance and the extent to which practitioners perceive the technology as trustworthy and reliable in clinical contexts. While these studies show clinicians tend to be optimistic about AI's potential to improve care, patients and the general public remain cautious. A global survey reported that 79 percent of healthcare professionals believe AI can enhance outcomes, compared with only 59 percent of patients (Philips, 2025).

National evidence reinforces this asymmetry. Nong and Platt (2025) found that 65.8 percent of U.S. adults expressed low trust in their healthcare systems to use AI responsibly, and 57.7 percent doubted that AI tools would avoid causing harm. Experiences of discrimination were significantly associated with lower levels of trust, while higher general trust in the healthcare system predicted greater confidence in responsible AI use (Nong & Platt, 2025). These findings reveal that attitudes toward AI mirror existing inequities in healthcare trust: AI inherits—not erases—the social context in which it is deployed.

At the same time, professionals' confidence in AI often exceeds their willingness to delegate clinical judgment. Studies indicate that clinicians tend to trust AI as an assistant rather than as a decision-maker, illustrating the persistence of a "trust paradox": stakeholders acknowledge AI's potential yet remain reluctant to rely on it fully in practice (Goisauf et al., 2025; Starke et al., 2025).

### 2.4.3 The Real Gap — From Measuring to Designing for Trust

While trust is widely recognized as critical, literature shows that most research merely measures trust instead of designing for it. Current studies typically assess trust through cross-sectional surveys that treat it as an individual psychological variable—how much a person “trusts” an AI system—without addressing the contextual or relational conditions that make trust sustainable (Steerling et al., 2023; Starke et al., 2025). This approach, though useful for benchmarking attitudes, fails to explain how trust develops over time or through user experience.

Authors therefore call for a shift from trust metrics to trust frameworks. Rather than asking how much people trust AI, research should examine how systems and organizations can become trustworthy in practice (Starke et al., 2025; Goisauf et al., 2025). Starke et al. (2025) warn that attempts to “increase” trust without increasing trustworthiness risk producing unwarranted trust—confidence in systems that may not deserve it—or unwarranted distrust, where users reject beneficial technologies due to poor design or communication. This is visualized in Figure 2.2

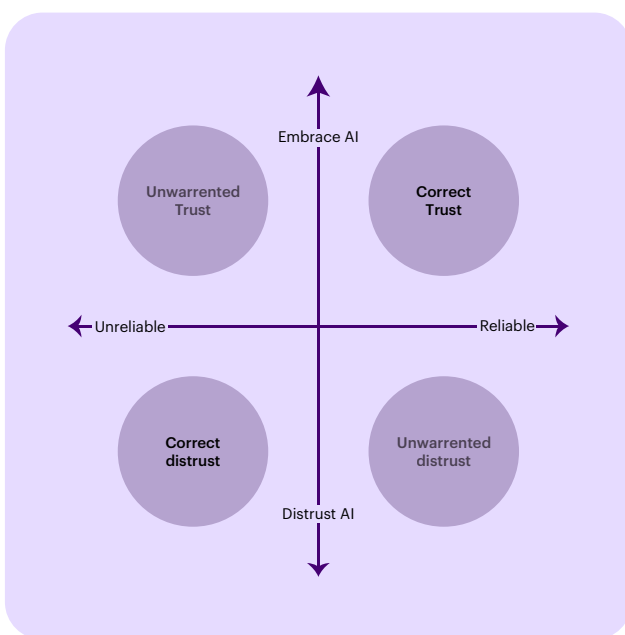


Figure 2.2: Trust / Distrust Matrix

The emerging consensus emphasizes the need for design for trust, approaches that integrate human experience, communication, and interaction into the AI development lifecycle (Goisauf et al., 2025; Lekadir et al., 2025). Human-centred frameworks such as FUTURE-AI – which provides guidance for the development and deployment of trustworthy AI tools in healthcare – identify usability and stakeholder involvement as core principles for building trustworthy systems (Lekadir et al., 2025). These approaches highlight that trust is not simply given but co-produced through transparent processes, accessible interfaces, and consistent collaboration with users.

However, this design perspective remains underdeveloped. Few studies systematically explore how trust is shaped by user experience, workflow integration, or communication design (Starke et al., 2025). The absence of empirically tested design models constitutes the most significant gap in current research. Addressing this gap requires exploratory and case study research that examines how trust is affected through interaction, rather than treating it as a static input.

### 2.4.4 Synthesis

Current literature converges on the view that trust is critical yet fragile. Patients remain cautious, clinicians selectively optimistic, and researchers acknowledge that trust cannot be mandated, it must be earned. The persistence of low public trust, evidenced by the findings of Nong and Platt (2025), underscores a disconnect between technological advancement and human acceptance. Conversely, the association identified by Nong and Ji (2025) between trust and positive expectations of AI demonstrates that strengthening trust can enhance engagement and perceived value.

Overall, addressing the trust paradox requires moving from measuring attitudes to designing for justified trust, a process that aligns system behaviour, communication, and user experience. This transition forms the conceptual foundation for the next chapter, which explores explainability as a design mechanism for fostering trustworthy and transparent healthcare AI.

## 2.5 Explainable AI as a Human-Facing Design Challenge

Explainable Artificial Intelligence (XAI) is increasingly recognised as a crucial mechanism for cultivating trust and trustworthiness in automated systems (Adeniran et al., 2024; Sadeghi et al., 2024). In healthcare, where decisions directly affect safety and wellbeing, explainability is essential not only for validating performance but also for ensuring ethical acceptance and user confidence. Clinician perspectives on AI are not uniformly positive: The opacity of complex AI models—often referred to as “black boxes”—does create uncertainty and scepticism among clinicians and patients alike (Miller, 2019; Hulsen, 2024). By making the reasoning behind AI outputs transparent and understandable, XAI provides the cognitive and ethical clarity necessary for users to rely on intelligent systems (Sadeghi et al., 2024; Loh et al., 2022).

In this thesis, XAI is approached not as a purely technical exercise in model interpretability but as a human-facing design challenge. The focus lies on how AI systems communicate their reasoning to non-expert users (particularly patients) and how such communication shapes comprehension, trust, and agency. Rather than dissecting algorithms, this perspective examines explainability as an interactional property: the interface through which people make sense of AI recommendations and decide whether to accept, question, or act upon them.

### 2.5.1 Role and Effects of Explainability in Healthcare

Explainable AI refers to strategies that make AI decision-making processes transparent and comprehensible to end-users, including clinicians and patients (Sadeghi et al., 2024; Adeniran et al., 2024). Beyond performance validation, it enhances transparency and accountability in systems that influence clinical outcomes (Adeniran et al., 2024). By revealing the logic, data, and uncertainty behind predictions, XAI allows users to scrutinise outputs and verify whether they align with professional judgement or personal context (Sadeghi et al., 2024).

Empirical evidence consistently links explainability to higher user trust and calibrated reliance. Clinicians report greater confidence in AI systems when explanations are clear, relevant, and concise (Rosenbacke et al., 2024), whereas opaque or overly complex explanations reduce confidence and adoption (Corti et al., 2024). Explainability thus acts as a trust calibration mechanism, encouraging appropriate reliance rather than blind acceptance (Sadaka & Qureshi, 2025). For patients, intelligible explanations enhance understanding, participation, and treatment adherence by clarifying why specific recommendations are made (Adeniran et al., 2024). In this sense, explainability is a functional requirement, ensuring that decision-support tools remain accountable and comprehensible to those they affect.

### 2.5.2 Patient-Facing Explainability

Designing explanations for patients introduces distinct communication and usability challenges. While clinician-facing XAI emphasizes evidence provenance and diagnostic reasoning, patient-facing XAI must translate algorithmic logic into accessible, actionable, and emotionally safe information (Adeniran et al., 2024). Scholars highlight the importance of human-centred and participatory design methods that adapt

explanations to users' literacy levels, cultural contexts, and emotional readiness (Saeed & Omlin, 2023; Corti et al., 2024).

Effective explanations prioritise meaning over mechanism, focusing on what the output implies for the individual rather than how it was computed (Adeniran et al., 2024). They clarify how personal health data—such as lifestyle patterns, test results, or sensor readings—inform predictions and advice, thereby supporting agency and informed consent (Corti et al., 2024). Multi-modal presentation formats combining textual, visual, and interactive elements improve comprehension by allowing users to navigate between surface summaries and deeper insights depending on interest and capacity (Corti et al., 2024;

Behaviourally, explainability fosters engagement and adherence. Patients are more likely to follow treatment recommendations and perceive AI outputs as credible when they understand their rationale (Sadaka & Qureshi, 2025). Furthermore, well-designed explanations reinforce rather than replace the clinician–patient relationship, facilitating shared decision-making and preserving empathy within digital healthcare interactions (Hildt, 2024).

### 2.5.3 Design Principles for Patient-Facing Explainability

Research across human–computer interaction (HCI) and user experience (UX) design offers clear guidance on how patient-facing XAI should be structured. The following design principles synthesise key insights from the literature and consolidate overlapping concepts into five overarching themes that capture the cognitive, interactive, and relational dimensions of explainability.

**Cognitive clarity** combines the use of plain, everyday language with the principle of low cognitive load. Explanations should be phrased in accessible, non-technical terms and structured concisely to prevent information overload. This clarity supports comprehension and ensures that users can focus on meaning rather than deciphering terminology (Adeniran et al., 2024; Eiband et al., 2018).

**Multi-modal accessibility** emphasises presenting explanations through multiple complementary formats—such as text, visuals, and summaries—to accommodate different levels of health literacy and information-processing preferences. A well-designed multimodal interface allows users to engage with explanations in the form that best suits their abilities and context (Corti et al., 2024).

**Interactive depth control** integrates the ideas of layered explanations and on-demand interactivity. It enables users to explore reasoning at their own pace, navigating from a high-level overview to more detailed or technical rationales as desired. This progressive disclosure supports curiosity, autonomy, and trust by aligning the level of detail with the user's cognitive and emotional readiness (Eiband et al., 2018; Saeed & Omlin, 2023).

**Contextual relevance** merges actionable framing with personalisation. Explanations should not only describe what the AI has concluded but also what it means for the individual—linking outputs to actionable insights or next steps. Tailoring content to the user's health condition, data profile, and prior knowledge increases perceived usefulness and fosters engagement (Adeniran et al., 2024).

**Two-way transparency** extends explainability beyond one-way communication. By embedding opportunities for feedback and contestability, patients can ask questions, correct inaccuracies, or express uncertainty. Such reciprocal interaction transforms explanation into dialogue, reinforcing accountability and mutual trust between users and intelligent systems (Saeed & Omlin, 2023).

Together, these five principles define a user-centred foundation for explainable healthcare AI. They emphasise that explainability is not merely a technical function but a design challenge grounded in communication, interaction, and participation. These five principles are used as design criteria in Chapter 4 and guide concept ideation and comparison in Chapter 5.

#### 2.5.4 Synthesis

Explainable AI serves as the practical bridge between algorithmic reasoning and human understanding, transforming opaque computation into meaningful, actionable information. Literature confirms that effective explainability directly fosters trust, confidence, and patient engagement, particularly when designed around human cognitive and emotional needs. Within patient-owned data ecosystems and increasingly agentic AI systems, explainability becomes not optional but essential: it safeguards user autonomy and ensures that individuals can comprehend, question, and participate in algorithmic decision-making.

Yet current approaches remain limited by their technical orientation. Few frameworks translate explainability into concrete design practices that address patient comprehension, interaction, and emotion. This gap positions human-centred design and UX research as key enablers for the next generation of explainable healthcare systems.

## 2.6 Key Takeaways

### ***AI transforms healthcare, but its success depends on patient understanding and trust.***

The literature review outlines a healthcare landscape undergoing a profound transformation through the rise of agentic Artificial Intelligence (AI). These systems extend beyond traditional automation to reason, learn, and act autonomously across complex, data-rich environments. They promise a shift from reactive to preventive care, from episodic interactions to continuous, adaptive relationships between patients and intelligent systems. Yet, as healthcare becomes increasingly mediated by AI, its success no longer depends primarily on computational capability but on the human capacity to understand and trust what these systems do.

### ***Trust in data sharing is the foundation of patient participation in intelligent healthcare.***

A central insight concerns the evolving role of health data. The shift from institution-controlled records to patient-held ecosystems redefines how individuals participate in data-driven care. However, ownership alone does not guarantee engagement. For agentic AI to deliver meaningful insights, patients must be willing to share their personal data and trust that doing so is safe, transparent, and beneficial. Trust thus becomes the bridge between private data and intelligent care. Design should make these relationships visible—clarifying how personal data contribute to AI-generated insights, what safeguards exist, and how users retain agency. Transparent communication of data use transforms sharing from a passive act into an informed and empowering choice.

### ***Designing for trust requires focusing on the patient's experience of explainability.***

Equally, the literature identifies trust as the decisive factor for patient adoption and engagement. People do not simply trust AI because it is accurate; they trust it when it is understandable, transparent, and aligned with their values. Explainable AI (XAI) provides the conceptual foundation for this, yet current research remains dominated by technical interpretability. The gap lies in the user experience of explanation—how reasoning and uncertainty are communicated to patients. Designing for trust therefore involves crafting visual and interactive explanations that are clear, contextual, and actionable. This is not about exposing algorithmic logic but about shaping how explainability is experienced: how patients see, feel, and understand what an AI system recommends.

Collectively, the literature provides clear direction for this thesis. Future patient-facing AI must combine autonomy with transparency, intelligence with intelligibility. This framing justifies the thesis's focus on designing for trust in agentic AI systems, with particular attention to how user-centered visualization can foster patient understanding, confidence, and agency in the face of algorithmic complexity.

# Exploratory Research

# 3

Chapter 3 presents the exploratory research phase, which translates the abstract challenges identified in the literature into practice-grounded insights about how trust, understanding, and transparency are experienced and interpreted in real healthcare contexts. The goal of this chapter is not to evaluate a specific interface solution, but to surface design-relevant mechanisms and tensions—such as reassurance versus detail, accountability versus overload, and transparency versus interpretability—that must be addressed when communicating agentic reasoning to patients.

The chapter outlines the research approach, sampling strategy, and analysis process, and then synthesises the core insights that directly inform the framing decisions and refined design principles introduced in Chapter 4.



## AI used in this chapter:

- ChatGPT for rewriting of text
- Otter.ai for transcribing audio files

### 3.1 Research Approach

This chapter presents the exploratory phase of the study, which aimed to uncover how trust, understanding, and interaction are experienced and constructed in healthcare communication, and how these insights can inform the design of patient-facing artificial intelligence (AI) systems.

Given the human and contextual nature of the topic, a qualitative approach was adopted. The study combined semi-structured interviews with a participatory workflow mapping session to examine both reflective and practical perspectives on how trust is formed in healthcare settings. The interviews explored professional experiences and views on communication, data transparency, and user confidence, while the workflow session visualised how empathy and reasoning are embedded in daily routines. The interview scripts can be found in Appendix A.

The goal was not to produce generalisable results but to derive design-relevant insights that explain how trust and understanding are created through communication, tone, and structure in care interactions.

### 3.2 Sampling of Participants

Three professionals participated in the data collection phase, representing complementary perspectives on trust and communication in healthcare:

- **Clinical Expert:** A psychiatrist and researcher involved in digital health innovation, offering a perspective on patient communication, transparency, and ethical boundaries.
- **GP Triagist:** A general-practice triagist working in out-of-hours primary care, contributing practical insights into workflow, empathy, and communication structure.

- **UX/UI Design Expert:** Specialised in designing User Interface designs for clients in which trust occasionally was an important factor.

Sessions lasted between 45 and 60 minutes and were all conducted online via a Teams call. All conversations were audio-recorded and transcribed using Fireflies.ai, after which transcripts were reviewed manually for accuracy. While the automated transcription was not flawless, it provided sufficient detail to capture key ideas and expressions.

### 3.3 Data Analysis Process

Quotes and insights were organised into clusters in Miro, resulting in visual groupings of recurring patterns such as empathy and reassurance, clarity of communication, feedback mechanisms, and ethical transparency.

This process was interpretive and comparative rather than formal coding: connections between statements were mapped, compared, and refined iteratively until thematic coherence emerged.

Cross-case comparison between the interviews and the workflow mapping session revealed recurring mechanisms that describe how trust, understanding, and perceived credibility are built in human communication.

### 3.4 Design Criteria Derived from the Exploratory Findings

The exploratory phase suggests that trust in patient-facing healthcare interactions depends less on technical detail and more on how guidance is communicated: tone, clarity, structure, and the user's ability to participate. The following six design criteria translate the qualitative insights into actionable requirements for the interface concepts developed in later chapters.

#### 3.4.1 Design Criterion 1: Prioritise reassurance and emotional safety in default communication

Participants described worry as a primary driver for seeking help, meaning that the first interaction moment often carries emotional load. One practitioner summarised this directly:

*"Most people call because they're worried... Taking away that worry is really what builds trust."*

This implies that a patient-facing agent should default to calm, supportive communication that reduces uncertainty rather than escalating it through dense information or overly clinical wording. In interface terms, this prioritises restrained language, clear framing, and a default presentation that feels emotionally safe before offering optional depth.

#### 3.4.2 Design Criterion 2: Make "what happens next" explicit

Trust increased when users understood what to do and what to expect. The exploratory findings suggested that people do not necessarily want every detail, but they do need clear direction and a sense of safety in the recommendation:

*"People don't need to know every detail; they just need to understand why something is safe and what will happen next."*

This points to the need for interfaces to communicate next steps in concrete terms and to keep those steps visible and unambiguous, even when additional explanation is available elsewhere in the interface.

#### 3.4.3 Design Criterion 3: Design transparency as "helpful visibility," not total disclosure

Participants framed transparency as useful only when it supports understanding. Exposing too much detail—particularly raw clinical information without interpretation—was seen as confusing and potentially distressing: "Transparency should help, not overwhelm." This implies that transparency should be staged and curated. Supporting evidence should be summarised and contextualised, and deeper detail should appear through purposeful entry points rather than being displayed by default.

#### 3.4.4 Design Criterion 4: Provide adaptable explanation depth aligned with user needs

The interviews indicated that users vary strongly in health literacy, emotional state, and desire for detail. A fixed explanation depth therefore risks under-supporting some users while overwhelming others. One expert explicitly linked understanding and confidence:

*"Trust grows when people feel confident and understand what's going on; comprehension and confidence always go hand in hand."*

This supports designing explanation as adjustable: a shallow, reassuring default paired with optional deeper layers for users who want justification or verification.

### **3.4.5 Design Criterion 5: Maintain predictable structure and stable interaction patterns**

Trust was reinforced when communication felt structured and systematic. In real care contexts, step-wise questioning and clear progression help users feel guided rather than left uncertain. This implies that digital interfaces should keep interaction patterns stable—consistent placement of actions, predictable navigation, and clear structuring of information—so that users can form reliable expectations about how the system behaves.

### **3.4.6 Design Criterion 6: Enable participation through feedback and correction mechanisms**

Trust was described as something that emerges through two-way communication rather than passive reception. Users need ways to clarify, correct, and ask follow-up questions so that understanding becomes co-created rather than imposed. This implies that a patient-facing agent should provide explicit opportunities for participation, such as confirming interpretations, correcting data, and asking questions, so that users remain involved in how the system forms and explains its guidance.

# Define

## Framing the Design Direction

Chapter 4 presents the Define phase, in which the exploratory findings from the Discover phase are translated into a clear design direction. The purpose of this chapter is to move from what was learned to what should be designed: it consolidates the most consequential empirical insights into a focused problem framing, clarifies the boundaries and objectives of the project, and formalises the requirements that will guide concept development and evaluation.

The chapter proceeds by first articulating the key insights derived from the research, then narrowing these into a concrete focus area that defines what the project will and will not address. It subsequently formulates a design brief that specifies the intended contribution and design challenge in actionable terms. Finally, it operationalises the findings into a set of design principles, which function as evaluative criteria and as a bridge to the Develop phase, ensuring that subsequent design decisions remain traceable to the empirical basis of the thesis.

# 4



### AI used in this chapter:

- ChatGPT for rewriting of text

## 4.1 Insights

The exploratory research demonstrated that trust in patient-facing AI is not a function of computational accuracy but of how information is exchanged between people and technology. The findings collectively point to a broader insight: trust emerges from communication design, not from data processing alone. For designers, this implies that the development of explainable and trustworthy AI requires a focus on how the system speaks, listens, and adapts rather than how it calculates.

In healthcare contexts, communication between user and system operates at two intertwined levels: the content of what is said and the manner in which it is delivered. The research made clear that the success of an AI system depends on both. Users are not only interpreting outcomes; they are also evaluating the tone, structure, and credibility of the interaction. This means that the quality of trust will rely on how effectively the AI can represent its reasoning, acknowledge uncertainty, and respond in ways that feel humanly coherent.

From a design perspective, this translates into a shift of attention. The task is not to reveal the full mechanics of algorithmic decision-making but to define how those mechanisms are communicated. This requires understanding what types of data the AI must draw from to communicate meaningfully — for instance, when personal or contextual data may be necessary to make explanations relatable — and how such data should be presented transparently and ethically. The designer's responsibility, therefore, lies in translating computational logic into communicative logic.

These insights underline that designing for trust in AI involves balancing the following tension:

- Abstraction vs. specificity: Determining how much of the system's reasoning needs to be surfaced for the user to feel informed but not overloaded.

