


SpringerBriefs in Applied Sciences and Technology

PoliMI SpringerBriefs

**Davide Spallazzo · Martina Sciannamè ·
Mauro Ceconello**



User Experience + Artificial Intelligence

Assessing the Qualities
of AI-infused Systems



POLITECNICO
MILANO 1863

OPEN ACCESS

 **Springer**

SpringerBriefs in Applied Sciences and Technology

PoliMI SpringerBriefs

Series Editors

Barbara Pernici, DEIB, Politecnico di Milano, Milano, Italy

Stefano Della Torre, DABC, Politecnico di Milano, Milano, Italy

Bianca M. Colosimo, DMEC, Politecnico di Milano, Milano, Italy

Tiziano Faravelli, DCHEM, Politecnico di Milano, Milano, Italy

Roberto Paolucci, DICA, Politecnico di Milano, Milano, Italy

Silvia Piardi, Design, Politecnico di Milano, Milano, Italy

Gabriele Pasqui , DASTU, Politecnico di Milano, Milano, Italy

Springer, in cooperation with Politecnico di Milano, publishes the PoliMI Springer-Briefs, concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 (150 as a maximum) pages, the series covers a range of contents from professional to academic in the following research areas carried out at Politecnico:

- Aerospace Engineering
- Bioengineering
- Electrical Engineering
- Energy and Nuclear Science and Technology
- Environmental and Infrastructure Engineering
- Industrial Chemistry and Chemical Engineering
- Information Technology
- Management, Economics and Industrial Engineering
- Materials Engineering
- Mathematical Models and Methods in Engineering
- Mechanical Engineering
- Structural Seismic and Geotechnical Engineering
- Built Environment and Construction Engineering
- Physics
- Design and Technologies
- Urban Planning, Design, and Policy

<http://www.polimi.it>

Davide Spallazzo · Martina Sciannamè ·
Mauro Ceconello


User Experience + Artificial Intelligence

Assessing the Qualities of AI-infused Systems



POLITECNICO
MILANO 1863

 Springer

Davide Spallazzo 
Department of Design
Politecnico di Milano
Milan, Italy

Martina Sciannamè
Department of Design
Politecnico di Milano
Milan, Italy

Mauro Ceconello
Department of Design
Politecnico di Milano
Milan, Italy



ISSN 2191-530X ISSN 2191-5318 (electronic)
SpringerBriefs in Applied Sciences and Technology
ISSN 2282-2577 ISSN 2282-2585 (electronic)
PoliMI SpringerBriefs
ISBN 978-3-031-77520-8 ISBN 978-3-031-77521-5 (eBook)
<https://doi.org/10.1007/978-3-031-77521-5>

This work was supported by Politecnico di Milano.

© The Editor(s) (if applicable) and The Author(s) 2025. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The rapid integration of artificial intelligence (AI) into everyday products has brought both excitement and challenges, transforming the ways in which users interact with technology. From smart home devices to self-driving vehicles, AI-infused products promise unprecedented levels of autonomy, adaptability, and personalisation. Yet, despite the sophistication of the underlying technologies, the user experience (UX) of these products often lags behind, revealing significant gaps in interaction quality, usability, and meaningful engagement.

This essay emerges from the extensive work conducted during the Meet-AI research project, funded by the Department of Design at Politecnico di Milano. As a pioneering effort, this project set out to systematically address the UX challenges posed by AI-infused products, recognising that the conventional tools and methods for UX assessment are inadequate for the unique complexities of these dynamic and evolving systems. In a field where rapid technological advancements have often outpaced human-centred design considerations, this project seeks to restore balance by focusing on how users experience and interact with AI-infused systems, trying to understand what they value the most in this new setting.

To the best of the authors' knowledge, the Meet-AI project is the first to take a comprehensive look at AI-infused products from a user-centred design perspective. The research team, comprised of eight experts from the institution, delved deeply into the design, usability, and interaction challenges that AI introduces, with the ultimate goal of creating and validating a new method for assessing the UX of AI-based products. This is meant to be a valuable resource for researchers, designers, and companies to evaluate the kind of products under investigation and possibly improve them so that they are impactful and meaningful to users.