In essence, trust in healthcare AI depends on the system's ability to communicate its intelligence, to make its reasoning perceivable, its boundaries visible, and its tone empathetic.

## 4.2 Focus Area

This project draws on the Personal Health Agent (PHA) framework proposed by Heydari et al. (2025) as a representative use case for exploring human–AI interaction in healthcare. The PHA describes a multi-agent architecture in which different specialised agents — including data science, domain expertise, and health coaching agents — collaborate to analyse personal health data and provide personalised feedback to the user. The Architecture of the PHA can be seen in Figure 2.1. It offers a realistic and well-documented example of how future healthcare AI systems may operate as distributed reasoning networks, where data, clinical knowledge, and behavioural guidance come together in one communicative interface.

Although it is important to understand the back-end structure of such systems, this thesis does not aim to optimise or technically redesign it. The research underpinning this project showed that trust in AI does not stem primarily from clinical accuracy or algorithmic performance, but from the quality of the user’s experience — how information is presented, how reasoning is explained, and how the system’s tone and behaviour are perceived. For that reason, the design phase focuses on the output layer of the PHA: the moment in which the system translates computational reasoning into human-facing communication.

The output interface is where understanding and confidence are formed. It is the place where the AI explains its reasoning, visualises data, acknowledges uncertainty, and provides guidance. Designing for this stage means defining how explanations should look, sound, and feel: how they convey empathy, credibility, and transparency without overwhelming the user. This focus directly links to the design principles established in the literature review and translates them into a concrete design challenge.

By drawing on the PHA as a use case, this thesis positions its design focus within a credible and future-oriented healthcare scenario while keeping attention on the communicative surface — where human experience, design, and artificial reasoning meet. The following section presents the design brief, outlining the goals, purpose, and target group of the framework developed in this thesis.

## 4.3 Design Brief

### Design Goal

Design a framework that guides designers how to design user-centred interfaces to enhance patients' trust and understanding in an agentic AI health assistant, supported by a demonstrating prototype. The framework will guide designers in creating transparent, empathetic, and trustworthy user interactions with AI agents in healthcare. The prototype will serve as a practical validation—demonstrating how communication, interaction, and visualization can foster trust and understanding in patient-facing AI tools.

### Specific goals and design criteria:

- The framework and demonstrator should show how trust can be supported through communication design (tone, structure, responsiveness) rather than claims of computational accuracy.
- The prototype concepts should make outcomes and next steps immediately understandable and actionable in patient-friendly language.
- To be tested: whether users trust and understand the system more when explanations are more abstract (summary-first) versus more specific (detail-rich), and how far specificity can increase before it causes overload.
- To be tested: whether transparency is experienced as more trustworthy when supporting information is curated and contextualised versus when more complete/raw information is disclosed.
- The concepts should support predictable interaction patterns and user participation (clarify, confirm, correct) so trust can be co-created rather than one-way.

### Target Group

1. Designers and design teams – UX and interaction designers (e.g., Accenture Song Design & Digital Products team) who will use the framework to design explainable AI systems.
2. End-users – as test participants who interact with the prototypes; their feedback ensures that the framework reflects real user needs and levels of understanding.

### What It Will Be

A design framework for creating explainable and trustworthy AI interfaces in healthcare, consisting of:

- A **framework** explaining how to design user-centred interfaces to enhance trust and understanding in agentic systems.
- A **designer-facing playbook** describing how, when and why to use the framework.
- A **demonstrator prototype** showing how the framework can be instantiated in an end-to-end interaction.
- A **roadmap** outlining how the approach could evolve from designer-led interfaces toward AI-assisted and governed explanation behaviour over time.

## 4.4 Design Principles

The literature review (Chapter 2.5.3) identified five core design principles for patient-facing explainability: Cognitive Clarity, Multi-modal Accessibility, Interactive Depth Control, Contextual Relevance, and Two-way Transparency. These principles provided a strong theoretical foundation for understanding how patients can be supported in interpreting AI reasoning.

Building on the exploratory research (Chapter 3), these principles were revisited and refined to include the communicative, empathic, and ethical dimensions of trust that emerged from expert and practitioner insights.

The resulting set of refined design principles forms the content foundation of the framework for designing explainable and trustworthy AI in healthcare.

### 4.4.1 Cognitive Clarity

Cognitive Clarity refers to the use of plain, concise, and non-technical communication to make complex reasoning understandable. Literature on XAI highlights clarity as a prerequisite for comprehension (Adeniran et al., 2024; Eiband et al., 2018).

The interviews reinforced this, showing that clarity is not only cognitive but also emotional: both experts stressed that patients feel confident when explanations are short, neutral, and easy to follow. As one expert phrased it, “People don’t need to know every detail; they just need to understand why something is safe and what will happen next.” In practice, this principle implies using familiar language, structuring explanations in small, scannable chunks, and avoiding clinical jargon or over-technical visuals. Clarity becomes the foundation for all other principles, enabling users to focus on meaning rather than interpretation.

### 4.4.2 Multi-modal Accessibility

Multi-modal Accessibility ensures that information is presented in multiple complementary formats—for example, combining text, icons, and visuals. This principle, widely discussed in explainable AI research (Corti et al., 2024), recognises that users have different levels of health literacy and ways of processing information.

The workflow mapping session confirmed that visual and textual redundancy supports comprehension. The practitioner referenced *Thuisarts.nl* as a trusted example, where visual cues and simple language help users feel informed.

For healthcare AI, multi-modal accessibility means designing interfaces that allow users to switch between explanation modes—visual summaries, text narratives, or voice assistance—depending on preference or ability. This not only improves understanding but also fosters inclusion and accessibility.

### 4.4.3 Interactive Depth Control

Interactive Depth Control combines layered explainability with user-driven exploration. The literature describes this as enabling users to “drill down” from general summaries to more detailed reasoning (Eiband et al., 2018; Saeed & Omlin, 2023).

The interviews expanded this principle by showing that depth must also adapt to the user’s emotional state and literacy. As one expert observed, **“Some people want to read everything; others just need reassurance.”**

Designing for depth control therefore means offering a clear top-level explanation while allowing optional exploration of details through expandable sections, tooltips, or follow-up prompts. This flexibility helps maintain comfort and curiosity, aligning transparency with the user’s readiness to engage.

#### 4.4.4 Contextual Relevance

Contextual Relevance connects explanations to the user's personal situation and goals, turning information into insight. Literature emphasises personalisation as a driver of engagement (Adeniran et al., 2024).

The interviews added an empathic dimension to this: relevance is not only about personalised data but also about tone and timing. The practitioner showed that empathy often stems from contextual sensitivity — recognising when reassurance, rather than information density, is most needed. For design, this means tailoring explanations to the user's context (e.g., recent symptoms, demographics, or prior interactions) and phrasing feedback in an emotionally attuned way. Contextual relevance thus bridges factual accuracy and human connection, making AI communication feel considerate and meaningful.

#### 4.4.5 Two-Way Transparency

Two-Way Transparency extends explainability into reciprocal interaction. Existing studies highlight feedback and contestability as mechanisms for trust (Saeed & Omlin, 2023). The empirical findings confirmed this, showing that trust grows when users can ask questions, correct data, or confirm understanding. As one interviewee described, ***“If something seems wrong, people want to say so, that’s when they start trusting it.”***

Designing for two-way transparency involves providing channels for clarification, feedback, and reflection. This could take the form of editable data fields, reassurance prompts (“Does this sound right?”), or short feedback loops after explanations. By turning one-way communication into dialogue, systems become accountable and co-creative, fostering mutual trust between users and AI.

### 4.4. Integration and Refinement

Together, these five refined principles form a user-centred foundation for designing explainable and trustworthy AI interfaces. Each principle operates at a different layer of the interaction:

- Cognitive Clarity ensures understandable language and structure.
- Multi-modal Accessibility diversifies explanation formats.
- Interactive Depth Control adapts explanation detail to user readiness.
- Contextual Relevance personalises and humanises information.
- Two-Way Transparency enables dialogue and accountability.

The exploratory research added a strong relational and emotional dimension to these principles, showing that explainability in healthcare is not merely about information clarity but also about how communication makes people feel—safe, respected, and in control.

These principles form the content layer of the framework presented in the next section (4.5), where their interrelations and roles across the interaction stages of AI communication are structured into a coherent model.

# Develop

## Concept Development and Evaluation

Following the problem framing and synthesis in the previous chapters, Chapter 5 marks the Develop phase of the Double Diamond. The goal of this phase was to translate the derived design principles into testable interface concepts for patient-facing agentic AI explanations. Rather than aiming for one “best” explanation screen, the chapter documents how multiple concept directions were explored to cover different trade-offs in explanation depth, interaction effort, and visual support. This phase therefore connects research insights to design decisions through iterative ideation, prototyping, and preparation for comparative evaluation in the subsequent chapter.

# 5



### AI used in this chapter:

- Google Gemini for storyboard creation
- DALL-E for visual idea generation
- Perplexity for quantitative scale research
- Google Gemini for Quantitative Analysis
- Fireflies.ai for audio transcriptions
- ChatGPT for rewriting of text

## 5.1 Exploring and Ideating a Healthcare assistant

Section 5.1 outlines the exploratory steps that shaped the initial solution space. First, existing patient-facing health applications were analysed to identify interface conventions for turning personal data into understandable insights. Second, earlier findings were translated into practical design prompts (principles and “How Might We” questions) to guide ideation. These inputs informed early sketches and concept directions, ensuring that later prototypes were grounded in established interaction patterns while still testing distinct approaches to explanation depth, visual support, and user control.

### 5.1.1 Current Patient-Facing Healthcare apps

To understand how users already interact with health information in everyday digital products, I reviewed four widely used patient-facing healthcare applications: WHOOP, Strava, Sleep Cycle, and Apple Health (see Figure 5.1). These applications are not AI assistants and they do not provide medical recommendations. However, they are relevant because they show how personal health-related data can be made understandable through summaries, structure, and controlled access to detail.

Across these apps, users typically encounter an overview-first structure: a small set of key metrics or short summaries is shown first, followed by more detailed breakdowns when users scroll, tap, or open a secondary view. For example, Strava’s Athlete Intelligence component provides a brief interpretation of an activity and then offers an explicit option to expand the content (“Vertel me meer”) to read a longer explanation. This does not function as medical guidance, but it demonstrates an interaction pattern in which interpretation can be offered in a short form first, with optional depth available through a clear trigger.

A second recurring pattern is the use of simple visual cues to reduce the amount of text needed for interpretation. In Sleep Cycle, icons communicate the meaning of events such as bedtime, waking up, and snoring duration (e.g., moon, alarm, “zzz”; see Figure 5.1). WHOOP similarly presents high-level indicators prominently (e.g., recovery or sleep state) while placing more granular metrics in supporting sections that users can explore further (see Figure 5.1). Apple Health complements charts and values with brief comparative framing (e.g., a message indicating the user walked less than usual), which helps users interpret what a number means in context (see Figure 5.1).

Design implications. Although these apps are not agentic assistants, they provide useful interface conventions for explainable patient-facing AI: (1) present a clear overview before detailed evidence; (2) make depth optional through explicit “read more / drill down” interaction; and (3) support interpretation using lightweight visual cues and short contextual framing. These conventions informed how the concepts in this thesis stage information and expose supporting evidence without overwhelming users.

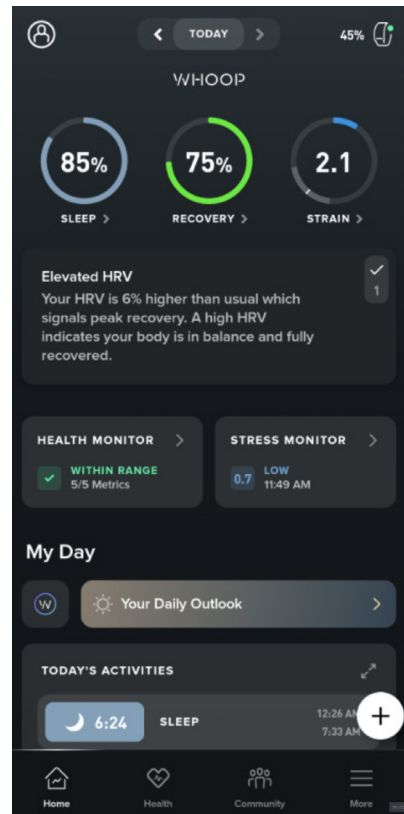


Figure 5.1: Screenshots of different UI's. (Strava, WHOOP, SleepCycle, Apple Health)

### 5.1.2 Scenario and User Journey

All concepts were designed around a single, fixed scenario to ensure that later differences in user responses can be attributed to explanation strategy rather than to changes in medical content, severity, or user goal. The scenario and the corresponding interaction moment are visualised in Figure 5.2.

The scenario depicts a user consulting My Health Agent about a recognisable, non-acute concern: a mild fever increasing from 38.3°C to 38.7°C over 48 hours, accompanied by fatigue, headache, and a slight cough. The system has access to relevant context data: Vaccination status, absence of chronic conditions, recent wearable signals such as elevated resting heart rate and reduced sleep. This scenario was selected because it creates a realistic need for reassurance and guidance, while still supporting explanation: multiple signals could plausibly influence a recommendation, making it suitable for testing how reasoning is communicated.

To evaluate the explainability strategy of three different interface concepts (A/B/C) under controlled conditions, the scenario was kept consistent and intentionally relatable. A fixed scenario reduces confounding influences such as differences in perceived severity, prior knowledge, or personal relevance, which could otherwise dominate participants' trust and understanding ratings. By holding the medical context constant, any differences observed during evaluation can be attributed more plausibly to the design of the explanation interface—including information hierarchy, visualisation choices, interaction affordances, and how supporting evidence is surfaced, rather than to the content itself.

Design implications. Fixing the scenario supports internal validity for concept comparison: it enables the study to test whether participants respond differently to the explainability design of Concepts 1–3, instead of reacting to a different health situation. This directly aligns the evaluation with the research goal of comparing how interface level explainability influences perceived trust and understanding.



Figure 5.2: Storyboard of the test scenario (Generated with Nano Banana)

### 5.1.3 Ideating

Building on the design principles established in Chapter 4, ideation focused on developing three distinct interface concepts that could be compared in the A/B/C study. The goal was to explore different ways of communicating AI health guidance—particularly how much explanatory support is needed, and what form that support should take—without changing the underlying scenario. To explore visual direction efficiently, generative AI imagery was used as inspiration material for layout and component ideas. Selected elements were then assembled into three low-fidelity digital concept directions (“spuugmodellern”), shown in Figure 5.3.

The three directions represent different hypotheses about what patients need in the moment of receiving AI health guidance:

- Concept 1 intentionally provides only the outcome, why it is safe, and immediate next steps. It tests whether a concise, action-oriented interface can still feel clear and trustworthy in a mild, relatable scenario.
- Concept 2 keeps the outcome central but adds a compact explanation layer through lightweight visuals, colour and short supporting cues. It tests whether providing limited, easy-to-scan context improves perceived understanding and trust without increasing cognitive load.
- Concept 3 keeps the outcome central, but provides in-depth inspectability, offering explicit pathways into contributing factors and underlying evidence with a knowledge graph, medical timeline, and raw-data. It tests whether deeper transparency increases trust and understanding, and for whom it becomes beneficial versus overwhelming.

At this stage, these concepts are not final prototypes; they serve as early directional models that define the intended differences in explainability approach. Section 5.2 refines these directions into coherent, comparable prototypes by standardising shared interface elements and interaction structure, while preserving the conceptual differences shown in Figure 5.3

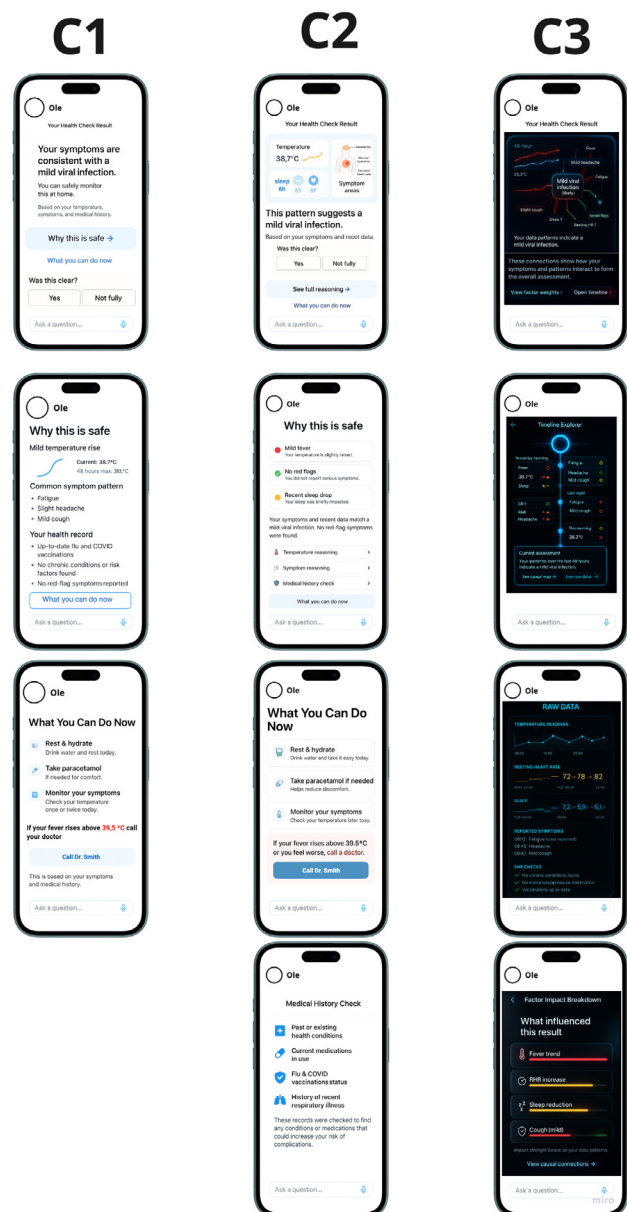


Figure 5.3: LoFi prototypes generated with DALL-E and assembled in Miro

## 5.2 From Idea to Final Concepts

The initial digital concept directions were translated into interactive prototypes and refined to function as credible, comparable test artefacts. Refinement focused on stabilising baseline interface quality—clarity, familiarity, and perceived trustworthiness—so that the A/B/C study evaluates differences in explainability design (depth, interaction, evidence visibility) rather than differences in polish or visual style.

### 5.2.1 Iterative Design Decisions in Figma

Refinement prioritised consistency, following Shneiderman’s (2016) rule to strive for consistency. Navigation patterns, button placement, terminology, icon usage, and layout structure were aligned across the three concepts wherever the user task is equivalent. This alignment reduces “learning effects” between conditions and keeps participants’ attention on the conceptual differences in explainability rather than on inconsistent interaction behaviour.

### 5.2.2 Colour Choice and Perceived Trust

A shared colour system was applied across all concepts. Blue served as the primary colour to support a calm, trustworthy tone (JCT CoatingsTech, 2007) and to avoid urgency signalling that could bias perceived risk (Figure 5.4). Using the same palette across conditions prevents colour from becoming an unintended factor in trust ratings.

### 5.2.3 Typography and Familiarity

All prototypes use the typeface Inter for readability and familiarity in screen-based interfaces. A shared typographic hierarchy supports scanning and reduces cognitive effort, while also keeping perceived professionalism consistent across the three concepts so that differences in evaluation relate to explanation design rather than text presentation.

### 5.2.4 Consistency Across Concepts

All visual and interaction design decisions (colour palette, typography, spacing, and basic call-to-actions) were applied consistently across the three concepts. This consistency was essential to ensure that differences observed in later testing could be attributed to variations in explanation style and interaction depth, rather than to differences in visual design quality.

Through this iterative refinement process, the prototypes evolved into coherent and realistic representations of patient-facing AI interfaces. They were sufficiently refined to be used as test artefacts in the subsequent synthetic user testing and user evaluation phases described in the following sections.

Red	Energy, Passion, Power, Excitement
Orange	Happy, Confident, Creative, Adventurous
Yellow	Wisdom, Playful, Satisfying, Optimistic, Cheerful
Green	Health, Regeneration, Contentment, Harmony, Nature, Relaxation
Blue	Honesty, Integrity, Trustworthy, Loyalty
Violet	Regal, Mystic, Beauty, Inspiration, Luxury, Sophistication
Brown	Reliable, Genuine, Sad, Wishful
Black	Authority, Power
White	Purity, Innocence, Light

Figure 5.4: Color Psychology (JCT CoatingsTech, 2007)

## 5.3 Synthetic User Testing

Before conducting user testing with real participants, the first iterations of the three concepts were evaluated through a synthetic user testing phase. The aim of this phase was to identify which user interface elements were experienced as effective or problematic, and to use these insights to improve the designs before entering the empirical testing phase.

the interface solely based on what was visible in the screenshots, assuming a first-time encounter with the system. This setup ensured that feedback reflected immediate reactions to layout, visual structure, wording, and interaction cues, rather than learned behaviour or comparison between concepts. The full persona definitions, evaluation protocol, and detailed outputs are included in the appendix B.

### 5.3.1 Rationale and Approach

Initially, the intention was to conduct this exploratory evaluation using synthetic persona software from Het Volkspportret. However, this option was not pursued further, as the costs associated with using the platform were disproportionate to the scope and budget of this thesis project. Neither me nor the project stakeholder, Accenture Song, considered this investment appropriate for an exploratory design phase.

As an alternative, self-made synthetic personas were used to evaluate the concepts in a structured and controlled way. A predefined set of personas was created to represent variation across age, digital familiarity, health literacy, and general attitudes toward healthcare and technology. These personas were used strictly as fixed viewpoints to simulate first-time interactions with the interfaces.

### 5.3.2 Evaluation Setup

The evaluation was conducted using Google Gemini. Screenshots of the Figma prototypes were uploaded into Gemini, and a separate Gem was created for each persona. Each Gem started with a blank context, ensuring that no information, judgments, or interpretations could be transferred between personas or between concept evaluations.

Each synthetic persona was exposed to one concept only, following a between-subjects setup. The personas evaluated

### 5.3.3 Key UI-Level Findings

The synthetic evaluations revealed several recurring patterns related specifically to interface elements and presentation choices. These insights were used directly to guide the next design iteration.

- **Clear main messages and action buttons were consistently valued:** Interfaces that presented a clear outcome at the top of the screen, followed by visible next steps, were generally experienced as strong and reassuring. Large, clearly labeled buttons—such as options to view next steps or contact a doctor—were repeatedly described as helpful. These elements gave the interface a sense of structure and direction, reducing ambiguity about what the user should do next.
- **Step-by-step layouts worked better than dense screens:** Screens that separated information into distinct sections or steps were preferred over screens that presented many elements at once. When too much information was visible simultaneously, participants described the interface as busy or overwhelming. In contrast, interfaces that revealed information gradually, or grouped related content into clear blocks, were experienced as easier to follow.
- **Visual indicators needed clear meaning:** Icons, colours, and small data indicators were only effective when their meaning was immediately clear. Several personas commented that small icons or numerical indicators without labels were confusing, even when they were visually consistent. This indicated that visual elements should either be self-explanatory or supported by short textual clarification.

- **Showing irrelevant data reduced confidence in the interface:** When interfaces displayed data that did not clearly relate to the current health situation, this was perceived negatively. For example, historical or unrelated health information included without explanation raised questions about why it was shown and whether it was relevant. This insight led to a stricter selection of which data points were surfaced in later iterations.
- **Complex visualisations were not always helpful:** Highly visual explanations that attempted to show many relationships at once were often experienced as difficult to interpret. While the intention of such visualisations was to increase transparency, they sometimes had the opposite effect, drawing attention away from the main message. Simpler visual structures with a clear focal point were generally perceived as more effective.

### 5.3.4 Outcome of the Synthetic Testing Phase

The synthetic user testing phase functioned as a design filter rather than an evaluation of effectiveness. It helped identify which interface elements should be retained, simplified, clarified, or removed before testing with real users.

Based on these findings, the concepts were iterated to:

- Simplify visual structure,
- Reduce the amount of information shown at once,
- Clarify the meaning of icons and indicators,
- Ensure that all visible data directly supported the main message.

This resulted in more focused and refined prototypes, which were then suitable for validation through user testing with real participants. The final versions of the concepts are presented in the following section.

## 5.4 Final Concepts

This section presents the three final interface concepts and explicitly links their design features to the literature-derived design principles (cognitive clarity, multi-modal accessibility, interactive depth control, contextual relevance, and two-way transparency) and the exploratory design criteria (reassurance and emotional safety, explicit next steps, and transparency).

Across all concepts, predictable structure and stable interaction patterns are maintained through the same typography, colour palette, spacing rules, and consistent placement of primary actions. In addition, all concepts incorporate two-way transparency in the specific form of a feedback affordance that asks users whether the explanation is clear (“Was this clear?”), allowing them to select Yes/No. This mechanism does not enable open dialogue, but it operationalises a minimal two-way loop by capturing whether the system’s communication succeeded from the user’s perspective.

### 5.4.1 Concept 1 Simple / low-interaction / text-dominant

As shown in figure 5.5, Concept 1 was intentionally designed as the simplest possible interface. It prioritises cognitive clarity through strict reduction: a single outcome message and immediately visible next steps in plain language. The design avoids explanatory widgets and evidence views on purpose, so it can function as a baseline condition.

This concept operationalises reassurance and emotional safety through calm framing and low information density, while making “what happens next” explicit through clearly stated actions and monitoring advice. Contextual relevance is maintained by grounding the message in the user’s reported symptoms and scenario context.

Concept 1 therefore tests whether trust and understanding can be supported through clarity, reassurance, and actionability alone, without additional transparency features.

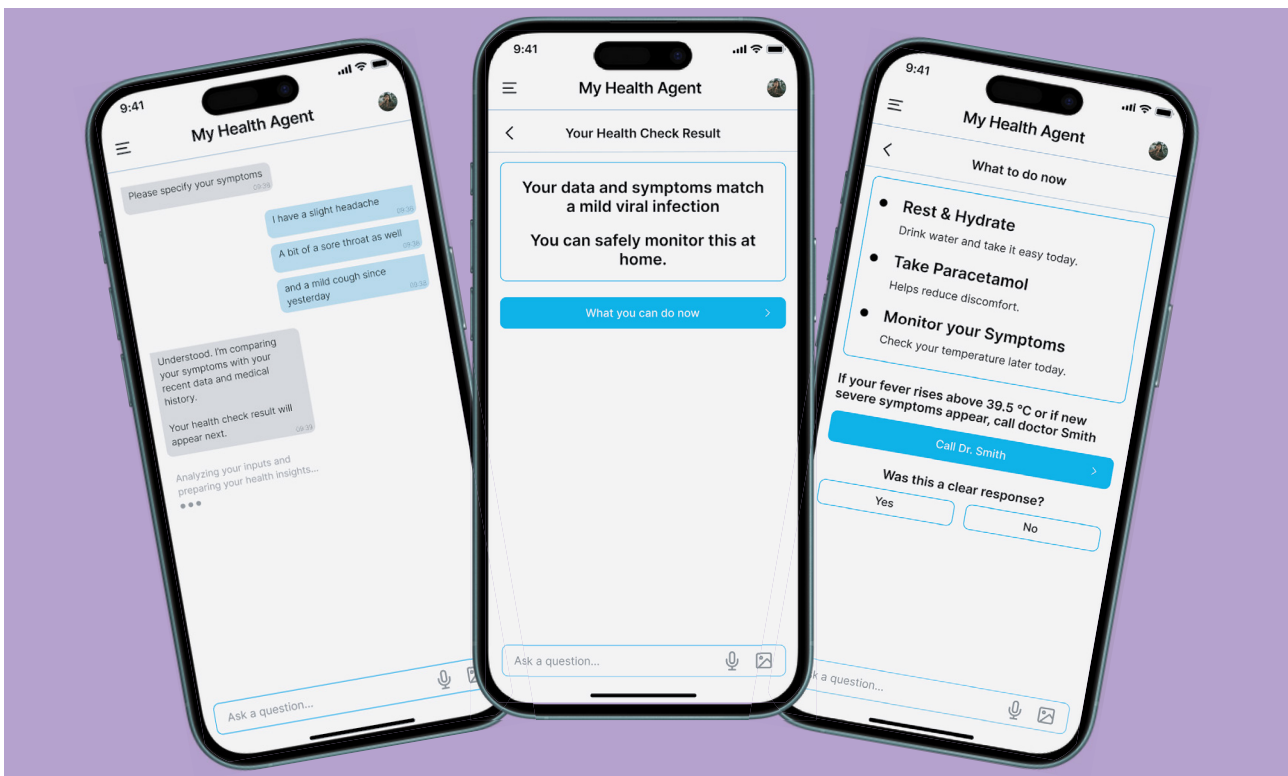


Figure 5.5: Concept 1 Figma mockups

### 5.4.2 Concept 2 Simple / mid-interaction / strong visual support

Concept 2, shown in figure 5.6, retains the same outcome-first structure as Concept 1, but introduces a lightweight explanation layer through compact multi-modal cues. The intent is to strengthen understanding without turning the interaction into a detailed investigation. Cognitive clarity remains central: outcome and next steps stay dominant, and explanatory content is designed to be scanned quickly rather than read as a long narrative.

The central component is a swipeable widget that allows users to move through three distinct views: (1) an illustration representing the identified condition, (2) a visualisation of temperature changes over time, and (3) a set of key contributing data points. In the presented scenario, these contributing data points include heart rate, sleep duration, and energy level. Each element is displayed in a compact and visually distinct manner.

This supports multi-modal accessibility by combining recognition (illustration), temporal context (temperature trend), and selective quantitative signals (key data points) in a bounded interaction.

Concept 2 operationalises transparency as helpful visibility: it surfaces only a small set of cues that support the user's interpretation of the outcome, without exposing raw records or exhaustive detail. It also includes limited interactive depth control through the swipe interaction, enabling modest user choice in which explanatory view is shown while keeping interaction effort low.

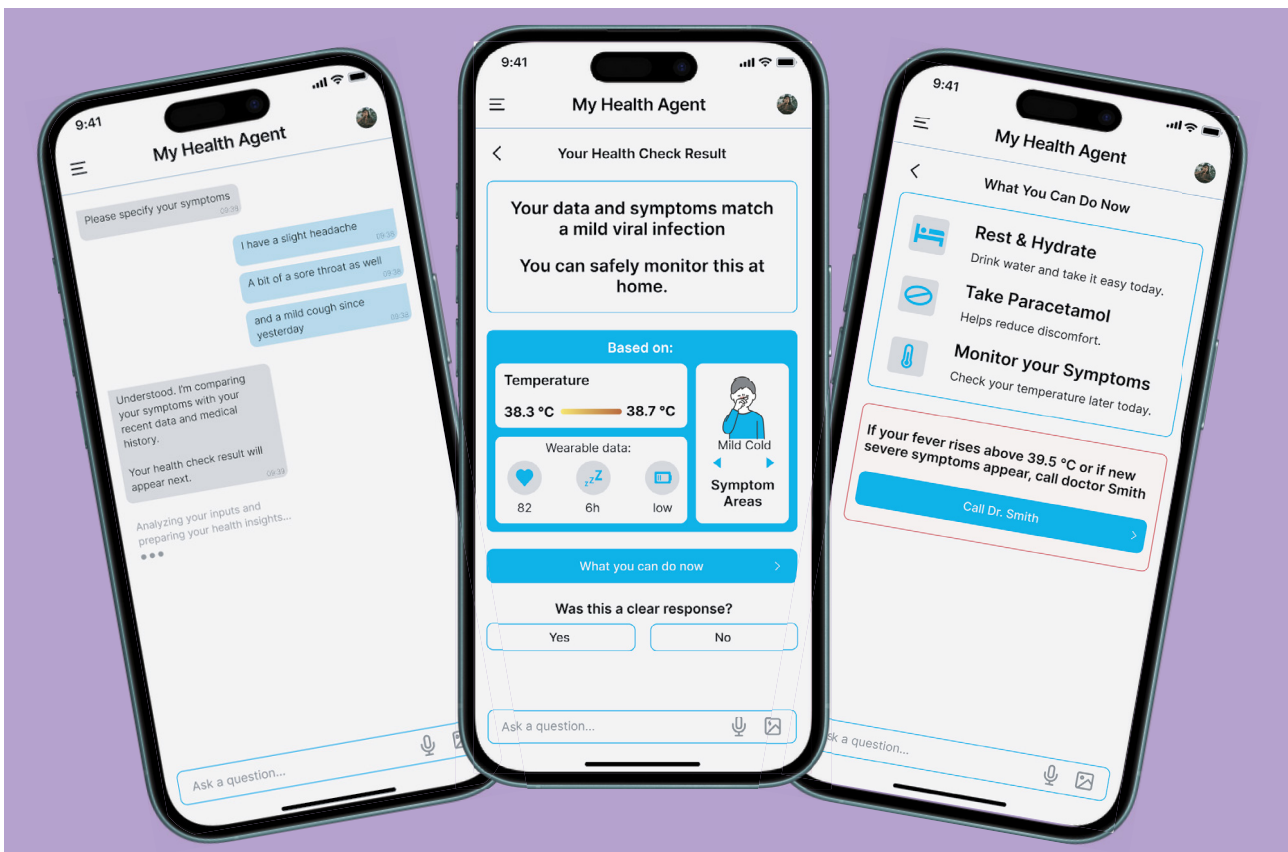


Figure 5.6: Concept 2 Figma Mockups

### 5.4.3 Concept 3 In-depth / high-interaction / detail-rich / strong visual support

Concept 3, illustrated in figure 5.7, was intentionally designed as a very detailed concept to test the upper bound of transparency and explanation depth. It functions as a “maximum detail” condition: the design explores whether deep inspectability is experienced as reassuring and trustworthy, or instead perceived as excessive and cognitively overwhelming in a patient-facing context. This concept therefore pushes strongly on the abstraction–specificity tension identified earlier, by making specificity and evidence access highly available.

Full disclosure is operationalised through a knowledge-graph-style data hub that serves as the central entry point for deeper evidence. From this hub, users can access their medical timeline (events and measurements over time) and contextual background such as vaccination history, enabling traceability beyond the immediate symptom report. Interactive depth control is therefore not

limited to a single widget, but implemented as a structured navigation approach to deeper layers of information.

To make reasoning perceivable already on the first screen, Concept 3 includes symptom impact bars. These bars visualise how each reported symptom influences the system’s calculation and show a reference range indicating what is typical for a mild viral infection. This provides an immediate explanation cue that links the outcome to specific inputs, while also signalling how “typical” the current pattern appears compared to the reference range.

Despite the depth, the concept maintains explicit next steps and keeps actionability reachable, ensuring that transparency does not displace guidance. However, unlike Concept 2, this concept does not prioritise minimising information; instead it tests whether users prefer deep, fully inspectable evidence access—or experience it as “too much”—compared to the simpler designs in Concepts 1 and 2.

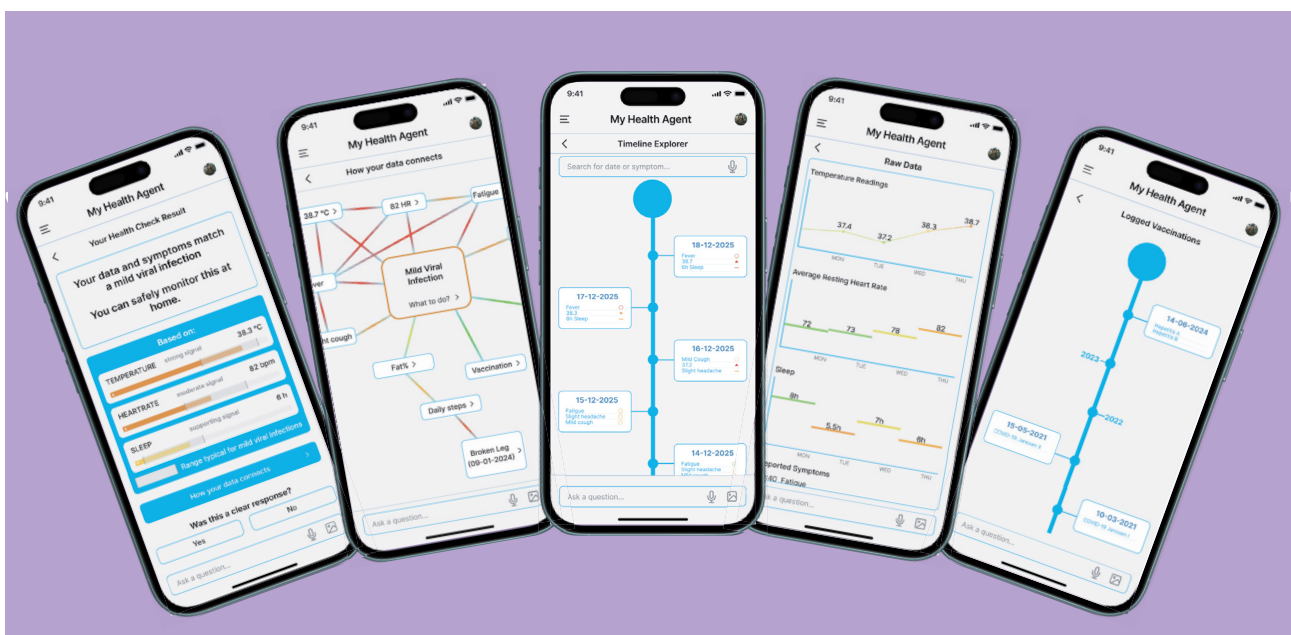
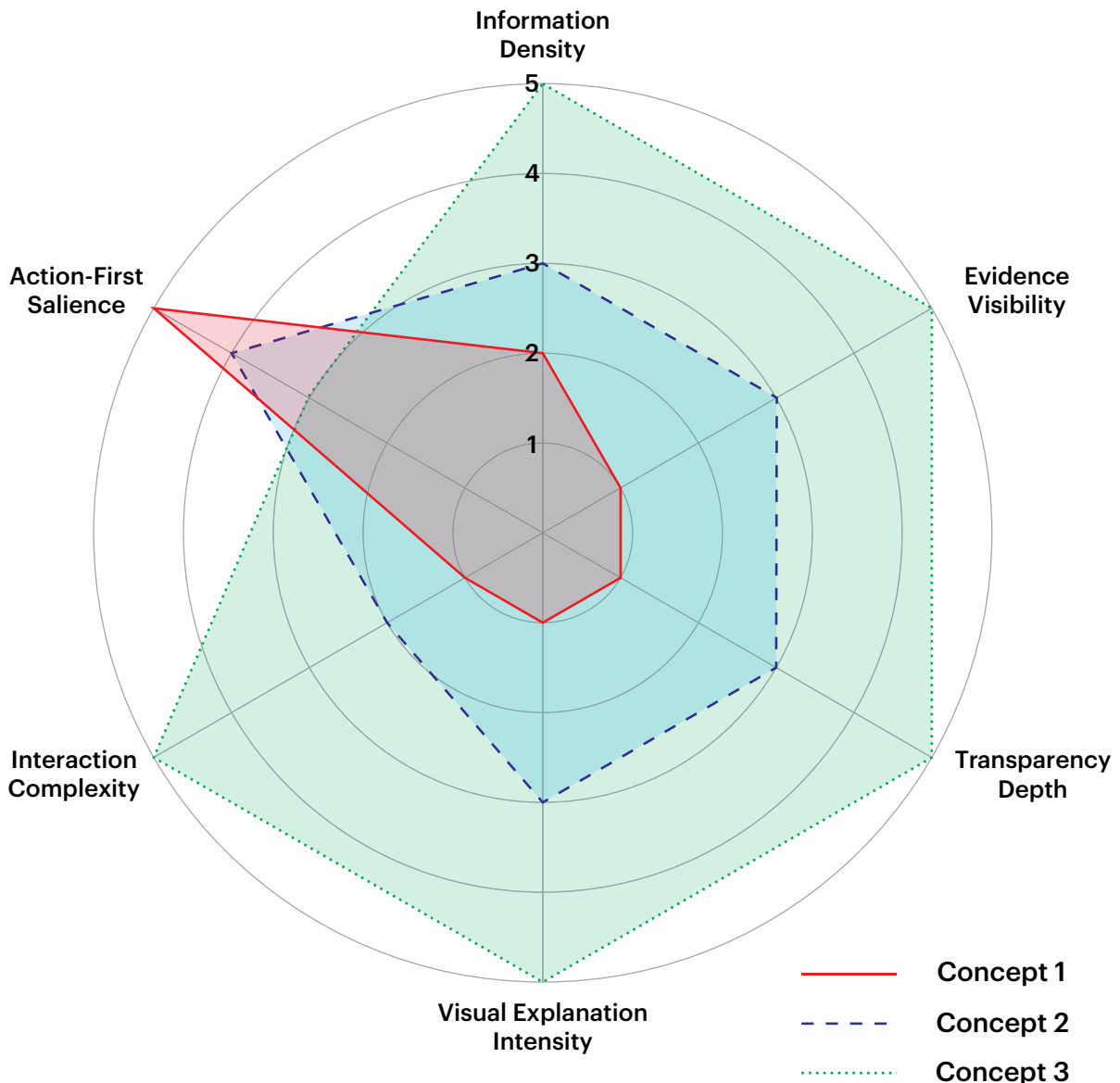


Figure 5.7: Concept 3 Figma Mockups

## 5.5 Rationale for Concept Diversity

The three final concepts were intentionally designed to differ along clearly defined interface dimensions. These differences were not introduced arbitrarily, nor were the concepts intended to represent incremental improvements of one another. Instead, they were constructed as distinct design variations that allow for systematic comparison of how different approaches to explanation, interaction, and visualisation influence user experience in patient-facing AI systems. Figure 5.8 showcases how the concepts differ from each other.

The rationale for this concept diversity is rooted in the central research aim of this thesis: to understand how different forms of explaining AI reasoning affect users with varying characteristics, needs, and levels of familiarity with digital health technologies. By deliberately varying interface complexity, interaction depth, and visual support across the concepts, the study creates controlled conditions under which these factors can be examined in isolation and in combination.



**Figure 5.8:** Spider diagram visualizing the intended experimental contrast between the 3 concepts

### 5.5.1 Variation in Interaction Depth and Interface Complexity

A primary axis of differentiation between the concepts is interaction depth. Concept 1 offers a minimal interaction model, limiting user engagement to reading a concise result and following a small number of clearly defined action steps. Concept 2 introduces moderate interaction through swipeable visual elements and supporting widgets, while maintaining a relatively shallow navigation structure. Concept 3, by contrast, provides extensive interaction possibilities, allowing users to explore relationships between data points, navigate timelines, and access multiple layers of underlying information.