This essay reflects the outcomes of that research and presents three key contributions. First, it offers a thorough examination of the UX challenges associated with AI-infused products, helping to understand the gap between technological capabilities and user expectations. Second, it introduces a set of new UX dimensions, including *trustworthiness*, *intelligence*, *meaningfulness*, and *conversational* qualities, which are crucial for evaluating AI-driven systems. Lastly, it presents the AIXE

scale, a quantitative, statistically validated tool designed to holistically evaluate the UX of AI-infused products.

As AI continues to transform the way we live and interact with technology, it is imperative that designers and researchers work together to ensure that these products not only function but also provide satisfying, seamless, and meaningful user experiences. The insights presented in this essay are meant to fuel that conversation, bridging the gap between technological advancement and UX design to create a future where AI-infused products serve their users in more human-centred and impactful ways.

This work also recognises the multidisciplinary nature of the conversation around AI and UX. Hence, it aims to engage a broad audience, including professionals across various fields such as ethics, psychology, social sciences, and jurisprudence, all of whom play a role in shaping the future of AI-infused products.

By bringing together insights from design research and AI technology, this essay seeks to contribute meaningfully to ongoing discourse, providing actionable knowledge and practical tools for those developing the next generation of AI-based artefacts. In doing so, it paves the way for further research, experimentation, and collaboration that will ensure AI-infused products not only innovate but also enrich and empower the human experience.

Milan, Italy

Davide Spallazzo
Martina Sciannamè
Mauro Ceconello

Acknowledgements

The Meet-AI project was funded by the Department of Design at Politecnico di Milano through the “FARB-Fondo alla Ricerca di Base” grant.

The project was carried out by a dedicated research team consisting of Marco Ajovalasit, Venanzio Arquilla, Mauro Ceconello, Giuseppe Fazio, Martina Scianamè, Ilaria Vitali, Francesco Zurlo, and coordinated by Davide Spallazzo. We extend our gratitude to the researchers not listed as authors of this book for their valuable contributions.

We would also like to express our sincere thanks to Emma Zavarrone for her invaluable assistance in the statistical validation process, and to Anton Krassa for his support with the systematic literature review.

Our appreciation goes to Demetra Opinioni for their help in administering the questionnaires, and we are especially grateful to the anonymous respondents whose participation helped validate the AIXE scale.

Contents

1	Introduction	1
1.1	Rationale of the Essay	1
1.2	Essay Structure	4
2	Making Sense of AI-Infused Systems. Framing Current Design Challenges	7
2.1	Introduction	7
2.2	The Nature of AI-Infused Products	9
2.2.1	Baseline Definitions	9
2.2.2	The Characterizing Features of AI-Infused Products	10
2.3	UX Challenges Posed by AI-Infused Products	13
2.3.1	Consequences of Their Peculiarities	13
2.3.2	Communication-based Issues	14
2.3.3	Repercussions on the Design Process	15
2.3.4	Ethical and Societal Impacts	16
2.4	New Interaction Paradigms	16
2.4.1	Design for Something Else	16
2.4.2	Reimagining Interfaces	17
2.4.3	New Forms of Assistance	18
2.5	Conclusion	19
	References	20
3	UX Dimensions for AI. Past and Future Perspectives	25
3.1	Introduction	25
3.2	The Evolution of UX Assessment	26
3.3	Methodology	28
3.4	Results	30
3.4.1	Umbrella Review	30
3.4.2	UX Evaluation Methods Scoping Review	34
3.4.3	AI-Infused Systems Systematic Review	38