This progression from low to high interaction depth reflects the assumption that users differ in how much information and control they desire when engaging with health-related AI systems. Some users may prefer brief, outcome-focused communication, while others may wish to explore how conclusions were formed and how different data sources relate to one another. Designing concepts across this spectrum makes it possible to examine how increasing complexity and interactivity shape user engagement and perception during the evaluation phase.

### 5.5.2 Variation in Use of Visual Support

A second dimension along which the concepts differ is the use of visual support. Concept 1 relies almost entirely on textual explanation, using layout and typographic hierarchy rather than imagery or data visualisation to communicate meaning. Concept 2 introduces visual elements that complement the text, such as icons, condition illustrations, and simplified data indicators. Concept 3 extends this approach further by incorporating multiple forms of visualisation, including signal bars, network-style representations of data relationships, timelines, and charts.

This deliberate variation allows the study to explore the role of visualisation as a

communicative aid. Visual elements can support comprehension by making abstract information more concrete, but they can also introduce additional interpretive effort when they become too dense or unfamiliar. By comparing concepts with differing levels of visual support, the study can examine how visuals function as explanatory tools for users with different backgrounds and preferences.

### 5.5.3 Variation in Transparency and Data Exposure

The concepts also differ in how explicitly they expose underlying data and reasoning processes. Concept 1 provides a high-level outcome and guidance without revealing much detail about how the conclusion was reached. Concept 2 partially exposes this reasoning by highlighting selected data points that contributed to the result. Concept 3 makes the reasoning process highly visible by presenting data relationships, historical context, and raw measurements across multiple views.

This variation reflects different design interpretations of transparency in AI systems. Rather than assuming that more transparency is always beneficial, the concepts explore a range of disclosure strategies, from summary-level communication to full data exploration. This enables the study to investigate how different degrees of transparency are experienced by users and how they interact with personal characteristics such as familiarity with technology or comfort with data-driven systems.

### 5.5.4 Alignment with Study Design

Together, these three dimensions—interaction depth, visual support, and transparency—form the foundation of the experimental design described in the next chapter. The concepts were constructed to differ clearly on these dimensions while remaining consistent in scenario, content, and medical context. This ensures that any observed differences during

evaluation can be attributed to interface design choices rather than to variations in use case or informational content.

By establishing concept diversity in a structured and intentional manner, the study creates a coherent basis for comparative evaluation. The following chapter outlines how these concepts were tested with participants and how the study design was structured to examine the effects of these design variations systematically.

## 5.6 Testing Approach and Study Design

This sub-chapter describes the empirical evaluation of the three final interface concepts. It outlines the study design, participant sample, procedure, measurement instruments, and hypotheses. The study was designed to systematically compare different approaches to explanation, interaction, and visualisation in a patient-facing AI interface under controlled conditions.

### 5.6.1 Study Design

To enable a direct comparison between the three interface concepts, the evaluation employed a within-subject (A/B/C) study design. Each participant interacted with all three concepts and completed the same set of questions after each concept. This approach was chosen to reduce variability caused by individual differences, as each participant served as their own reference point across conditions.

The study was conducted in person using a smartphone and the Figma mobile application, allowing participants to interact with the prototypes in a realistic, touch-based context. All concepts were evaluated within the same health scenario and differed only in interface design, interaction depth, and visual presentation.

### 5.6.2 Participants

A total of 11 participants took part in the study (target = 12). Participants varied in age, gender, educational level, place of residence, and self-reported medical or health literacy. Recruitment aimed to achieve heterogeneity across these characteristics to reflect a broad range of potential users of patient-facing health applications.

Given the exploratory and design-oriented nature of this thesis, the sample size was considered sufficient to identify meaningful patterns and differences between interface concepts, particularly within a within-subject design.

### 5.6.3 Procedure

Each testing session followed a fixed structure consisting of three phases: background questions, concept evaluations, and comparative reflection.

#### **Background phase.**

Before interacting with the prototypes, participants completed the first few questions of the questionnaire capturing demographic information and individual characteristics. Health literacy was assessed using brief subjective screening items based on the Brief Health Literacy Screen, which has been shown to reliably capture functional health literacy in applied settings (Wallston et al., 2014). Participants also answered questions regarding their familiarity and comfort with AI-based systems.

#### **Concept evaluation phase.**

Participants then evaluated the three concepts sequentially using the Figma app on the researcher's phone. After interacting with Concept A, participants completed a set of questions regarding concept A. The same questions were then completed after Concept B and Concept C, using identical wording and scale formats. This ensured consistency across conditions and enabled structured within-participant comparison.

Participants were instructed to explore each concept as if they were receiving advice from a personal health assistant in a real situation. No time limits were imposed, allowing participants to interact with the interfaces at their own pace.

#### ***Comparative and qualitative phase.***

After completing all three concept blocks, participants answered two comparative questions in which they chose their preferred version and the version they considered most trustworthy. This was followed by a semi-structured conversation in which participants were asked to elaborate on their choices, reflect on specific interface elements, and discuss which aspects they would keep, remove, or combine. These qualitative insights were used to complement and contextualise the quantitative results.

The full survey structure, question wording, and scales are provided in appendix C.

#### **5.6.4 Measures**

All outcome measures were collected immediately after interaction with each concept using Likert-type scales. The questionnaire combined validated measurement approaches with short, repeated item sets suitable for within-subject comparison.

**Perceived trust** was measured using the Short Trust in Automation Scale (S-TIAS), a validated short-form instrument developed for repeated measurement in human-AI interaction contexts (McGrath et al., 2025).

**Perceived ease of use** was measured using items adapted from the Technology Acceptance Model (TAM), which conceptualises ease of use as a key determinant of user acceptance of interactive systems (Davis, 1989).

**Perceived understanding** was measured using items adapted from prior research on subjective understanding of complex health-related information, drawing on the work of Roberts et al. (2018). These items assess how well participants felt they understood the information presented and whether they perceived their understanding as sufficient to act upon it.

**Perceived clarity** was measured using items grounded in the concept of visual and structural clarity, informed by the work of Schrepp et al. (2020). In this context, clarity refers to the perceived order, structure, and comprehensibility of an interface, rather than objective correctness of the information.

### 5.6.5 Hypotheses

The study was structured around four hypotheses that link differences between the interface concepts to expected differences in participant responses, while accounting for individual characteristics.

#### ***H1: Age as a moderator of interface complexity and trust.***

Older adults are expected to report higher perceived trust in the simple, low-interaction concept (Concept 1) than in the in-depth, high-interaction concept (Concept 3), whereas younger adults are expected to report equal or higher trust in Concept 3 compared to Concept 1

#### ***H2. Interface complexity and cognitive demand.***

Across participants, the in-depth, high-interaction concept (Concept 3) is expected to be experienced as more cognitively demanding and to result in lower perceived clarity and lower perceived understanding than the simple, low-interaction concept (Concept 1)

#### ***H3. Visual support and health literacy.***

The concept with strong visual support (Concept 2) is expected to lead to higher perceived clarity and higher perceived understanding than the text-dominant concept (Concept 1). This effect is expected to be stronger for participants with lower health literacy than for participants with higher health literacy.

#### ***H4. AI knowledge or comfort as a moderator of trust.***

Participants with lower AI knowledge or comfort are expected to report lower perceived trust in the in-depth, high-interaction concept (Concept 3) compared to the simple concept (Concept 1), whereas participants with higher AI knowledge or comfort are expected to report equal or higher trust in Concept 3.

## 5.7 Quantitative Results and Comparative Analysis

This section presents the findings from the quantitative user testing (N=11). The data was analyzed to evaluate the three distinct concepts and to test the formulated hypotheses regarding trust, clarity, and behavioral intention.

### 5.7.1 Data Preparation and Sample Characteristics

Prior to analysis, the raw data set was screened for quality and completeness. Since I sat next to the participants during the tests no participants had to be excluded, resulting in a final valid sample of N=11.

Composite scores were calculated for the key constructs. Health Literacy was computed by averaging the three items from the Brief Health Literacy Screen (BHLS), where a higher score indicates a lower level of literacy (need for help). AI Comfort was calculated as the mean of the two attitude items (M = 4.95, SD = 1.34). The internal reliability of the construct scales (Clarity, Trust, Ease of Use, Intention) was considered acceptable for an exploratory sample.

### Demographics

The sample consisted of a balanced gender split (6 females, 5 males) with a median age of 31 years (Range: 23–79). Geographically, the participants were split between urban residents (Randstad: Amsterdam/Utrecht, n=6) and rural residents (Countryside, n=5), allowing for an exploratory analysis of regional differences.

### 5.7.2 Descriptive Statistics

Preliminary observation of the means indicates a divergence between performance and preference. While Concept A scored highest on Clarity (M=4.27) and Ease of Use (M=4.55), it received low scores for Intention to Use (M=3.36). Conversely, Concept B emerged as the most balanced option, achieving the highest Trust (M=3.70) and Intention (M=3.82) scores. Concept C scored notably lower on usability metrics (M=2.73 for both Clarity and Ease).

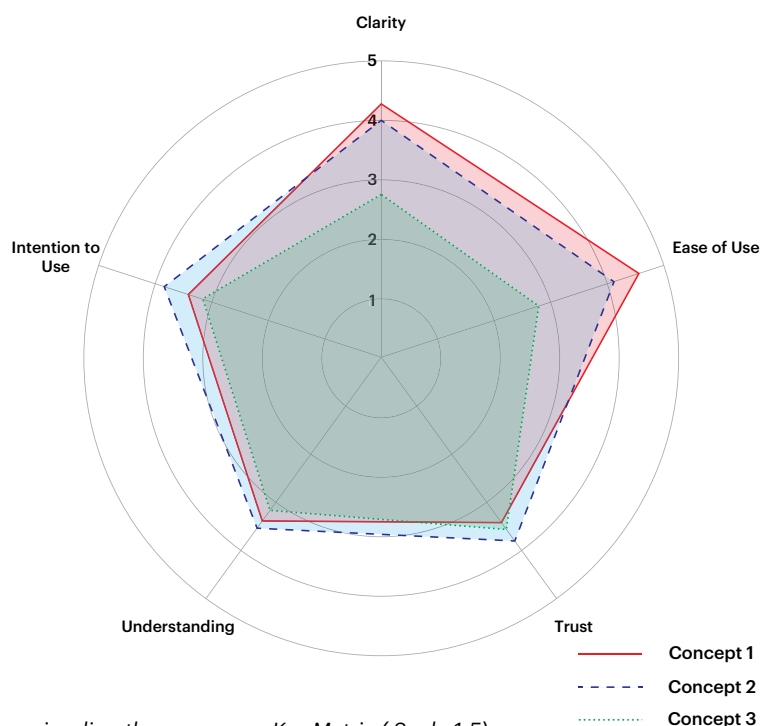


Figure 5.9: Spider diagram visualizing the means per Key Metric (Scale 1-5)

### 5.7.3 Hypothesis Testing

To test the specific effects of interface design and user characteristics, a series of Paired Samples T-Tests (for within-subject effects) and Independent Samples T-Tests (for between-subject moderators) were conducted.

#### **H1: The Moderating Effect of Age on Trust**

Hypothesis 1 posited that age would moderate the relationship between complexity and trust, specifically that older adults would prefer the simple Concept A, while younger adults would prefer the complex Concept C.

A median split was performed to categorize participants into “Younger” (<31 years) and “Older” (>31 years) groups. Independent t-tests on the Trust Differential score (Trust\_C - Trust\_A) revealed no significant difference between the age groups ( $p > .05$ ). Contrary to the hypothesis, older adults did not penalize the complex interface; in fact, the older group reported slightly higher trust in Concept C ( $M=3.67$ ) than Concept A ( $M=3.61$ ). Thus, H1 is rejected.

#### **H2: Complexity, Clarity, and Understanding**

Hypothesis 2 predicted that the high-interaction Concept C would be experienced as more cognitively demanding, leading to lower clarity and understanding compared to Concept A. A Paired samples t-tests confirmed a main effect for Clarity. Participants rated Concept A clearer ( $M=4.27$ ) than Concept C ( $M=2.73$ ),  $t(10) = 4.22$ ,  $p = .002$ . The large mean difference (-1.54) indicates a substantial penalty for the complexity of Concept C. However, regarding Perceived Understanding, there was no significant difference between Concept A ( $M=3.36$ ) and Concept C ( $M=3.27$ ),  $t(10) = 0.20$ ,  $p = .848$ . This suggests that while users found C “messy” or unclear, they were still able to extract the core medical message. Consequently, H2 is partially supported.

#### **H3: Visual Support and Health Literacy**

Hypothesis 3 proposed that Concept B (Visual) would outperform Concept A (Text) in clarity, particularly for users with lower health literacy.

The sample was split into High and Low Health Literacy groups based on the BHLS scores. An independent samples t-test on the Visual Benefit Score (Clarity\_B - Clarity\_A) revealed a marginally significant interaction effect in the opposite direction of the hypothesis ( $p = .06$ ).

- **High Literacy Group:** Experienced a clarity benefit from visuals (+0.29).
- **Low Literacy Group:** Experienced a clarity penalty from visuals (-0.95).

Rather than aiding vulnerable users, the icons and visual layouts appeared to add cognitive noise for those with lower literacy, who rated the simple text (Concept A) as clearer. Therefore, H3 is rejected.

#### **H4: AI Comfort and Trust**

Hypothesis 4 suggested that lower AI comfort would lead to lower trust in the complex Concept C. An independent samples t-test comparing Low vs. High AI Comfort groups showed no significant difference in their trust preference ( $p = .988$ ). Both groups displayed a slight preference for the trustworthiness of the complex concept, suggesting that the “expert aesthetic” of Concept C signals competence universally, regardless of the user’s technical confidence. H4 is rejected.

### 5.7.4 Drivers of Behavioral Intention (Repeated Measures Correlation)

To identify which factors most strongly predict a user's Intention to Use the assistant, a statistical approach suited for small-sample, within-subject designs was employed. Due to the limited sample size (N=11), a standard linear regression on aggregated means would lack sufficient statistical power. Therefore, a Repeated Measures Correlation (rmcorr) was conducted to examine the relationship between the predictors and intention across the three concept iterations (df=21). The results can be seen in the graph below

The analysis revealed that Trust was the strongest statistical predictor of adoption ( $r_{rm} = .71, p < .001$ ). This indicates an observed positive correlation: as a user's trust in a specific prototype increased, their intention to use it rose disproportionately.

Clarity also demonstrated a significant positive correlation with intention ( $r_{rm} = .45, p = .032$ ), suggesting that clearer interfaces effectively boost user willingness. Perceived Ease of Use showed only a moderate and marginally significant correlation ( $r_{rm} = .36, p = .095$ ).

Additionally, Perceived Understanding was not a significant predictor of intention ( $r_{rm} = .33, p = .121$ ).

To assess the impact of stable user traits, Spearman's Rank Correlations ( $r_s$ ) were conducted on the aggregated scores. Neither AI Comfort ( $r_s = -.32, p = .33$ ) nor Health Literacy ( $r_s = -.02, p = .96$ ) showed a significant relationship with Intention to Use.

These findings suggest a shift in the behavioral model: Trust is the primary gatekeeper for adoption, significantly outweighing friction (Ease of Use) or user personality (AI Comfort). Users appear willing to adopt a tool if they trust its output, even if the interaction is not perfectly frictionless, but they will not adopt a tool solely because it is easy to use.

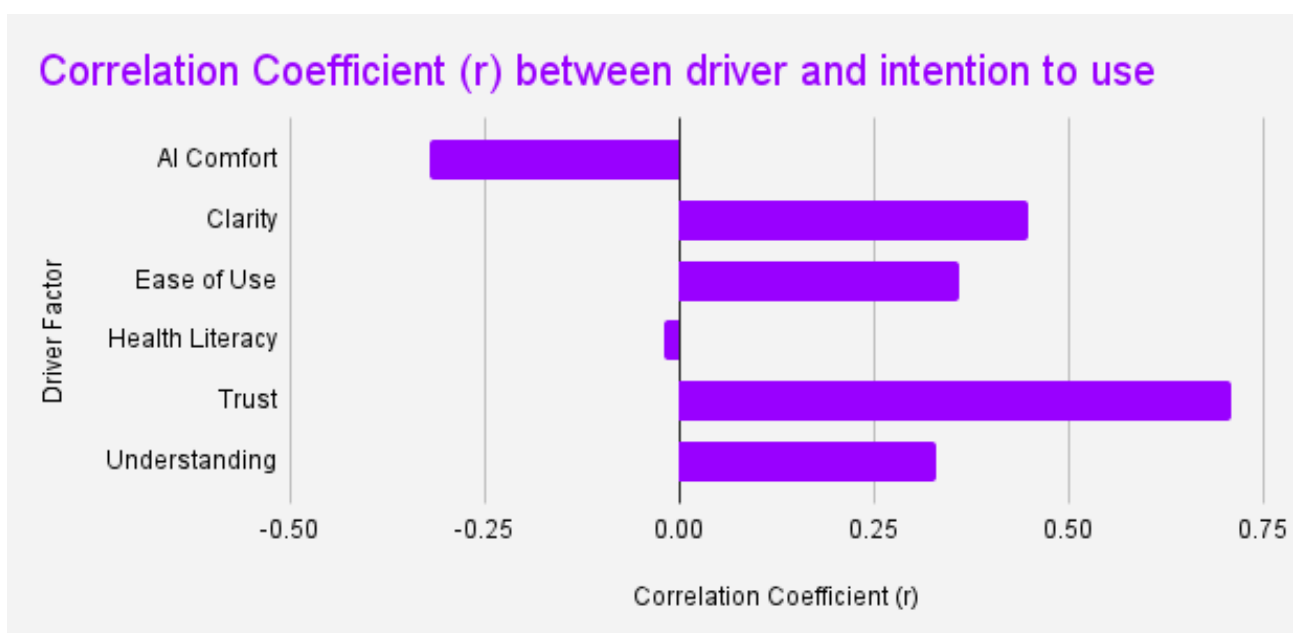


Figure 5.10: Correlation Coefficient between driver and intention to use

### 5.7.5 Preference and Regional Differences

Finally, participants were asked to make a forced choice regarding their preferred version for real-life use (Q33) and the version they found most trustworthy (Q34).

- **Real-Life Preference: Concept B** was the clear winner, selected by 73% (n=8) of participants.
- **Trustworthy Choice: Concept C** was selected as the “most trustworthy” by 55% (n=6), despite its poor usability scores.

A notable difference emerged between the perceived trust scores (S-TIAS) and the explicit choice. While participants rated Concept B highest on the multi-item trust scale (M=3.70 vs. M=3.58 for C) during the evaluation, the majority selected Concept C when forced to pick the “most trustworthy” option. This suggests a “halo effect” where the complexity of Concept C signaled competence in the abstract, even though users felt more reliable trust towards Concept B during actual interaction.

### Exploratory Analysis: The Urban-Rural Gap

A post-hoc analysis revealed a significant divide based on place of residence. Participants from the Randstad region reported higher Intention to Use scores across all three concepts compared to participants from the Countryside ( $p < .05$  for all versions). For Concept A, the gap was most pronounced (M\_Urban = 4.08 vs. M\_Rural = 2.50,  $p = .006$ ), suggesting a fundamental skepticism toward AI health tools in the non-urban demographic that interface design alone did not overcome.

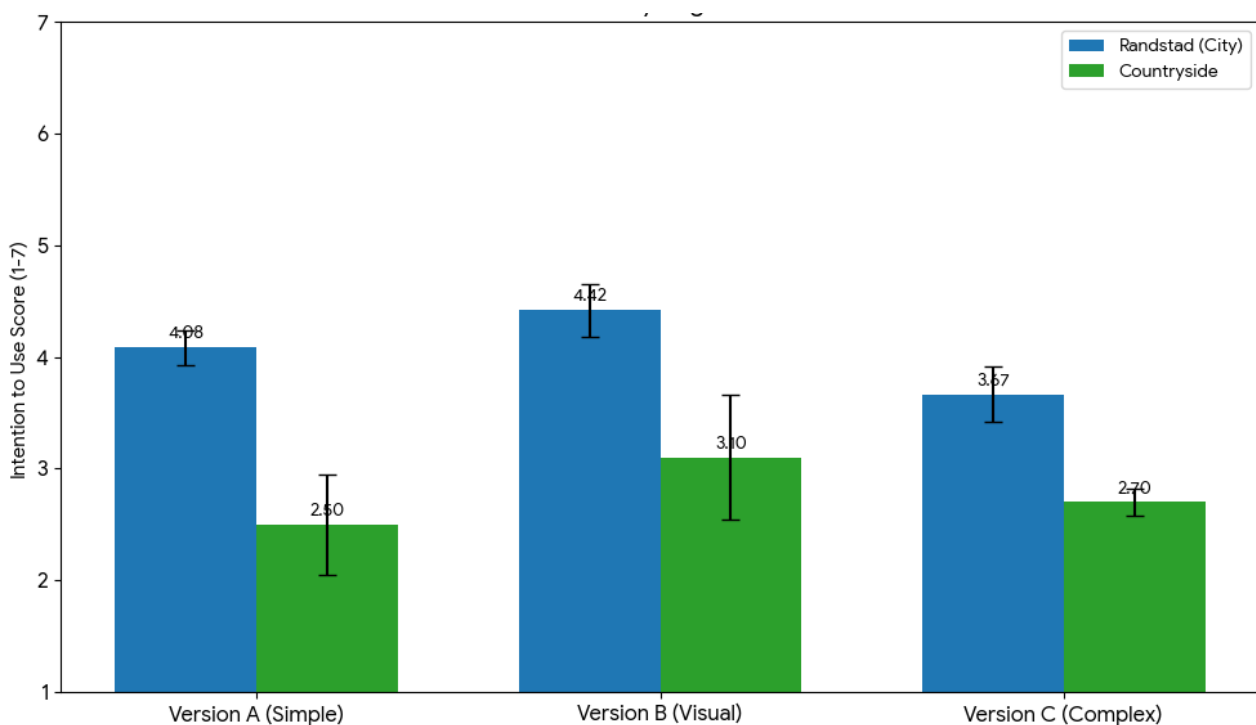


Figure 5.11: Mean Intention to Use scores (1-5) for Concepts A-C by place of residence (Randstad vs. Countryside).

## 5.8 Qualitative Results

This section presents the results of the qualitative analysis of the semi-structured follow-up interviews conducted after the user testing sessions (N = 11). The qualitative findings complement the quantitative results by providing deeper insight into how participants experienced trust, interface clarity, and information presentation across the three concepts. A full analysis of the results can be found in Appendix D.

### 5.8.1 Trust and Transparency of Data Use

A recurring theme across the interviews was the relationship between trust and transparency. Five participants explicitly stated that their trust in the system increased when the application displayed the specific data points used to reach a conclusion, such as heart rate, vaccination status, or elements of medical history. Seeing this data helped participants understand that the output was grounded in personal information rather than a generic response.

Related to this, four participants emphasized that trust increased when they could clearly see which data the AI had taken into account. These participants described data visibility as reassuring, as it demonstrated that the system was “looking at the right things” when generating medical advice.

In contrast, two participants expressed skepticism when the application provided a diagnosis or conclusion too quickly without asking sufficient follow-up questions. For these users, limited interaction and rapid conclusions reduced confidence in the system’s reliability.

### 5.8.2 Visual Support and Interface Preferences

Six participants strongly preferred the use of icons and images, as implemented in Concept B, to supplement textual explanations. These participants reported that visual elements helped them process information more easily and made the interface feel clearer and more approachable, particularly when explaining symptoms or next steps.

In addition, four participants highlighted the importance of prominent, single-purpose action buttons that clearly communicated what to do next (e.g., taking medication or contacting a doctor). These buttons were described as helpful in reducing uncertainty and making the advice feel actionable.

### 5.8.3 Perceived Information Overload and Interface Complexity

At the same time, multiple participants reported negative reactions to overly complex or visually dense interfaces. Four participants stated that screens containing many lines, data points, or visual elements felt “busy,” leading them to disengage from the information rather than examine it in detail. Similarly, four participants criticized Concept C for appearing cluttered and overwhelming.

More specifically, four participants described the Knowledge Graph or Mindmap used in Concept C as messy, confusing, or unreadable on mobile devices. These participants indicated that the visualization required too much effort to interpret, particularly in a health-related context where clarity and reassurance were expected.

Finally, three participants noted that irrelevant medical history, such as past injuries unrelated to the current complaint, should be hidden by default unless it was clearly applicable to the present condition. These participants expressed a preference for relevance over completeness in the presentation of personal health data.

## 5.9 Key Insights

This section synthesises the quantitative outcomes (Section 5.7) with the qualitative findings (Section 5.8) into key insights about how explanation design influences perceived trust, clarity, and intention to use patient-facing agentic health assistants. The aim is not to repeat the results, but to interpret the patterns across concepts and explain what they reveal about communicating AI guidance in healthcare.

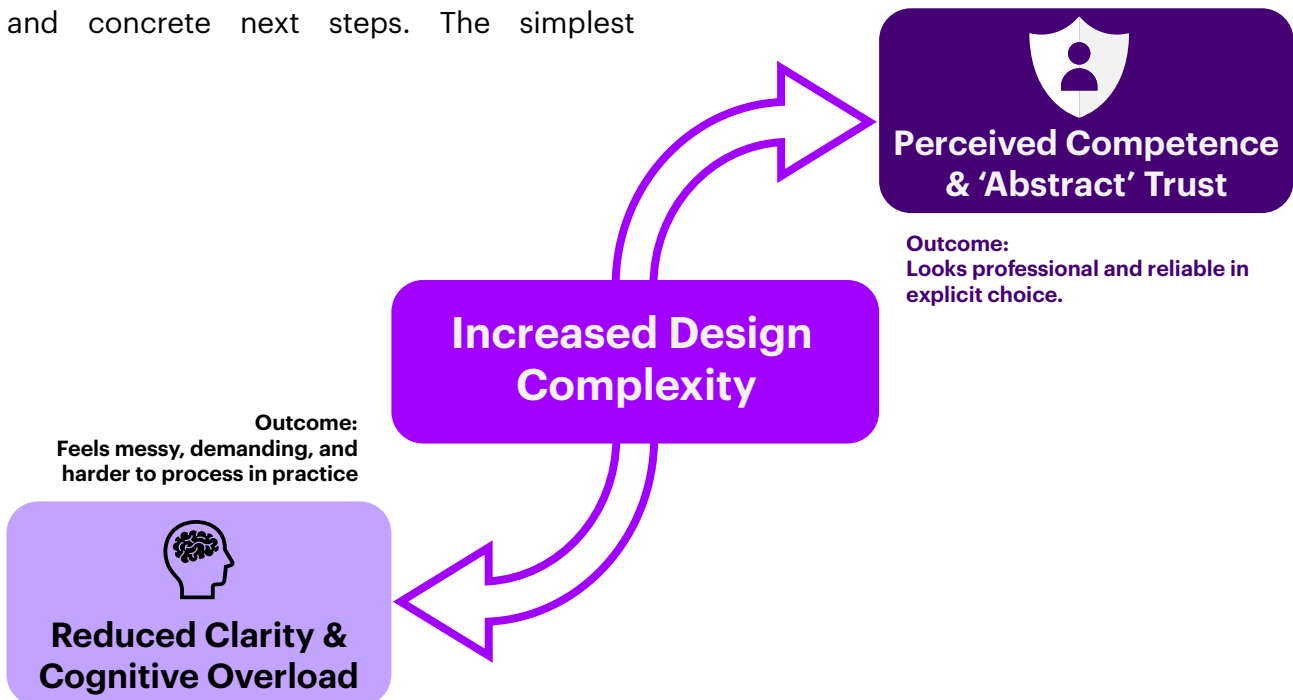
### 5.9.1 Outcome-first communication supports immediate usability

Across concepts, participants consistently valued interfaces that provide a complete, usable message immediately: a clear outcome and concrete next steps. The simplest

concept performed best on clarity and ease of use, indicating that reducing information density helps users quickly understand what the system is suggesting and what they should do. Qualitative comments reinforce this preference, describing reassurance and clear direction as the most valuable elements in the moment of uncertainty. Taken together, the findings suggest that outcome-first communication functions as the “anchor” of the interaction, even when additional explanation may exist elsewhere in the interface.

### 5.9.2 The Competence-Usability Paradox

A defining characteristic of the user testing results was the tension between credibility signalling and interaction cost. As detailed in the quantitative analysis, the complex concept was experienced as significantly less clear and harder to use. Yet, paradoxically, participants frequently selected it as the “most trustworthy” option when asked to make an explicit, abstract choice. This phenomenon is visualized in Figure 5.12.



**Figure 5.12:** *The Competence-Usability Paradox. Increased design complexity creates diverging outcomes: it signals competence in the abstract but increases cognitive burden during actual interaction.*

As illustrated in the figure, increasing the design complexity—through denser information and a more “expert-like” presentation—successfully creates a “credibility signal.” It leads to higher perceived competence and a form of abstract trust where the system looks professional. However, the simultaneous outcome is a substantial increase in interaction cost. The complex interface felt “messy” and demanding, leading to reduced clarity and cognitive overload.

Crucially, while clarity decreased strongly with complexity, users’ perceived understanding of the core medical message remained broadly similar. This suggests that the cognitive overload shown in the lower half of Figure 5.9. affected how clean and confidently interpretable the interaction felt, rather than whether the basic message was ultimately understood.

### **5.9.3 Visual support is not universally beneficial**

While Concept B (Visual) emerged as the most balanced option overall, the results indicate that visual support did not improve clarity for all users. For participants with lower health literacy, the visual layer appeared to add to the cognitive overload depicted in Figure 5.9.1 rather than alleviating it. This suggests that visuals help when they support rapid recognition and reduce reading burden, but become distracting noise when they introduce additional elements that vulnerable users feel they must interpret.

### **5.9.4 Transparency preferences depend on how evidence is presented**

Participants valued seeing that the assistant used relevant personal signals, as this made the guidance feel grounded rather than generic. However, the manner of presentation matters. The dense evidence presentation and extensive disclosure used in the complex concept contributed directly to the “Reduced Clarity” outcome in the paradox diagram,

especially on mobile devices. This explains why the complex concept could be perceived as trustworthy in the abstract, while still being experienced as difficult to navigate in practice.

### **5.9.5 Trust is the primary gatekeeper for adoption**

Across all concepts, willingness to use the assistant aligned most strongly with how much participants trusted a concept. While clarity and ease of use contributed to willingness, they appeared less decisive than trust. This indicates that simply making an interface easy is not enough; users must also feel confident in the system’s competence. The challenge for design is that the easiest way to signal that competence (complexity) often degrades the ease of use.

### **5.9.6 Synthesis: Implications for the Framework**

Collectively, these insights—particularly the paradox illustrated in Figure 5.9.1—point to the need for a staged explanation approach in patient-facing agentic AI.

Because a single interface style cannot simultaneously maximize immediate clarity and deep competence signaling, explanation needs to be structured into progressive levels. Design must protect the clarity of the default layer with outcome-first guidance, while providing accessible pathways to deeper layers where complexity and evidence can be presented without overwhelming the user at the point of primary interaction. Chapter 6 builds on these empirically grounded requirements to formalise a framework for governing these transitions.

# Deliver

## Framework, Roadmap and Prototype

Chapter 6 presents the Deliver phase of the project, in which the research outcomes are translated into concrete and implementable design outputs. In this chapter, the empirical insights and requirements from the earlier phases are consolidated into a staged approach to explainability that supports immediate user orientation while keeping deeper justification available when needed. The chapter's work centres on producing and formalising the Explanatory Depth Framework, which structures explanations into four progressive layers—Outcome & Next Steps, Brief Explanation, Traceability, and Deep Evidence—to help designers calibrate explanation depth to user needs and context without forcing complexity by default.

In addition to defining the framework, this chapter develops two supporting deliverables to enable practical adoption: a designer-facing playbook that specifies application logic, triggers, transition principles, and anti-patterns, and an implementation roadmap that outlines how the approach can be operationalised and scaled within Accenture Song's Design & Digital Product teams as agentic systems evolve. The chapter also applies the framework in a final prototype to demonstrate how layered explanation behaviour can be realised in a healthcare interaction and supported by system architecture, and it concludes by validating the framework through expert review and reflecting on implementation implications.

# 6



### AI used in this chapter:

- ChatGPT for rewriting of text

## 6.1 From Insights to Deliverables (Development Process)

This section explains how the empirical insights from the Develop phase were translated into four deliverables: (1) the Explanatory Depth Framework, (2) a designer-facing playbook, (3) a working demonstrator prototype, and (4) an implementation roadmap. The development logic was intentionally traceable: rather than “adding transparency,” the aim was to design how understanding is staged, so that the interface remains calm and actionable by default, yet becomes inspectable when users need reassurance, justification, or evidence. This direction directly follows the synthesis in Chapter 5, which concluded that staged explanation is required because a single fixed explanation layer cannot simultaneously reduce anxiety, maintain clarity, and support contestability.

### 6.1.1 Revising the design principles based on Develop-phase evidence

The deliver phase began by revisiting the five design principles established earlier in the project and refining them based on the observed breakdowns and preferences in user testing. The goal was not to replace the principles, but to make them operational. In other words, specific enough to drive concrete decisions about what should be shown first, what should remain optional, and how depth should be introduced without increasing cognitive load unnecessarily.

Three refinements were especially consequential for the final framework:

- Clarity as an adoption condition, not only a comprehension variable. Testing indicated that even when users appreciated transparency intentions, they disengaged when explanations felt effortful or visually dense. This positioned clarity and low effort as gatekeeping requirements for adoption, reinforcing the need for a calm default layer.

- Multimodality reframed as “modular fit” rather than universal benefit. User responses and earlier synthesis showed that visuals support understanding only when they enable rapid recognition (clear labels, familiar metaphors) rather than demanding interpretation. This resulted in global rules for the framework: text-first by default, self-explanatory visuals, and a fallback “text-only” mode for low health literacy or high stress.
- Depth control sharpened into outcome-first staging. Both interviews and testing supported the idea that users typically want orientation first and justification only when needed. The practitioner insight—“Most people call because they’re worried... Taking away that worry is really what builds trust”—clarifies why reassurance must precede detail in healthcare contexts.

These refinements shifted the design target from “maximal transparency” toward responsible transparency: make what matters visible, keep effort low, and ensure deeper evidence is available without being forced into the primary interaction.

### 6.1.2 Translating principles into a staged framework and supporting artefacts

Following this refinement, the insights were translated into a staged model with four explanation layers (Outcome & Next Steps, Brief Explanation, Traceability, Deep Evidence). The layers were arranged from simple to detailed to reflect a progression in user need: immediate action and reassurance first, followed by optional reasoning access, then verification of what inputs mattered, and finally full evidence access for auditability and dispute.

A deliberate design choice was to treat the deepest layer (Layer 4) as a core app capability rather than an exceptional transparency feature. This reflects the reality of agentic assistants that operate on longitudinal personal data: the system must be able to show the underlying data hub (records, timelines, linkages) when accountability requires it. At the same time, Layer 4 is explicitly framed as data access rather than explanation, keeping routine interactions psychologically safe and readable.

However, staged layers alone do not determine when depth should be shown. The central contribution of the framework is therefore not only the existence of layers, but the mechanism that selects the appropriate default layer in a way designers can understand and apply.

### 6.1.3 The Question Moderator: making depth selection designer-readable

The project's most consequential finding was that explanation depth is not merely a preference; it is state-dependent. Interviews and testing jointly indicated that the same "why" question can signal very different needs: reassurance under anxiety, verification under uncertainty, learning under curiosity, or contestability under dispute. The clinical expert summarised this principle succinctly: "People don't need to know every detail; they just need to understand why something is safe and what will happen next."

To operationalise this insight in a way that remains legible to designers, the framework introduces a Question Moderator that connects user state to explanation depth. In the healthcare demonstrator, four moderators were defined:

1. **Anxiety** (highest priority): because a large share of health queries are worry-driven, and anxiety increases the risk that additional detail becomes overload rather than reassurance.
2. **Uncertainty** (second priority): Because once anxiety is addressed, reducing "I'm not sure what this means" is the next most common barrier to trust—so a brief, plain-language "why" and interpretation should come before deeper evidence or data access.
3. **Curiosity** (medium priority): because curiosity should unlock depth, but not override emotional safety.
4. **Dispute** (lowest priority as a default driver): because dispute primarily functions as an escalation trigger to Layer 4 evidence access and correction pathways, rather than as a routine starting state.

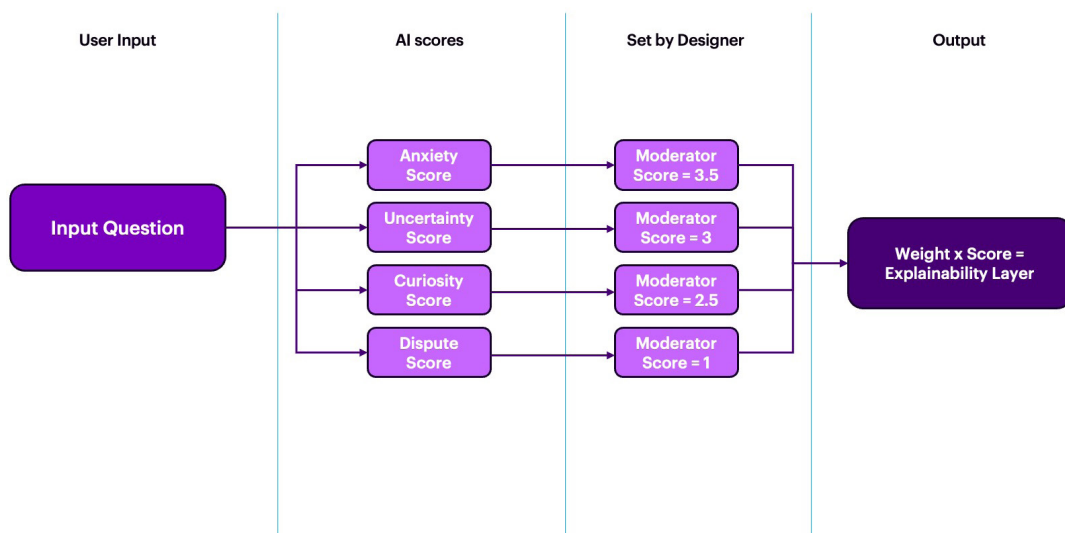
In this thesis depth selection is operationalised as a weighted moderator model. The moderator agent assigns each moderator a score from 1–5 per user turn. Designers define context-specific weights that reflect the relative importance of each moderator for safe communication in healthcare (Anxiety 35%, Uncertainty 30%, Curiosity 25%, Dispute 10%). For each moderator, a strength value is calculated as score multiplied by weight, and the moderator with the highest strength determines the default explanation layer. This approach keeps the mechanism interpretable and tunable: anxiety dominates unless another moderator is substantially stronger (e.g., uncertainty must score 5 to override anxiety at 4), while deeper layers remain available on demand.

Although the selection logic is visualised in figure 6.1 and further specified in Appendix E, it is intentionally described here in designer-readable terms:

- Each moderator is rated on a simple 1–5 scale using observable cues (e.g., calm vs distressed language, informational vs urgent situation, “just tell me” vs “show me everything,” uncertainty vs explicit disagreement).
- Each moderator has a weight that reflects its relative importance in the healthcare context, grounded in the practitioner insights above.
- The system then selects the dominant moderator and uses that to decide the default layer.

detail can increase anxiety rather than trust. In the demonstrating prototype, the moderator scoring and layer selection were implemented as a working computation to illustrate how a future “moderator agent” could govern explanation staging in real time. Importantly, the weighting scheme is presented for this thesis as a healthcare-specific parameterisation derived from qualitative (expert) logic, not as a universally validated model; other domains would require their own moderator research and calibration.

This “dominant moderator” logic is crucial for interpretability: it makes it clear why the interface stays shallow under distress even if a user is also curious. For example, if anxiety and curiosity are both high, the system defaults to Layer 1 to preserve calmness and reduce effort; deeper layers remain available but are not made the standard. This aligns with the playbook rule that users should not be forced through depth linearly and that unnecessary



**Figure 6.1:** How AI Defines which Layer to show

#### 6.1.4 Why four deliverables were needed

The deliver phase produced four deliverables because each addresses a different gap between insight and real-world uptake:

- **The Framework** defines the layer structure, global design rules (text-first, labelled visuals, fallback modes), and the Question Moderator as the mechanism that connects user state to depth.
- **The Playbook** translates the framework into designer responsibilities: When and how to use the framework, defining default layers, designing transitions, and identifying additional contextual moderators.
- **The Prototype** serves as a concrete example in a healthcare scenario, demonstrating how staged depth and the moderator logic could appear in interaction if the Framework is followed (the implementation is detailed in Section 6.3).
- **The Roadmap** addresses feasibility: it frames how the framework can be applied today as a designer-led scaffold (Horizon 1) while evolving into a governance mechanism as agentic systems increasingly assemble interfaces dynamically (Horizons 2–3).

Together, these deliverables ensure the contribution is not only conceptual but also actionable: the framework explains what staged explainability is, the moderators explain how and when depth is selected, the prototype demonstrates what it looks like, and the roadmap clarifies how implementation can become realistic over time.

The next section (6.2) details the framework components and structure, including the four layers, the global rules, and the full moderator specification used in the healthcare demonstrator.

## 6.2 The Explanatory Depth Framework: Components and Structure

Figure 6.2 provides an overview of the Explanatory Depth Framework and its components. To support readability, the figure is annotated with numbered callouts (red boxes). The section below follows the same numbering to allow direct cross-referencing between the visual and the explanation.

### **(1) Explanatory Depth: Simple to Detailed**

The depth layers express the core organising principle of the framework: explanations can be delivered at different depths, from minimal and action-oriented to evidence-rich and inspectable. The continuum is included to clarify that “more depth” is not inherently better; it is appropriate only when it matches the user’s need and context.

### **(2) Layer 1 — Outcome & Next Steps**

Layer 1 is the most concise layer and communicates the conclusion and immediate next steps in plain language. Its purpose is to provide fast orientation and enable safe action without requiring the user to interpret underlying reasoning.