- 3.5 Discussion and Findings 40
 - 3.5.1 RQ1—Are Current UX Assessment Methods Enough for AI-Infused Products? 40
 - 3.5.2 RQ2—Are New UX Dimensions Needed for These Products? 41
 - 3.5.3 RQ3—What Characteristics Should the New Method Have? 42
- 3.6 Conclusion 45
- References 45
- 4 Unpacking AI-Infused Systems Qualities. Building a UX Evaluation Method 49**
 - 4.1 Actively Exploring the Qualities of AI-Infused Products. Overview of the Research Methods 49
 - 4.2 Phase 0: Broadening the Boundaries of AI-Related Qualities Through a Survey 50
 - 4.3 Phase 1: UX Dimensions of AI-Infused Products According to Advanced Users 51
 - 4.4 Phase 2: Insights from an Intertwined Analysis of AI-Related Descriptors 54
 - 4.5 Phase 3: A Research Workshop to Systematise the Findings 61
 - 4.6 Conclusions, Limitations, and Further Actions 62
 - References 63
- 5 AIXE. A Method to Evaluate the UX of Systems Integrating AI 65**
 - 5.1 Introduction 65
 - 5.2 Methodology 67
 - 5.2.1 Items and Questionnaire Elaboration 67
 - 5.2.2 Statistical Validation 68
 - 5.3 Results 69
 - 5.3.1 Items and Questionnaire Generation 69
 - 5.3.2 Statistical Validation 74
 - 5.4 Discussion 77
 - 5.4.1 Reflecting on the Results 77
 - 5.4.2 Strengths and Limitations 79
 - 5.5 Application of the AIXE Scale 80
 - 5.5.1 Setting 80
 - 5.6 Conclusion and Future Work 80
 - References 81
- 6 Applying AIXE to Compare Domestic Smart Speakers 83**
 - 6.1 Introduction 83
 - 6.1.1 Sample of Respondents and Methodology 84
 - 6.2 How Smart Speakers Performed 85
 - 6.2.1 The Big Picture 85
 - 6.2.2 Performances Over Time: 2021 Versus 2023 86
 - 6.2.3 Performances in the UK 88

- 6.2.4 Performances in the USA 90
- 6.2.5 UX Results by Age Group 93
- 6.2.6 UX Results by Device 94
- 6.3 Discussion 96
- 6.4 Conclusion 98
 - 6.4.1 Limitations and Future Research 99
- References 99
- 7 Conclusions 101**
 - 7.1 Summarizing the Contribution 101
 - 7.1.1 Contribution to Design and UX Assessment 101
 - 7.1.2 Implications for AI-Infused Products Design 102
 - 7.2 Strengths, Limitations and Future Opportunities 103
 - 7.2.1 A Broader Access 103
 - 7.2.2 Transcending Conventions to Embrace Evolution
and the Whole Design Process 104
 - 7.2.3 A Multidimensional and Multi-method Approach 105
 - 7.2.4 A Broader Range of AI-Infused Products 106

Chapter 1

Introduction



Abstract This chapter introduces the essay, by explaining its purpose and significance, and clarifying its background. It further presents the overall methodology, specifying the aims of the essay, the target audience and the expected impacts.

1.1 Rationale of the Essay

This essay is grounded in the extensive work conducted as part of the Meet-AI research project. It was funded by the Department of Design at Politecnico di Milano as a timely initiative to address the urgent need to explore the intricacies of products integrating artificial intelligence (AI) from a user experience (UX) design perspective. The project brought together a team of eight researchers from the institution, whose expertise in product, interaction, and UX design enabled a human-centred deepening and understanding of the topic. Specifically, the project goal was to investigate and develop novel methods for assessing the user experience of these game-changing interactive products, providing researchers, designers, and companies with tools to create, evaluate, and deploy products that are both meaningful and impactful.

To the best of our knowledge, the Meet-AI project stands as the first systematic effort to address the peculiar challenges that AI systems pose to the UX of the products in which they are embedded. Hence, the essay brings an original contribution to the ever-growing discourse around AI.

The research project stems from the critical observation of a phenomenon that is unfolding in a contradictory manner. A new and complex category of interactive systems that blend advanced AI functionalities with everyday user tasks is rapidly expanding, and it is transforming how users interact with technology. From voice assistants to smart home devices, AI-infused products have quickly captured the interest of the public and sparked hype towards unprecedented promises of autonomy and adaptability. Also from a UX standpoint, the unique traits of these artefacts seemed to cater exciting and transformative possibilities. Yet, while the technology is indeed sophisticated, the UX of these products has not evolved at the same pace as the algorithms. Soon, the shortcomings in this realm became patent and the expectations

built around AI-infused products were failed. In fact, these often result in opaque and confusing interfaces that betray basic usability and interaction quality standards, which are cornerstones of good UX design. This discrepancy has led to widespread difficulties in delivering the seamless, intuitive, and satisfying interactions that users expect from modern technology.