### **(3) Layer 2 — Brief Explanation**

Layer 2 provides a short, non-technical justification that answers “why” at a high level. It exists to reduce uncertainty and improve perceived legitimacy while remaining skimmable and low-effort.

### **(4) Layer 3 — Traceability**

Layer 3 exposes what inputs mattered and how they relate to the outcome. Its role is to support verification (relevance, fairness, correctness) without requiring full raw-data inspection.

### **(5) Layer 4 — Deep Evidence**

Layer 4 provides access to underlying evidence and raw data (e.g., records, timeline views, data sources). It is positioned as a core capability for accountability, auditability, and dispute handling, rather than as the default explanation layer.

### **(6) Question Moderators**

The Question Moderator is the framework component that connects user state and intent to the selection of an appropriate default depth. In the healthcare demonstrator, it is represented through a small set of interpretable moderators (anxiety, uncertainty, curiosity, dispute), which designers can use as a practical lens when deciding what layer should be surfaced by default in a given interaction.

### **(7) UI Elements**

This component specifies the interface building blocks that instantiate each layer. Rather than treating layers as abstract labels, the framework ties them to concrete UI patterns (e.g., outcome statements and actions in Layer 1, “because of...” summaries in Layer 2, input-to-outcome linkages in Layer 3, and data hub/timeline views in Layer 4). This is included to reduce ambiguity and make the framework directly usable in design work.

### **(8) Global interaction rules + two-way transparency**

This component captures rules that apply across all layers to preserve clarity and user control. It includes principles such as progressive disclosure (depth remains optional), consistent tone, and persistent two-way transparency features (e.g., a mechanism to ask questions, correct inputs, or express doubt regardless of depth).

**(9) Design requirements (quality criteria across layers)**

The design requirements define what “good” looks like beyond simply showing more information. They function as quality criteria that guide evaluation and implementation—covering aspects such as clarity, low effort, relevance filtering, actionability, comprehensibility, user control, traceability, and auditability.

**(10) Cross-domain translation layer**

The cross-domain layer demonstrates how the same layer logic can be applied to healthcare, finance, and insurance by mapping each layer to domain-appropriate content and evidence types. Its purpose is to show that the framework governs how explanations are staged, while allowing the content and evidence to vary by domain.

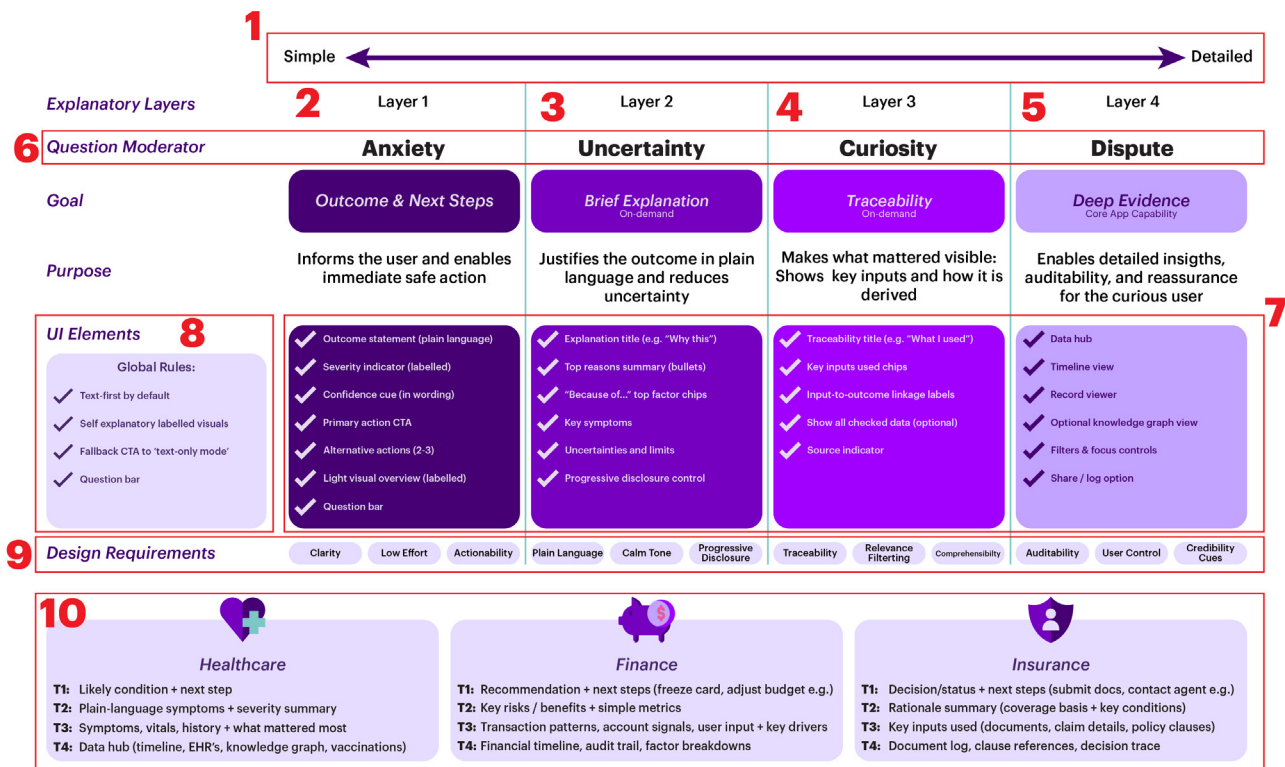


Figure 6.2: The Explanatory Depth Framework with numbered callouts

# Explanatory Depth Framework



Explanatory Layers	Layer 1	Layer 2	Layer 3	Layer 4
<b>Question Moderator</b>				
<b>Goal</b>	<b>Outcome &amp; Next Steps</b>	<b>Brief Explanation</b> <small>On-demand</small>	<b>Traceability</b> <small>On-demand</small>	<b>Deep Evidence</b> <small>Core App Capability</small>
<b>Purpose</b>	Informs the user and enables immediate safe action	Justifies the outcome in plain language and reduces uncertainty	Makes what mattered visible: Shows key inputs and how it is derived	Enables detailed insights, auditability, and reassurance for the curious user
<b>UI Elements</b>	<ul style="list-style-type: none"> <li>Outcome statement (plain language)</li> <li>Severity indicator (labelled)</li> <li>Confidence cue (in wording)</li> <li>Primary action CTA</li> <li>Alternative actions (2-3)</li> <li>Light visual overview (labelled)</li> <li>Question bar</li> </ul>	<ul style="list-style-type: none"> <li>Explanation title (e.g. "Why this")</li> <li>Top reasons summary (bullets)</li> <li>"Because of...": top factor chips</li> <li>Key symptoms</li> <li>Uncertainties and limits</li> <li>Progressive disclosure control</li> </ul>	<ul style="list-style-type: none"> <li>Traceability title (e.g. "What I used")</li> <li>Key inputs used chips</li> <li>Input-to-outcome linkage labels</li> <li>Show all checked data (optional)</li> <li>Source indicator</li> </ul>	<ul style="list-style-type: none"> <li>Data hub</li> <li>Timeline view</li> <li>Record viewer</li> <li>Optional knowledge graph view</li> <li>Filters &amp; focus controls</li> <li>Share / log option</li> </ul>
<b>Design Requirements</b>	Clarity, Low Effort, Actionability	Plain Language, Calm Tone, Progressive Disclosure	Traceability, Relevance Filtering, Comprehensibility	Auditability, User Control, Credibility Cues



## Healthcare

- T1:** Likely condition + next step
- T2:** Plain-language symptoms + severity summary
- T3:** Symptoms, vitals, history + what mattered most
- T4:** Data hub (timeline, EHR's, knowledge graph, vaccinations)



## Finance

- T1:** Recommendation + next steps (freeze card, adjust budget e.g.)
- T2:** Key risks / benefits + simple metrics
- T3:** Transaction patterns, account signals, user input + key drivers
- T4:** Financial timeline, audit trail, factor breakdowns



## Insurance

- T1:** Decision/status + next steps (submit docs, contact agent e.g.)
- T2:** Rationale summary (coverage basis + key conditions)
- T3:** Key inputs used (documents, claim details, policy clauses)
- T4:** Document log, clause references, decision trace

## 6.3 Applying the Framework

This section explains how the Explanatory Depth Framework is intended to be used in practice: its purpose, intended users, applicable domains, and the designer responsibilities it implies. In contrast to Section 6.2 (which defines the framework components), the emphasis here is procedural—i.e., how teams should apply the framework during design and delivery of agentic AI interfaces.

### 6.3.1 Purpose and intended users

The framework is intended as a design scaffold for creating explainable, trustworthy, and controllable user interfaces for agentic AI in high-stakes contexts. It supports designers in staging explanations over time—balancing clarity, emotional safety, and transparency without overwhelming users.

The primary intended users are digital product, UX, and interaction designers working on agentic AI experiences. Within this thesis, the audience is explicitly framed as designers at Accenture Song, where the framework is positioned as a reusable method for shaping human-facing explanations as systems move from advisory assistance toward more autonomous decision-making.

A key boundary condition is that the framework does not prescribe a fixed set of UI components or a visual style. Instead, it defines design intent, interaction logic, and explanation strategy that can be adapted to a product's existing design language and technical constraints.

### 6.3.2 When to use the framework

The framework is designed for scenarios in which an AI system influences user decisions under conditions of:

- High sensitivity of personal data, and
- Meaningful consequences (health, financial, or legal), and
- Asymmetric expertise between the system and the user, and
- Accountability expectations, including recourse and escalation.

In such contexts, explainability must be treated as an interaction strategy rather than an “information dump.” The framework is therefore most relevant when design teams must decide what to show by default, what to keep optional, and how to enable contestability without increasing anxiety or cognitive load.

### 6.3.3 Domains and scope of applicability

The playbook specifies three target domains for cross-domain applicability:

Healthcare, Finance, and Insurances.

The rationale for choosing these domains is not that they share identical content, but that they share a similar interaction problem: users must provide intimate information, accept outcomes based on reasoning they may not fully understand, and trust that their data is used responsibly and proportionally.

This creates a recurrent tension between (a) the need for usable orientation and (b) the need for accountability and evidence access. The framework is positioned as a response to that tension by separating explanation from raw data exposure and enabling controlled, contextual access to understanding.

### 6.3.4 Core application principle

A guiding principle for applying the framework is that full transparency is rarely the default user goal in high-stakes situations. Instead, users typically seek orientation first, understanding second, and justification only when needed. Operationally, this means teams should design Layer 1 as the stable default and treat deeper layers as purposeful options—invoked when the user’s intent, uncertainty, or risk context requires it, rather than as a static “always-on” layer.

### 6.3.5 Designer responsibilities and application workflow

Applying the framework shifts the designer’s role from “screen design” toward explanation orchestration: designing not only what is shown, but when and under what conditions depth is revealed.

In the playbook, this is translated into five practical responsibilities:

- 1. Identify contextual moderators (beyond the generic moderator set).** Before implementing layer behaviour, designers should identify context-specific factors that constrain or shape explainability—such as regulatory requirements, domain risk levels, organisational escalation policies, and cultural sensitivities. These moderators affect when layers may be shown, how data can be exposed, and when human escalation is required.
- 2. Define the default layer.** Layer 1 should be the system’s first response unless there is a clear reason otherwise. This ensures the experience remains readable and actionable for the majority of situations and prevents premature complexity.
- 3. Design transitions between layers.** Movement to deeper layers should be intentional, user-initiated, and reversible. In other words, progression must be designed as a deliberate interaction

pattern, not left to emergent system behaviour.

- 4. Maintain tone consistency across layers** Deeper layers should add clarity and evidence without changing emotional framing or implying stronger certainty than warranted. This prevents a common failure mode in which Layer 1 feels calm and supportive but deeper layers feel clinical, defensive, or contradictory.
- 5. Ensure two-way transparency at every layer** Each layer must allow users to ask questions, correct inputs, or express doubt. This embeds contestability and participation as interaction properties rather than “end-of-flow” features.

### 6.3.6 Anti-patterns to avoid

To prevent explainability from undermining trust, the playbook identifies three anti-patterns that should be treated as explicit design “don’ts”:

- Exposing raw data without context (increases interpretive burden and anxiety)
- Forcing users into technical explanations (breaks accessibility and often fails to reassure)
- Treating explainability as a one-time disclosure (ignores the conversational nature of user doubt and evolving need)

In combination, these anti-patterns reinforce the central application stance of the framework: explainability should function as controlled access to understanding, with depth staged to preserve usability and emotional safety while still enabling accountability when needed.

## 6.4 Final Prototype (Demonstration of the Framework)

To demonstrate the practical implications of the Explanatory Depth Framework in a concrete interaction, a working prototype was developed as a domain-specific demonstrator. The prototype's purpose is not to validate clinical accuracy, but to show (a) how the four-layer explanation structure can be instantiated in an end-to-end interaction, (b) how the system can adapt the default explanatory depth to the user's query, and (c) how a "deep evidence" layer can exist as a core capability without becoming the default experience.

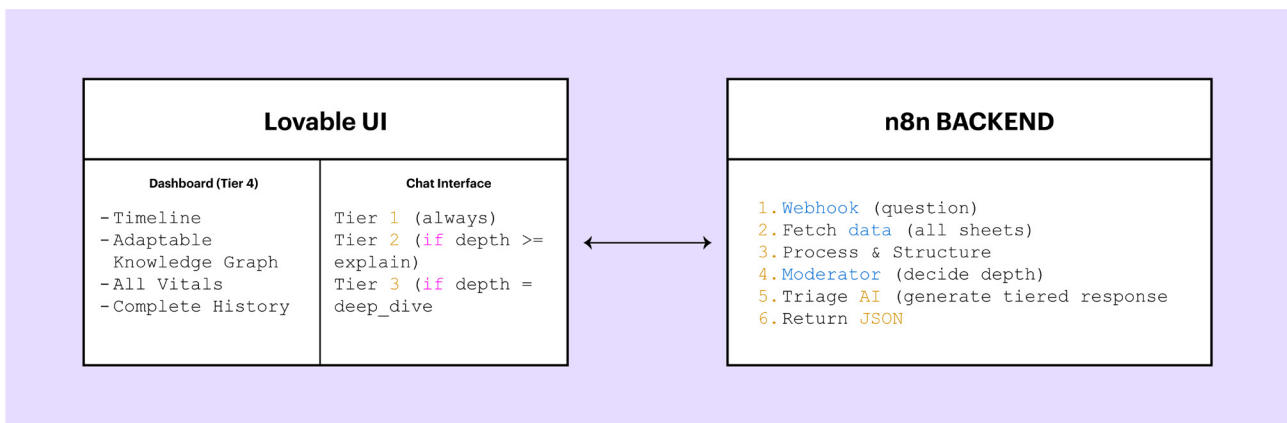
### 6.4.1 Synthetic patient dataset and scenario fidelity

Because the prototype required longitudinal, multi-source health data, a synthetic patient dataset was generated using Mostly AI. The synthetic persona represents a 30-year-old male and includes a 60-day log of wearable-derived signals, alongside structured health context such as vaccinations, allergies, EHR-style records, and laboratory results. The dataset was consolidated in a Google Spreadsheet to enable rapid retrieval. The design intent was to approximate a plausible "2028 scenario" in which patient-facing agents operate on integrated, patient-owned data ecosystems, without introducing privacy risk or relying on real patient data.

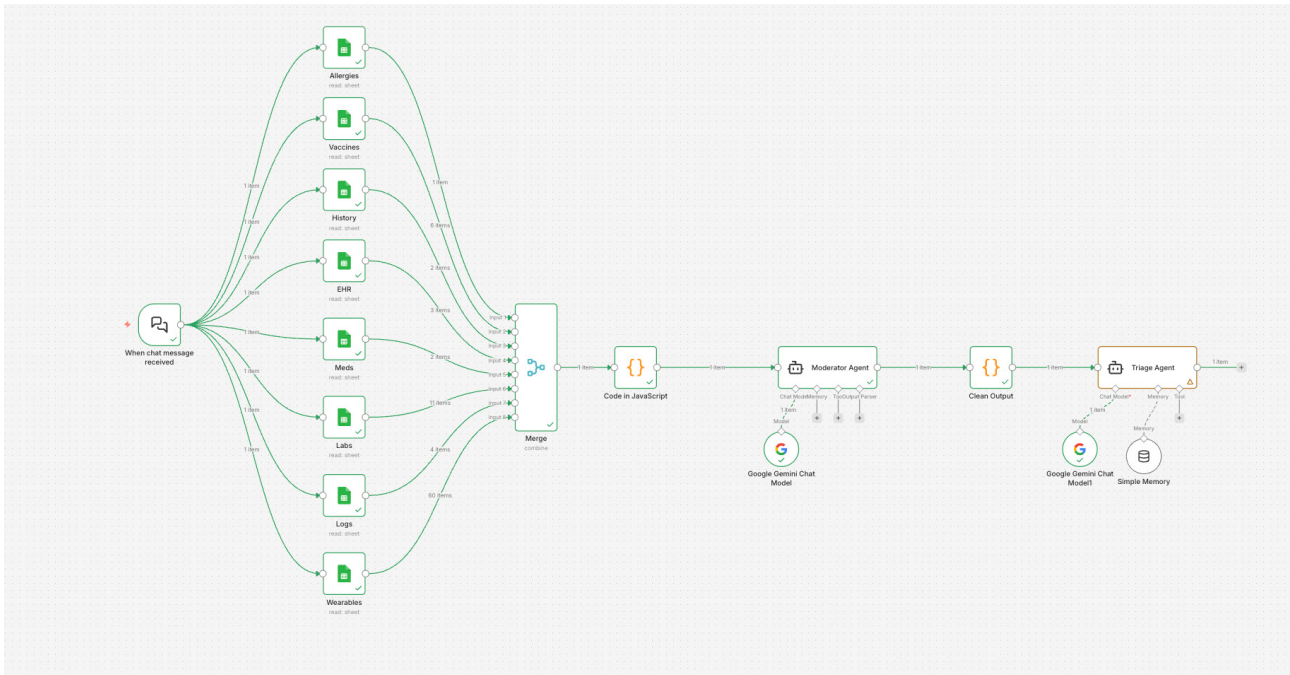
### 6.4.2 Prototype implementation and interaction flow

**Front-end (Lovable):** The user interface was created using Lovable, a prompt-driven "vibe coding" tool that generates a working web front-end from natural-language instructions (Lovable, 2026). In practice, Lovable functions as an AI code generator: it produces the layout, UI components, and interaction logic based on iterative prompts, without requiring manual front-end development. Within this prototype, Lovable served as the presentation layer that captures user queries, builds the Layer-4 data hub, and renders the returned response in the appropriate explanation layer (Layer 1–4) as specified by the back-end workflow.

**Back-end orchestration (n8n):** The back-end logic was implemented in n8n, a visual workflow automation platform that allows system behaviour to be defined as a sequence of connected steps ("nodes") rather than handwritten code (n8n, 2026). In this prototype, n8n functioned as the orchestration layer: it received the user's query from the front-end, retrieved relevant patient data from the Google Spreadsheet, triggered the triage and moderation steps, and returned the final layer-specific output to the interface in JSON code.



**Figure 6.3:** The architecture of the Lovable prototype connected to the n8n backend



**Figure 6.4:** *Prototype back-end n8n workflow*

**Triage agent:** The user’s prompt and the retrieved patient context were used to produce a triage-oriented response. This agent combined the patient-specific signals with general medical knowledge provided through a Google Gemini LLM.

**Moderator agent:** The triage output and user prompt were routed to a dedicated moderation step. To avoid “black box” behavior, this agent was not a generic LLM chat node, but a structured classifier. It utilized a specific system prompt to analyze the user’s input for semantic markers of the four moderators (e.g., detecting “anxiety” through repetitive phrasing or urgent terminology). It scored each moderator on the 1–5 scale defined in the framework and outputted a structured JSON object containing the winning layer, rather than generating free text.

**Layered response assembly:** The chosen layer was sent back to the Lovable front-end, where the response was generated and rendered in the corresponding depth format. This pipeline allowed the prototype to operationalise the framework’s core promise: users receive an outcome-first response when appropriate, while deeper reasoning and evidence can be staged or accessed when needed—without forcing complexity into every interaction.

To try the working prototype, scan the QR code below:



### 6.4.3 Demonstration outputs

The prototype was used to generate multiple example interactions that illustrate how default explanation depth changes across different user intents and states:

#### **Acute distress / high anxiety (Layer 1 default):**

Figure 6.5 shows the output for the prompt: “I almost can’t breathe!” In this case, anxiety cues were detected and the system defaulted to Layer 1, presenting immediate orientation and next-step guidance rather than extended justification.