This gap between technological advancement and user experience is not unusual. Historically, the early stages of technological revolutions have often been entrusted to technologists, who primarily focus on their domain of expertise and aim at showcasing their new discoveries. In the case of AI, technologists succeeded in developing sophisticated systems that present unprecedented capabilities, like the possibility to entertain colloquial conversations with machines, and managed to bring these advancements into commercial products, such as smart speakers. Nonetheless, they often lack the human-centred focus necessary to create products that meet user needs. As a result, AI-infused products were released with significant UX flaws, as the challenges to design for complex, unpredictable, and evolving entities have not been adequately addressed. Meanwhile, the design and HCI fields lagged behind the swift technological changes driven by AI and are currently trying to react, in the attempt of being increasingly involved in the decisional and development process of AI-infused products. Indeed, they have the potential to bring an essential human-centred perspective to the discourse, but only recently the effort put in design research and practice to address this new design material has started to provide some answers to the several questions and reflections related to the necessary interventions and possible approaches to a meaningful development of AI-based systems.

Undoubtedly, the probabilistic and evolving nature of this new category of products presents unique challenges to the UX, especially because their users generally lack a proper understanding of these systems and their implications. Indeed, they often face opaque and confusing interfaces, which can modify their behaviour over time, might spread across multiple touchpoints and raise ethical concerns at different levels.

It becomes increasingly clear that traditional UX evaluation methods—which were built for more static and predictable systems—might be insufficient and ineffective at targeting such dynamic and complex issues.

To tackle this challenge, the Meet-AI project employed a comprehensive multi-method research strategy, trying to cope with the difficulties posed by an emerging topic—with a consequent limitation of available resources—and dealing with the risk of subjectivity hindering the results.

The research effort built on a rigorous review of existing UX assessment methods—including umbrella, scoping, and systematic reviews—to understand their applicability and limitations when applied to AI-infused products. As a result, UX dimensions consistent with the assessment of AI-based artefacts have been derived and further investigated with a user-centred approach. Indeed, expert users were involved in a later stage of the study to further explore and validate the identification of relevant UX dimensions.

Building on the obtained results, and through a combination of qualitative and quantitative methods, the research team was able to delineate and statistically validate

a quantitative UX assessment scale tailored to the necessities of AI-infused products: AIXE (AI user eXperience Evaluation). It was ultimately outlined in the form of a comprehensive questionnaire, that provides a standardized tool to evaluate the UX of AI products holistically, considering the unique challenges and complexities they introduce.

The AIXE scale was then applied in a real-world scenario, to assess the UX of the most common materialization of AI systems in the domestic realm: smart speakers. These products serve as prime examples of how AI can pervade everyday life and highlight interaction and usability difficulties.

By portraying the outcomes of the Meet-AI project, this essay aims to make a significant contribution to the ongoing discourse surrounding AI and UX, and its relevance unfolds at different levels.

First, it sheds light on the challenges posed by AI-infused products from a UX perspective. It offers a deep exploration of their nature and the issues that arise both during users' interactions and the design process.

Second, it highlights and explains essential UX dimensions that are crucial for understanding and evaluating AI-based artefacts, also introducing new ones.

And lastly, it provides the AIXE scale, a robust tool for practitioners and researchers to assess the UX of AI products in a holistic and systematic manner.

These contributions are intended to serve a diverse range of audiences. In the essay, researchers in HCI and UX can find useful resources to study AI-infused products, including a novel tool and a set of UX dimensions and more granular qualities. Designers are offered insights that can support them in the evaluation of advanced and working prototypes, but also provide foundational and actionable knowledge to better understand and anticipate how AI-infused systems can behave and which potentialities they can exploit. Finally, companies may benefit from the possibility to more properly assess and benchmark their products over time or to compare them with the competition, employing the AIXE scale. As well, the main insights of the project can inform them about key areas for improvement in user interaction.

Moreover, as the topic is positioned at the intersection of multiple disciplines—not only computer science and design, but also ethics, psychology, social sciences, and jurisprudence—the contribution of the essay can extend to additional areas of research and practice, engaging professionals that deal with such kind of products, to fuel a multidimensional and multidisciplinary conversation.

In conclusion, this essay seeks to bridge the gap between technological innovation and UX design, to ensure that AI-infused products are not only innovative and powerful, but also user-friendly and meaningful.

1.2 Essay Structure

The essay structure follows the rationale and unfolding of the Meet-AI research project, aimed at addressing the growing complexities and challenges in assessing the user experience of AI-infused products.

Accordingly, Chap. 2 explores the core qualities and peculiarities of AI-infused products demanding new interaction paradigms, which pose significant challenges to their design and the methods traditionally used to evaluate UX. This chapter sets the foundation for the essay by establishing the need to rethink conventional UX evaluation approaches for AI-based systems.