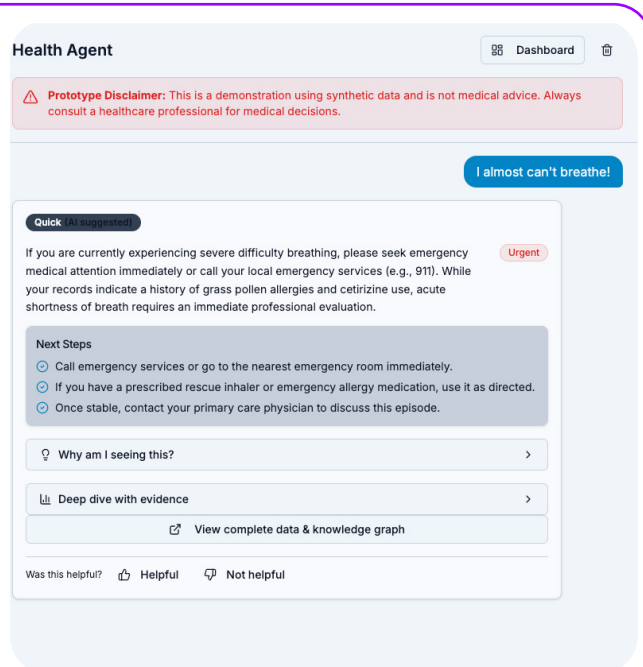


Figure 6.5: Prototype output of Layer 1

Health Agent

Dashboard

Prototype Disclaimer: This is a demonstration using synthetic data and is not medical advice. Always consult a healthcare professional for medical decisions.

I have pain in my knee, what can I do?

Explained

Show less

Your knee pain may be related to the new workout routine you started on December 22nd, which likely increased the load on your joints. While your activity levels (steps) remain consistent, your body may be adjusting to different types of movement or intensity.

Monitor

Next Steps

- Try the RICE method (Rest, Ice, Compression, Elevation) for 48 hours
- Modify or reduce the intensity of your new workout routine
- Monitor for swelling, redness, or inability to bear weight

Why am I seeing this?

At age 34 with no chronic conditions, this is likely an acute musculoskeletal strain, though it could impact your ability to maintain your 9,000+ daily step average.

Steps

Averaging 9,535 steps today, showing high activity despite potential discomfort.

HRV

68.2 today suggests your body is managing systemic stress reasonably well.

Temperature

36.8°C indicates no systemic fever or significant inflammation.

Low Risk

Moderate Confidence

Show evidence

View complete data & knowledge graph

Was this helpful? Helpful Not helpful

#### **Everyday, moderate-impact query (Layer 2 default):**

Figure 6.6 shows the output for: “I have pain in my knee, what can I do?” With no strong anxiety cues and no explicit demand for detailed reasoning, the system defaulted to Layer 2, providing a brief, plain-language rationale alongside actionable steps.

Figure 6.6: Prototype output of Layer 2

**Explicit request for data-based justification (Layer 3 default):**

Figure 6.7 shows the output for: “I have difficulty sleeping. Explain in detail why and what the data shows, and what I should do?” Here the prompt signals higher informational intent (curiosity/verification), leading the system to default to Layer 3, where key inputs and their relevance to the outcome are surfaced more explicitly

**Figure 6.7:** Prototype output of Layer 3

**I have difficulty sleeping. Explain in detail why and what the data shows, and what I should do**

**Deep Dive** | Management | Show less

Your data suggests a combination of lifestyle factors may be impacting your sleep quality, including recent work-related stress and a new workout routine initiated in December. While your current sleep duration averages around 7.4 hours, your logs indicate subjective struggles with noise and fatigue that correlate with periods of high activity and stress. Monitor

**Next Steps**

- Maintain a consistent sleep schedule to stabilize your 7.4-hour baseline
- Review the timing of your new workout routine to ensure it isn't too close to bedtime
- Discuss your slightly low Vitamin D levels (48 nmol/L) with your doctor, as this can sometimes impact sleep quality

**Why am I seeing this?**

Sleep disruption appears to be episodic rather than chronic, often linked to environmental or lifestyle stressors, which may affect your daytime energy and recovery.

<b>Sleep Duration</b> Averaging 7.4-7.8 hours recently, though logs show subjective poor quality on Dec 3rd.	<b>Heart Rate Variability</b> Recent rise to 88.2-76.8 suggests good recovery despite reported stress.	<b>Daily Steps</b> Consistently near 9,000-10,000 steps, indicating high physical activity.
---	---	--

Low Risk | High Confidence

**Show evidence**

- Lifestyle** (high)  
Stress and environment  
Logs from Dec 16 (work stress) and Dec 3 (neighbor noise) coincide with reports of poor sleep.
- Wearable** (medium)  
Activity correlation  
High step counts (e.g., 18,500 on Dec 28) followed by variable sleep suggests physical exertion impacts rest cycles.
- Labs** (medium)  
Nutritional deficiency  
Vitamin D measured at 48 nmol/L on Nov 5 is below the optimal range.

**Patterns Identified**

- Stress-induced sleep maintenance issues mapped to work deadlines
- Potential overstimulation from new physical workout routines started in late December

[View complete data & knowledge graph](#)

**Layer 4 evidence access and knowledge graph view:**

Figure 6.8 illustrates the “deep evidence” representation (Layer 4) in the sleep scenario, including a knowledge-graph style view that highlights which data points are treated as influential. In the framework logic, this layer functions as an inspectable evidence space rather than an explanation that must be consumed by default.

**Figure 6.8:** Prototype output of Layer 4, a core app capability

**Health Dashboard** | Chat with Age

**Prototype Disclaimer:** This is a demonstration using synthetic data and is not medical advice. Always consult a healthcare professional for medical decisions.

Overview | Timeline | Vitals | Symptoms | Graph

**Health Knowledge Graph**  
23 entities • 43 relationships • 5 focused

AI-identified entities: Sleep (100%) | Stress (90%) | Workout routine (80%) | Heart Rate Variability (80%) | Caffeine (60%) | Vitamin D (50%)

Filter (click to hide): Symptom (12) | Vital (2) | Condition (12) | Lab (7) | Medication (2) | Vaccine (1) | Allergy (1) | Cause (1)

**Relationship Types**

- Causes
- Correlates
- Improves
- Worsens
- Affects

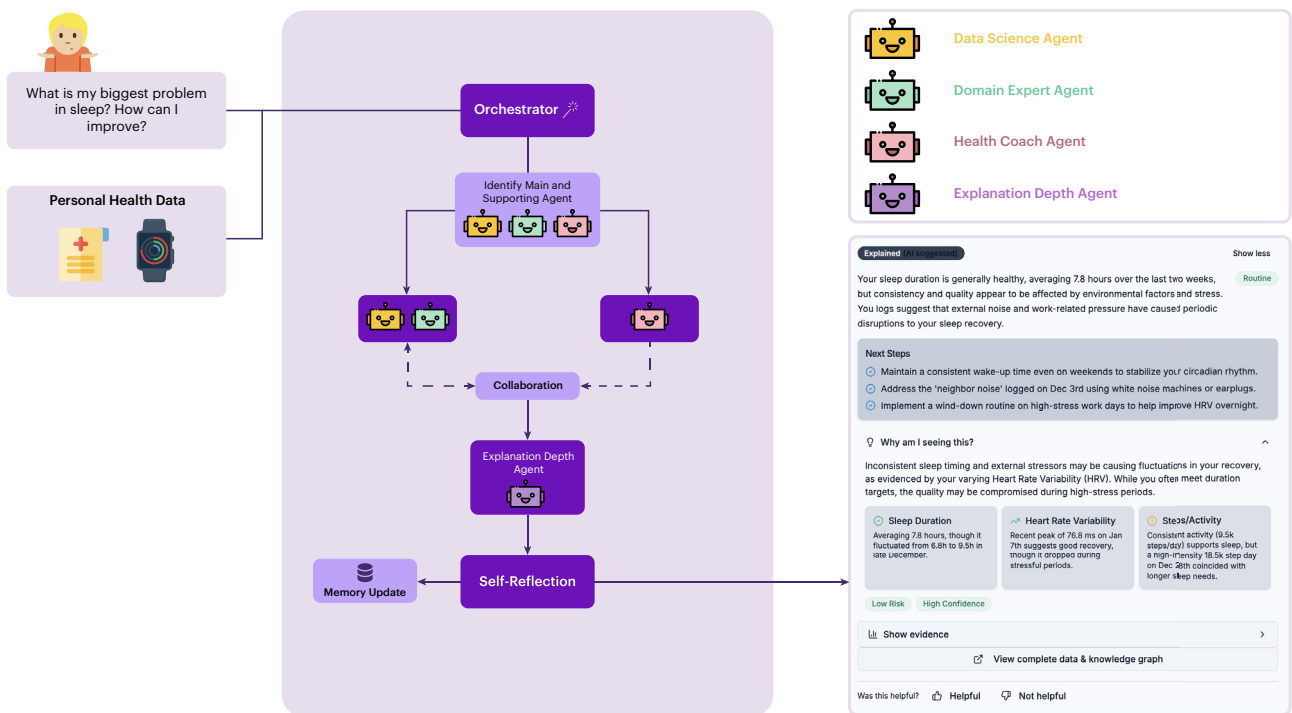
### 6.4.4 System architecture

Figure 6.9 presents the system architecture used for the demonstrator, adapted from the Personal Health Agent (PHA) reference model proposed by Heydari et al. (2025). In the original PHA framing, an orchestrator coordinates specialised agents (e.g., Data Science, Domain Expert, Health Coach) that collaborate over personal health data to produce user-facing support.

In this thesis prototype, the architecture is extended by explicitly introducing a Moderator agent responsible for staging how results are communicated to the user. Concretely, after the orchestrator identifies the main and supporting agent(s) and a collaborative output is produced, the explanation depth agent governs how that output is rendered (Layer 1–Layer 4), aligning the interface response with the framework’s staging logic.

This addition makes the explainability layer a dedicated architectural component—rather than an afterthought in presentation—supporting the thesis argument that explainability in agentic systems must be designed as interaction governance, not just information disclosure.

**Architecture of the Personalized Health Agent**



**Figure 6.9:** Architecture of the newly proposed Personalized Health Agent.

## 6.5 Roadmap for Implementation and Future Development

Page 77 presents a three-horizon roadmap that translates the Explanatory Depth Framework from a near-term, designer-led method into a longer-term governance mechanism for agentic user experiences. The roadmap is included because the framework's core value—staging explanation depth, tone, and evidence exposure—becomes increasingly relevant as interfaces shift from fixed screens to dynamically assembled experiences. At the same time, the roadmap acknowledges that organisational feasibility, tooling maturity, and risk constraints require a staged adoption approach rather than a single-step transformation.

### 6.5.1 Future vision

The roadmap is anchored by a 2028 future vision:

*By 2028, Accenture Song's Design & Digital Products teams will design and govern agentic user experiences that adapt how reasoning is communicated, improving trust and understanding in high stakes situations.*

This vision builds directly on the thesis' central premise: in high-stakes contexts, trust is not achieved by full transparency by default, but by controlled, contextual access to understanding—where orientation is provided first and justification is available when needed.

### 6.5.2 Why the roadmap uses horizons

The roadmap is structured into horizons to reflect three sequential maturity shifts that cannot responsibly be skipped:

1. Experience construction: from designer-authored interfaces to AI-assembled interfaces to generative UI.
2. Designer responsibility: from designing interfaces, to designing interaction logic, to governing AI behaviour.
3. Accountability needs: as adaptivity increases, the need for explicit boundaries, safeguards, and escalation workflows grows (especially in domains with sensitive personal data and consequential outcomes).

In other words, horizons are included to align ambition with feasibility and risk: the framework can be adopted immediately, but its organisational role changes as agentic systems mature.

### 6.5.3 Horizon-by-horizon: what changes

#### ***Horizon 1 (2026) — Designer-led adoption (foundation building).***

The primary change in Horizon 1 is standardisation, not automation. Designers apply the framework manually to establish a consistent layer vocabulary and to embed layered explanation into deliverables such as component libraries, interface designs, and end-to-end user flows. In practice, this horizon is about making “explanation staging” a repeatable design practice (and review criterion) rather than an ad hoc decision made per project.

#### ***Horizon 2 (2027) — AI-assisted assembly (designers as rule-makers).***

Horizon 2 shifts the centre of gravity from producing interface variants to producing rules and safeguards. The roadmap makes this explicit by introducing deliverables such as “rules, constraints & safeguards” and “AI workflow architecture.” The playbook clarifies the underlying behavioural shift: designers stop designing every instance and instead define constraints for layer selection, data exposure, and tone, then periodically review and tune system behaviour.

#### ***Horizon 3 (2028) — Generative UI (governance and escalation).***

In Horizon 3 the framework functions as a guardrail that governs how adaptive interfaces communicate decisions and expose evidence, including explicit routing of high-stakes situations to human experts. The roadmap therefore introduces governance-oriented deliverables—such as “experience boundary definitions” and “escalation & handover workflow”—reflecting the reality that in highly dynamic interfaces, accountability depends on boundaries and recourse mechanisms, not on fixed screens. Across all horizons, the layer structure itself does not change; what changes is who applies it and how often, progressing from continuous designer authoring to episodic policy updates.

## Role of the Framework

## Role of Designers

## Role of AI Systems

The framework helps designers structure clear, safe, and understandable interfaces for high-stakes AI decisions.

The framework provides a component set and interaction rules so AI can assemble tiered explanations consistently across contexts.

The framework acts as a guardrail that governs how agentic systems communicate decisions, adapt explanations, making it hyper-personal.

### Designing Interfaces

#### Designing Interaction Logic

#### Governing AI Behaviour

AI generates content within designer-defined interfaces and explanation levels, without deciding how or when explanations are shown.

AI assembles interfaces dynamically by selecting appropriate explanation tiers based on context, while following designer-defined rules and constraints.

AI autonomously adapts explanation depth and interaction flows within clear guardrails, and routes high-stakes situations to human experts.



## Future Vision

By 2028, Accenture Song's Design & Digital Products teams will design and govern agentic user experiences that adapt how reasoning is communicated, improving trust and understanding in high stakes situations

## Deliverable

### Component Library & Design Language

#### Interface Designs

#### End-to-end User Flows

### Rules, Constraints & Safeguards

#### AI Workflow Architecture

#### Component Usage Rules

### Experience Boundary Definitions

#### Escalation & Handover Workflow

### 6.5.4 Implications for Accenture Song's design process

The roadmap implies a gradual evolution in Accenture Song's design practice for agentic products:

- **From screen design to explanation orchestration.** The playbook explicitly frames this as a role shift: designers become responsible for orchestrating what is shown by default, what remains optional, and how users move between depths.
- **From artefacts to behavioural specifications.** Over time, teams increasingly deliver interaction policies (layer rules, transition constraints, disclosure boundaries) alongside traditional UI deliverables.
- **From usability-only evaluation to accountability evaluation.** Evaluation expands from “is this understandable?” toward “is this contestable, traceable, and safe under stress?”—particularly when personal, financial, or legal consequences are involved.
- **From isolated design decisions to shared governance.** The later horizons require tighter coupling between design, engineering, AI, and domain experts because safeguards (data exposure, escalation, logging) must be enforceable at system level.

A practical near-term implication (Horizon 1) is that the framework should be integrated into existing D&DP routines as an explicit checkpoint: defining the default layer strategy, specifying layer transitions, and documenting how evidence access and escalation are handled in the experience.

### 6.5.5 Kick-off Workshop

To operationalise Horizon 1 adoption, the roadmap was initiated through a structured kick-off workshop facilitated by the author. The workshop's primary purpose was knowledge transfer: to communicate the thesis insights

and the Explanatory Depth Framework to relevant stakeholders, so they can apply the underlying logic in future design decisions involving patient-facing agentic assistants.

The session consisted of a presentation and an open discussion. The presentation summarised the central problem identified in this thesis—how trust and understanding depend on how reasoning is communicated rather than on model performance alone—and introduced staged explainability as a practical design response. The discussion that followed allowed participants to ask clarifying questions and to challenge assumptions, which helped sharpen how the framework should be communicated and positioned for real organisational settings.

The workshop included members of Accenture Song's Design D&DP team, Accenture Health & Public Services (H&PS), and designers from the Amsterdam Health & Technology Institute (AHTI). This audience was relevant because it reflects the professional ecosystem in which such patient-facing interfaces are likely to be shaped: product designers, healthcare innovation stakeholders, and organisational actors who influence implementation constraints.

## 6.6 Validating the Framework (Expert Review)

This thesis did not conduct a formal expert validation study (e.g., semi-structured interviews with predefined protocols and systematic analysis). Nevertheless, the framework and roadmap were reviewed informally through open conversations and the kick-off workshop described in Section 6.5.5, involving Accenture Song D&DP, Accenture AI & Data, Accenture's H&PS, and AHTI designers.

The feedback gathered in these exchanges is therefore treated as formative plausibility input, not evaluative evidence. Its primary contribution was not to “prove” the framework’s effectiveness, but to assess whether the model is understandable, credible, and relevant to practitioners who design and govern high-stakes patient-facing experiences. The questions raised and reactions observed helped refine how the framework should be framed and communicated, and they served a pragmatic purpose: reducing the risk that the proposed design framework remains unused by connecting it better to the concerns, language, and way of working of the stakeholders who may apply it in practice.

### 6.6.1 Convergence on adaptive and malleable interface trajectories

Across the conversations, several D&DP members expressed that AI interfaces will become increasingly adaptive and malleable, meaning that UI composition and information density can change based on context and user preferences. This expectation aligns with emerging HCI work on generative and malleable UIs, which frames future interfaces as dynamically produced and iteratively adjustable rather than fixed and application centric (Cao et al., 2025)

Participants also noted that personalisation may occur through system-driven adaptation (the interface adjusts automatically), and user-driven configuration (users choose their preferred UI style and depth). Emotional-state-based staging was new to most, but multiple conversation partners reported that they definitely see its value, particularly in high-stakes interactions where anxiety and uncertainty strongly influence whether information is perceived as reassuring or overwhelming.

### 6.6.2 Alternative 2028 scenario: platform-controlled interfaces and reduced design control

The D&DP lead articulated an alternative future that materially shifts where design influence sits. In this scenario, large LLM platforms (e.g., Google, OpenAI, Anthropic) become the dominant front-end through which users interact, while domain services (healthcare providers, banks, insurers) are accessed “behind” the platform via APIs and authenticated data connections. The lead’s core claim is not only that LLMs will perform the reasoning, but that they will also increasingly control presentation, including interface generation and adaptation logic. Consequently, external designers would have limited influence over moderators, reasoning presentation, and front-end interaction patterns, because these would be implemented as platform capabilities.

This scenario has an important implication for Accenture Song: clients may no longer commission D&DP teams primarily to design separate domain front-ends. Instead, they may ask how they can remain visible, trusted, and relevant when the user’s primary interface is a third-party LLM platform. In such a landscape, differentiation shifts from “owning the UI” toward “owning the service logic, trust signals, and governance conditions under which the service appears and acts.”

### 6.6.3 Implication for the roadmap

Taken together, the conversations validate the roadmap's direction toward adaptive interfaces but highlight uncertainty about implementation locus (domain app vs platform front-end). This affects how Accenture Song should interpret the framework's long-term positioning. In a platform-dominated future, the framework is less useful as a set of UI prescriptions for domain apps and more useful as an interaction contract: a structured specification for how outcomes, reasons, traceability, and evidence access should be staged—regardless of the surface that renders it.

This also elevates a strategic capability for D&DP: AI architecture design. If UI composition and moderation increasingly occur within platform LLMs, the design work shifts toward architecting the end-to-end experience at system level—defining data access boundaries, evidence retrievability, escalation paths, auditability, and the “terms of explanation” under which a service can be safely represented in an LLM-mediated interaction. In other words, even if D&DP does not fully control the UI, it can still shape the experience through architecture, constraints, and governance mechanisms that determine what the LLM can do on the client's behalf.

## 6.7 Reflection on the Framework and Deliverables

The D&DP lead’s “platform-front-end” scenario raises a legitimate challenge: if a small set of LLM platforms controls both reasoning and interaction presentation, the practical influence of external design teams on moderation logic and layered explanation may be reduced. However, the thesis contribution remains meaningful under a more pluralistic view of adoption.

First, a single-interface future is not the only plausible trajectory. Platform interfaces may dominate some interaction types (quick questions, low friction queries), while dedicated applications remain relevant in other contexts—particularly where users prefer continuity, specialised workflows, regulated disclosures, or tailored service ecosystems. The likely outcome is therefore not a binary replacement but a segmented front-end landscape, where different user groups and contexts favour different interaction surfaces.

Second, even in a platform-front-end world, domain organisations still face accountability and trust requirements that cannot be fully delegated. If the primary interface is external, then the design problem shifts from “design the UI” toward “design the conditions under which the service is represented”: what data is shared, what evidence is retrievable, what escalation pathways exist, how the organisation’s decisions are justified, and how users can contest outcomes. In that sense, the framework can be interpreted not only as a UI staging model, but as a communication contract: a specification for how explanations and evidence access should be structured, regardless of whether the surface is a proprietary app or a platform-mediated interface.

Within this thesis, the roadmap should therefore be read as a directional artefact:

it describes an evolution from interface authoring to behavioural governance, while acknowledging that the “owner” of the interface may shift. The deliverables remain relevant because they provide a structured vocabulary and logic for staging understanding—either as a design method applied within domain apps, or as a set of principles and constraints that domain organisations negotiate and enforce in platform-mediated experiences.

# 7. Conclusion

This thesis set out to explore the following research question:

*How to design user-centred interfaces to enhance patients' trust and understanding in an agentic AI health assistant?*

Through qualitative and quantitative research, the study uncovered that patients' trust is shaped less by exposure to technical model details and more by the interactional quality of the explanation—whether it is reassuring, coherent, and aligned with users' immediate informational needs. The qualitative phase showed that healthcare interactions are frequently driven by worry and uncertainty, making empathetic tone, plain language, and predictable structure prerequisites for perceived trustworthiness. At the same time, transparency should be contextual and optional: users want concise “why” reasoning and actionable next steps by default, while retaining the ability to inspect underlying evidence and correct inputs when doubts arise.