To address this, Chap. 3 examines the state-of-the-art of UX evaluation methods, identifying gaps and insufficient approaches in capturing the nuances of AI-infused products. It also explores the developments in the assessment of AI-based systems, finding a still emergent field that tends to be more technology than UX-oriented. From the insights collected after umbrella, scoping, and systematic literature reviews, the chapter highlights eight essential UX dimensions, offering a starting point for more focused experimental research. The *pragmatic*, *aesthetic*, *hedonic*, *affective*, *intelligence*, *trustworthiness*, *conversational*, and *meaningfulness* dimensions, are proposed as critical elements in understanding the user experience of AI-infused products, laying the groundwork for further exploration.

Building on this foundation, Chap. 4 systematises the qualities of AI-infused products and proposes a user-centred approach to the identification of the building blocks for a new UX assessment method. Indeed, the chapter presents the findings from a study involving advanced users, who were engaged to detect the most significant UX qualities for AI-based products. The investigation process unfolded in different stages, and the outcomes emphasised how the dimensions that better addressed the unique traits of AI systems (*intelligence*, *trustworthiness*, *conversational*, and *meaningfulness*) also resonated the most with users. While still relevant, more traditional dimensions like the *pragmatic*, *aesthetic*, *hedonic*, and *affective* ones produced more modest results.

Chapter 5 introduces the AIXE (AI user eXperience Evaluation) scale, the statistically validated tool that, stemming from the previous research phases, has been specifically designed to assess the UX of AI-infused products. It details the development and validation of a 33-question scale, using a Likert-scale approach to evaluate 12 descriptors that cover 6 UX dimensions. This chapter crucially illustrates how the formalised method for assessing the UX of AI-infused products has been constructed and can be employed, offering a clear, practical framework for researchers and designers to quantify user experience in a field that is still lacking standardised tools.

Chapter 6 further dives into the application of the AIXE scale in a real-world context. A comparative study has been conducted to evaluate the UX of smart speakers, arguably the most pervasive and emblematic materialization of AI systems. With data from over 1600 respondents from the US and UK, this Chap. provides an in-depth analysis of how different smart speaker devices perform across the identified

UX dimensions. This study offers insights into user perceptions, revealing patterns and trends in the user experience.

Finally, Chap. 7 reflects on the main contributions of the essay and the Meet-AI project as a whole, summarising key insights while acknowledging the limitations of the research. It also provides guidance for future research directions, suggesting pathways to further refine the UX evaluation for AI-infused products.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 2

Making Sense of AI-Infused Systems. Framing Current Design Challenges



Abstract The chapter portrays the state of the art of AI-infused products, that while offering exciting novelties in terms of user experience, offer present difficulties to exploit their full potential. The core and peculiar qualities of AI-infused products are primarily explored. The chapter further frames the current and unique challenges they present to their design and new interaction paradigms they introduce in the user experience. The chapter finally questions the necessity of a rethinking of traditional UX evaluation methods to accommodate these emerging design challenges.

2.1 Introduction

Continuous technological developments in the computing power and storage capacity of modern computers have made possible the deployment of systems that for a long time were only an ambitious dream. Artificial intelligence (AI) is not a novelty, but the successful implementation of machine learning (ML), and their subset deep learning (DL), systems is quite revolutionary.

Already in 2020, Burr et al. [1] observed how AI and ML systems spread in different human domains. Their study highlighted four main ones: healthcare, education and employment, governance and social development, media and entertainment. However, the phenomenon has grown since then, and AI systems are expanding in several additional endeavours, from transportation, to industrial development, judicial system, house management, personal care, work efficiency, and even sustainability [2].

This boundless spread is also testified by people's perceptions around the world. As reported by 2024 AI Index Report, 66% of the respondents think that AI will dramatically affect their lives in the next three to five years [3].

Nevertheless, despite the abundance of examples of AI-infused products and services that surround us, research and experimentations are needed, especially in the fields of design and HCI. Indeed, the thorough work conducted by Yildirim et al. [4], mostly from Carnegie Mellon University, underlined how more than 85% of projects intended to foster innovation through AI fail. The main reason can be attributed to

a lack of human-centred design, resulting in products that are not able to address people's actual needs. In fact, professionals from the design or HCI field, if involved, often enter the discourse when the problem to tackle has already been identified and they do not have enough preparation about what AI can reasonably do. Because of this, they tend to envision solutions that are overly complex and not feasible, missing the opportunity to improve the UX by just implementing simple AI systems.