The quantitative analysis further indicates that trust is the dominant driver of adoption. Across the evaluated concepts, intention to use aligned most strongly with trust, while clarity also supported willingness and ease of use played a smaller, less decisive role. These results imply that adoption fails primarily when users do not feel confident in the system's competence and accountability—not merely when the interaction is imperfectly streamlined.

At the same time, the findings show that trust is not maximised by exposing more information. Complexity can increase abstract credibility while reducing experienced trust during use through cognitive overload. Users

differ substantially in their preferred depth and degree of control, indicating that trust emerges from reassurance and interpretability as defaults, while deeper justification is most valuable as an accessible option rather than a mandatory layer.

To address this, the thesis proposes adaptable interfaces that respond to users' emotional state and informational intent. This is operationalised through the Explanatory Depth Framework, a four-layer staging model: Outcome & Next Steps, Brief Explanation, Traceability, and Deep Evidence. The framework functions as a cognitive and emotional structure for regulating when complexity is introduced while maintaining user control and contestability through two-way interaction. It is supported by a designer-facing playbook and roadmap, and demonstrated through a prototype illustrating implementable layer selection and evidence access.

Ultimately, this thesis does not claim to deliver a finished health product. Instead, it provides an evidence-informed direction for designing agentic health interfaces in which trust is fostered through outcome-first orientation, optional depth, coherent multimodal support, and mechanisms for correction—aligning explainability with the realities of patient emotion, health literacy, and risk.

## 8. Discussion

This project initially approached the challenge of trustworthy agentic health AI through a visualisation-centric lens, aiming to translate explainable AI into concrete interface representations that would make reasoning visible to patients. This premise is visible in the early framing that the main gap in current XAI work is its technical orientation and the need to translate explainability into design practices that address patient comprehension and experience.

Over the course of the project, however, the empirical work reshaped this focus. The comparative concept evaluation demonstrated that explainability is not solved by identifying a single “best” visualisation style; instead, it behaves as an interaction challenge in which pacing, information density, and user effort determine whether explanations are helpful or burdensome.

A central interpretive contribution of the study is that adoption is primarily trust-gated. Quantitative modelling showed that trust was the strongest predictor of intention to use, with clarity contributing positively, while perceived ease of use had a weaker and less reliable relationship. This indicates that patient-facing explainability must first succeed as trust calibration: users need to feel that the outcome is grounded, accountable, and safe to act on. Ease of use remains important, but mainly as a design constraint—if the interaction imposes excessive cognitive effort, it can undermine the very trust that explainability aims to build.

This helps explain the observed competence–usability tension across concepts. A more complex presentation can signal professionalism and competence at a glance, yet still reduce experienced trust in interaction when it increases information

density, interpretive effort, or perceived risk of misunderstanding. In this sense, “more transparency” is not inherently better; it can become counterproductive if it shifts burden onto the user at the moment they seek reassurance or orientation.

These findings directly motivate the Explanatory Depth Framework as a synthesis device. If trust must be supported without overwhelming users, then explanation cannot be a single static layer. Instead, trust should be initiated through outcome-first guidance and calm interpretation, sustained through optional access to a brief plain-language rationale, and protected through structured pathways to traceability and deep evidence when users seek verification, correction, or contestation.

Finally, the framework helps bridge empathic AI and interface design. Current empathy-oriented AI research predominantly evaluates empathy as a property of generated conversational responses and notes limitations such as the absence of nonverbal cues and the context-dependence of empathy (Muthukumar, 2025). Building on this, this thesis argues that empathy-relevant outcomes in patient-facing agents are not achieved through wording alone: they require UI-level adaptation that regulates when reassurance versus justification is presented, so emotional safety is supported without adding interaction cost. This is operationalised by the framework’s staged disclosure and moderator-driven defaults.

## 9. Limitations

While the results of this thesis are promising, several limitations should be acknowledged. These limitations clarify what can—and cannot—be concluded from the study and indicate where further work is needed.

Firstly, the empirical evaluation was conducted with a small sample size ( $N = 11$ ) and did not involve real patients. Although the study design allowed consistent comparison between concepts, the limited number of participants reduces statistical power and restricts how far the findings can be generalised to broader patient populations. In addition, recruiting non-patient participants means the study cannot fully represent how lived experience with illness, varying health literacy, or heightened vulnerability may shape explanation needs and trust formation in real healthcare interactions.

Secondly, the concepts were evaluated in a single, non-urgent healthcare scenario. The mild viral infection context was chosen because it is relatable for most people, which supported participant engagement and helped avoid large variation in prior knowledge. However, this choice also limits ecological validity. In more severe, ambiguous, or emotionally charged situations, users may demand different forms of reassurance, may tolerate less friction, or may require stronger evidence access and escalation pathways—potentially changing how the layers are used and perceived.

Thirdly, the evaluation used a within-subject design, meaning each participant experienced all three interface conditions. This strengthens internal comparability, since participants can directly contrast concepts, but it can also introduce learning, contrast, and fatigue effects. For example, later concepts may

be evaluated relative to earlier ones rather than independently, and participants may become more familiar with the task over time, influencing perceived ease of use and trust.

Lastly, although the thesis argues that the framework may be relevant to other high-stakes domains, it was not empirically tested in finance or insurance. These contexts involve different evidential standards, regulatory requirements, and user expectations (e.g., auditability, liability, and dispute resolution). As a result, the transferability of the framework beyond healthcare remains a reasoned proposition rather than an empirically validated claim.

# 10. Recommendations

This chapter provides recommendations derived from the thesis outcomes and the limitations of the study. The first part translates the findings into actionable steps for Accenture Song in a healthcare context. The second part outlines recommendations for design researchers and design teams who aim to design or study explainable, agentic systems in high-stakes domains.

## 10.1 Recommendations for Accenture Song

- **Shift the centre of gravity from “AI implementation” to trust-led adoption.** Beyond integrating AI capabilities into client organisations, Accenture Song should prioritise end-user adoption as the success condition. The results of this thesis suggest that adoption is primarily trust-gated: users must feel confident that the system is competent, accountable, and grounded in relevant personal signals. Success criteria should therefore include trust calibration, contestability, and clear accountability mechanisms—alongside usability—rather than implementation and technical explainability alone.
- **Design for reassurance and interpretability by default, with optional depth.** In worry-driven contexts such as healthcare, the interface should provide orientation early: what the outcome is, what it means, and what to do next, supported by a brief plain-language rationale. The default experience should reduce uncertainty without forcing evidence parsing, while preserving clear pathways to traceability and deeper evidence when users need verification or correction.
- **Implement the Explanatory Depth Framework as product logic.** Accenture Song can use the four-layer model as a concrete interaction architecture: define what each layer contains, how users move between layers, and what triggers (e.g., dispute, confusion, repeated questioning) should surface deeper levels automatically. This makes explainability operational and testable, and ensures consistency across touchpoints, channels, and teams.
- **Invest in synthetic persona testing as a capability for rapid iteration.** Accenture Song should expand methodological capability in synthetic persona testing and AI-augmented evaluation. While synthetic personas cannot replace real users in high-stakes validation, they can accelerate early-stage concept exploration, edge-case discovery, and iteration cycles—reducing time spent on low-value revisions and supporting faster convergence before involving costly stakeholder sessions.
- **Strengthen AI architecture and governance expertise as a design constraint, not an afterthought.** For agentic systems, design decisions are inseparable from governance choices (e.g., escalation rules, audit trails, data provenance, consent boundaries, and accountability). Investing in architecture and governance knowledge will allow teams to design “Deep Evidence” and verification features responsibly, align with regulatory requirements, and prevent trust-breaking failures at deployment.

## 10.2 Recommendations for Design Researchers and Design Teams

1. Study learning effects on trust calibration, friction, and information needs.

The current results suggest trust is the primary adoption gatekeeper, but both trust formation and perceived effort may shift as users become familiar with agentic systems. Longitudinal research is needed to test whether growing competence changes (a) how much explanation is needed to establish trust, (b) the tolerance for interaction cost, and (c) preferences for traceability and evidence access.

2. Investigate “human-in-the-loop” interaction patterns for agentic care.

Further research should define how responsibility, consent, and escalation should be handled when agents take action on behalf of users. This includes design questions such as when an agent may suggest versus automatically book an appointment, how clinician choice should be presented, what confirmations are required, and how accountability and audit trails should be communicated to users and professionals.

3. Validate transferability across domains and evolving AI platform ecosystems.

The framework should be tested in other high-stakes contexts such as finance and insurance, where evidential standards, dispute processes, and regulatory constraints differ. In parallel, research should anticipate platform shifts in which large general-purpose LLM interfaces become the primary user gateway. This raises open questions about how smaller service providers can remain visible and trusted, how explanation and evidence should be routed across multi-provider systems, and what API and governance architectures are needed to ensure provenance, accountability, and consistent user experience across an ecosystem.

# Reflection

The thesis process was complex, but it also became a steep learning journey for me. At the start, I did not know much about the subject, and I needed time to build a solid understanding of the topic and its context. My supervisor Feline played an important role in that learning curve. She taught me a lot, and her feedback helped me develop my knowledge and think more critically about what I was doing and why.

One of the main challenges in the beginning was defining the scope of the project. I struggled to decide what my focus should be and how broad or narrow the research needed to be. This made it harder to move forward with confidence. Once I identified a clear gap between XAI and UX/UI design, the project started to “click.” I was able to shape the direction more clearly, make stronger decisions, and work with consistent momentum—taking concrete steps every day and week.

A part I genuinely enjoyed was exploring how AI could support the design process. I liked discovering where AI could accelerate work and where the process needed to remain human-led. At the same time, I learned that the design process requires not only generating ideas quickly, but also making trade-offs explicit. Especially when using AI during ideation, I had to consciously separate “many possible options” from “the best option for this context.” In the beginning, it was difficult for me to stay critical towards AI output, mainly because my understanding of the topic and context was not sufficient enough to challenge what the AI produced. As my knowledge improved, I became better at evaluating AI suggestions, questioning them, and using AI as a tool rather than treating its output as an answer. Over time,

I also improved at documenting my design rationale—why I selected certain patterns and rejected others—so that my supervisors could better follow my reasoning and see how feedback was incorporated.

Throughout the project, I learned more about myself as a worker. I realised that I work fast and usually know quickly what steps to take next. I planned well and that helped me reach my deadlines. However, I also learned something important about where my strengths and development areas lie. I learned that my main strength lies in the first diamond: quickly identifying the core problem and understanding what users need. The second diamond is where I can improve most—translating those insights into refined design decisions through stronger iteration and development.

Another important lesson was about communication and collaboration. I learned that I could be more transparent towards my supervisors about the steps I take and the decisions I make during the process. Sometimes it looked like I did not incorporate their feedback, simply because I did not clearly mention what I changed or how I applied their suggestions. Being clearer about this would have improved alignment and made the collaboration even stronger.

Finally, I am grateful for all the feedback I received from colleagues, friends, and family. Their input helped me stay motivated and improve the quality of the work.

# Reference List

- Adeniran, A. A., Onebunne, A. P., & William, P. (2024). Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. *World Journal of Advanced Research and Reviews*, 23\*(3), 2647–2658.
- Baumgart, D. C. (2024). An intriguing vision for transatlantic collaborative health data use and artificial intelligence development. *npj Digital Medicine*, 7\*(19). <https://doi.org/10.1038/s41746-024-01005-y>
- Boston Consulting Group. (2025). *How digital and AI will reshape health care in 2025*.
- Cao, Y., Jiang, P., & Xia, H. (2025). Generative and malleable user interfaces with generative and evolving task-driven data model. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM. <https://doi.org/10.1145/3706598.3713285>
- Corti, L., Oltmans, R., Jung, J., Balayn, A., Wijzenbeek, M., & Yang, J. (2024). “It is a moving process”: Understanding the evolution of explainability needs of clinicians in pulmonary medicine. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (pp. 1–21). ACM. <https://doi.org/10.1145/3613904.3642551>
- David, W. (2024). Interoperability and standardization challenges in health information exchange. *Journal of Health and Medical Informatics*, 15\*, 515.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Demiris, G., Iribarren, S. J., Sward, K., Lee, S., & Yang, R. (2019). Patient generated health data use in clinical practice: A systematic review. *Nursing Outlook*, 67(4), 311–330. <https://doi.org/10.1016/j.outlook.2019.04.005>
- Dingel J., Kleine A., Cecil J., Sigl A., Lerner E., Gaube S. (2024) Predictors of Health Care Practitioners’ Intention to Use AI-Enabled Clinical Decision Support Systems: Meta-Analysis Based on the Unified Theory of Acceptance and Use of Technology *J Med Internet Res* 2024;26:e57224 URL: <https://www.jmir.org/2024/1/e57224> DOI: 10.2196/57224
- Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing transparency design into practice. In *Proceedings of the 2018 International Conference on Intelligent User Interfaces (IUI '18)*. ACM. <https://doi.org/10.1145/3172944.3172961>
- Fang, H. S. A., Tan, T. H., Tan, Y. F. C., & Tan, C. J. M. (2021). Blockchain personal health records: Systematic review. *Journal of Medical Internet Research*, 23\*(4), e25094. <https://doi.org/10.2196/25094>
- Fraser, H., Crossland, D., Bacher, I., Ranney, M., Madsen, T., & Hilliard, R. (2023). Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: Clinical data analysis study. *JMIR mHealth and uHealth*, 11\*(1), e49995.
- Gartner, Inc. (2025). *Quick answer: How will agentic AI transform healthcare provider operations?* (G00830680).
- Goisauf, M., Cano Abadía, M., Akyüz, K., Bobowicz, M., Buyx, A., Colussi, I., Fritzsche, M.-C., Lekadir, K., Marttinen, P., Mayrhofer, M. T., & Meszaros, J. (2025). Trust, trustworthiness, and the future of medical AI: Outcomes of an interdisciplinary expert workshop. *Journal of Medical Internet Research*, 27\*, e71236. <https://doi.org/10.2196/71236>

- Haddad, A., Habaebi, M. H., Suliman, F. E. M., Elsheikh, E. A. A., Islam, M. R., & Zabidi, S. A. (2023). Generic patient-centered blockchain-based EHR management system. *\*Applied Sciences, 13\*(3), 1761.* <https://doi.org/10.3390/app13031761>
- Heydari, A. A., Gu, K., Srinivas, V., Yu, H., Zhang, Z., Zhang, Y., Paruchuri, A., He, Q., Palangi, H., Hammerquist, N., Metwally, A. A., Winslow, B., Kim, Y., Ayush, K., Yang, Y., Narayanswamy, G., Xu, M. A., Garrison, J., Lee, A. A., ... Xu, X. (2025, August 27). *\*The anatomy of a personal health agent.\** arXiv. <https://arxiv.org/abs/2508.20148>
- Hildt, E. (2025). What is the role of explainability in medical artificial intelligence? A case-based approach. *\*Bioengineering, 12\*(4), 375.* <https://doi.org/10.3390/bioengineering12040375>
- Hulsen, T. (2023). Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. *\*AI, 4\*(3), 652–666.* <https://doi.org/10.3390/ai4030034>
- JCT CoatingsTech. (2007). *Color Psychology.* Ipswich, MA: JCT CoatingsTech.
- LaBoone, P. A., & Marques, O. (2024). Overview of the future impact of wearables and artificial intelligence in healthcare workflows and technology. *\*International Journal of Information Management Data Insights, 4\*(2), 100294.* <https://doi.org/10.1016/j.ijime.2024.100294>
- Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., ... Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *\*BMJ, 388\*, e081554.* <https://doi.org/10.1136/bmj-2024-081554>
- Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *\*Computer Methods and Programs in Biomedicine, 226\*, 107161.*
- McGrath, M. J., Lack, O., Tisch, J., & Duenser, A. (2025). Measuring trust in artificial intelligence: Validation of an established scale and its short form. *Frontiers in Artificial Intelligence, 8, Article 1582880.* <https://doi.org/10.3389/frai.2025.1582880>
- McKinsey & Company. (2020). *\*Transforming healthcare with AI.\** EIT Health and McKinsey & Company.
- McKinsey & Company. (2024). *\*Why agents are the next frontier of generative AI.\**
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *\*Artificial Intelligence, 267\*, 1–38.*
- Mirchev, M., Mircheva, I., & Kerekovska, A. (2020). The academic viewpoint on patient data ownership in the context of big data: Scoping review. *\*Journal of Medical Internet Research, 22\*(8), e22214.* <https://doi.org/10.2196/22214>
- Mohammed, S., & Malhotra, N. (2025). Ethical and regulatory challenges in machine learning-based healthcare systems: A review of implementation barriers and future directions. *\*BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 5\*, 100215.* <https://doi.org/10.1016/j.tbench.2025.100215>
- Mollerus, F., Lynch, C., & Bruining, H. (2025). Data interoperability for a systems approach to developmental conditions. *\*Neuroscience and Biobehavioral Reviews, 176\*, 106245.* <https://doi.org/10.1016/j.neubiorev.2025.106245>
- Muthukumar, K. (2025). Empathy AI in healthcare. *Frontiers in Psychology, 16, 1680552.* <https://doi.org/10.3389/fpsyg.2025.1680552>

- Nong, P., & Ji, M. (2025). Expectations of healthcare AI and the role of trust: Understanding patient views on how AI will impact cost, access, and patient-provider relationships. *Journal of the American Medical Informatics Association*, 32\*(5), 795. <https://doi.org/10.1093/jamia/ocaf031>
- Nong, P., & Platt, J. (2025). Patients' trust in health systems to use artificial intelligence. *JAMA Network Open*, 8\*(2), e2460628. <https://doi.org/10.1001/jamanetworkopen.2024.60628>
- Philips. (2025). *Building trust in healthcare AI: Perspectives from patients and professionals – Global report.* <https://www.philips.com/c-dam/corporate/newscenter/global/future-health-index/report-pages/experience-transformation/2025/philips-future-health-index-2025-report-building-trust-in-healthcare-ai-global.pdf>
- Roberts, J. S., Robinson, J. O., Diamond, P. M., Bharadwaj, A., Christensen, K. D., Lee, K. B., Green, R. C., & McGuire, A. L. (2018). Patient understanding of, satisfaction with, and perceived utility of whole genome sequencing: Findings from the MedSeq Project. *Genetics in Medicine*, 20(9), 1069–1076. <https://doi.org/10.1038/gim.2017.223>
- Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI*, 3\*, e53207. <https://doi.org/10.2196/53207>
- Sadaka, F., & Qureshi, R. (2025). Ensuring trust in healthcare robotics: The essential role of explainable AI. In *Explainable AI in healthcare imaging for medical diagnoses* (pp. 1–28). Elsevier. <https://doi.org/10.1016/B978-0-443-23979-3.00001-4>
- Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhalwaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladik, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118\*, 109370. <https://doi.org/10.1016/j.compeleceng.2024.109370>
- Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263\*, 110273.
- Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Information Fusion*, 126\*, 103599. <https://doi.org/10.1016/j.inffus.2025.103599>
- Schrepp, M., Otten, R., Blum, K., & Thomaschewski, J. (2020). What causes the dependency between perceived aesthetics and perceived usability? *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(6), 78–85. <https://doi.org/10.9781/ijimai.2020.12.005>
- Shapiro, M., Johnston, D., Wald, J., & Mon, D. (2012). Patient-generated health data: White paper. RTI International (Prepared for the Office of Policy and Planning, Office of the National Coordinator for Health Information Technology).
- Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., and Elmqvist, N., *Designing the User Interface: Strategies for Effective Human-Computer Interaction: Sixth Edition*, Pearson (May 2016) <http://www.cs.umd.edu/hcil/DTUI6>
- Starke, G., Gille, F., Termine, A., Aquino, Y. S. J., Chavarriaga, R., Ferrario, A., Hastings, J., Jongasma, K., Kellmeyer, P., Kulynych, B., Postan, E., Racine, E., Sahin, D., Tomaszewska, P., Vold, K., Webb, J., Facchini, A., & Ienca, M. (2025). Finding consensus on trust in AI in health care: Recommendations from a panel of international experts. *Journal of Medical Internet Research*, 27\*, e56306. <https://doi.org/10.2196/56306>

- Steerling, E., Siira, E., Nilsen, P., Svedberg, P., & Nygren, J. (2023). Implementing AI in healthcare—the relevance of trust: A scoping review. *Frontiers in Health Services, 3*, 1211150. <https://doi.org/10.3389/frhs.2023.1211150>
- Taleb, H., Andrieux, G., Khalife, D., Ajami, A., & Nasser, A. (2025). Exploring healthcare in the 6G and AI era: Opportunities and challenges. *Internet of Things, 34*, 101744.
- Wallston, K. A., Cawthon, C., McNaughton, C. D., Rothman, R. L., Osborn, C. Y., & Kripalani, S. (2014). Psychometric properties of the Brief Health Literacy Screen in clinical practice. *Journal of General Internal Medicine, 29*(1), 119–126. <https://doi.org/10.1007/s11606-013-2563-0>
- Wang, Y., Wan, Y., Lei, X., Chen, Q., & Hu, H. (2025). A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models. *Array, 28*, 100504. <https://doi.org/10.1016/j.array.2025.100504>