Undoubtedly, the organization related to the design process of AI-infused products and an adequate preparation of the professionals dealing with it are urgent and foundational issues. The study presented in this book aims to support the development of solutions for these root problems by focusing on the UX of such artefacts. Indeed, this would allow fast interventions in this rapidly evolving endeavour, setting the premises for deeper changes.

A starting point for investigating the UX of AI-infused products is to get a deep understanding of their qualities and hindrances. Therefore, this chapter aims to provide a comprehensive overview of the characteristics that define AI-infused products and the UX challenges they present. It begins by exploring the nature of these products, clarifying their definition and identifying the features that set them apart from traditional systems. AI-infused products are designed to tackle uncertain, ill-defined problems, where only the goals are known but the exact path to achieve them is not. This capability makes them highly adaptable, learning from experience, yet also prone to flaws due to their probabilistic nature. Another key quality is their characterization as sociotechnical systems, implying the central role of people in their development and end goals.

The chapter then delves into the UX challenges posed by AI-infused products, grouping them into four major categories. First are the challenges that stem directly from their core characteristics, such as unpredictability, ergonomic issues and the agency and proactivity that AI systems may exhibit. Second, communication-related challenges make AI systems opaque, leading to misunderstandings and misalignments between user expectations and system behaviours. Unclear or insufficient communication can result in poor-quality interactions and suboptimal interfaces, further complicating the UX. Third, AI-infused products require designers to adjust the design process itself, as they are a new and evolving material for designers to work with. Lastly, ethical and societal concerns play a significant role in shaping the design of AI-infused products.

In addition to outlining these challenges, the chapter also examines the new interaction paradigms introduced by AI-infused products. These paradigms demand that designers adapt their approaches, considering ecosystems and human values as integral parts of the design process. Designers must view AI as a counterpart agent, embracing the inherent imperfection of these systems and allowing for flexible, adaptive designs. New interface typologies, such as conversational and gesture-based interactions, are emerging alongside AI's ability to assist with decision-making, personalized recommendations, and content generation.

Finally, the chapter concludes by questioning the need for new UX evaluation methods that address the unique and complex qualities of AI-infused products.

2.2 The Nature of AI-Infused Products

2.2.1 *Baseline Definitions*

Before diving into the specificities of AI-infused products, it is important to clarify what we mean with this term.

Indeed, a lot of ambiguity revolves around this topic and a possible motivation is the lack of agreement in the AI field itself. From its inception, at Dartmouth College, the concept of AI was linked to the objective of making

machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves [5].

Hence, its definition depended on a fuzzy reference to human capabilities, leaving multiple possibilities of interpretation. In their seminal textbook, Russell and Norvig [6], effectively synthesize these human-related denotations in two main typologies. The first, called “Turing test approach”, describes AI systems as capable to *act* like people, for instance, communicating by processing natural language (NLP), storing information (Knowledge representation), making inferences (automated reasoning), or adapting to new circumstances and recognizing patterns (ML systems). The second, the “cognitive modelling approach”, recognizes AI systems as able to *think* like humans, therefore demonstrating internal thought processes like those studied in psychology, cognitive and neural sciences.

However, these conceptions of AI systems are highly subjective and can be modified over time based on people’s inclination to accept machines as equal or superior entities in their own endeavours. Indeed, Turing [7] very early understood this as an emotional concept, and recognized that people would modify their definition of intelligence so that a computer could not pair it and humans could maintain their primacy.

For these reasons, we believe that the further definitions of AI systems, provided by Russell and Norvig [6], are more sound references. According to the “laws of thought approach”, they are characterized by reasoning processes derived by logic or probability, making them irrefutable and underlining their close relationship with statistics. The prevailing interpretation, which we embrace in this volume, is the “rational agent approach”. Based on this, AI systems are identified as agents able to perceive their (physical or digital) environment through sensors and act upon it through actuators, making their recognition easier and objective.

A further corroboration of this conceptualization is provided by the High-Level Expert Group (HLEG) on AI, appointed by the European Commission in 2018. They stated:

Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can

either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. [8]

Because there is no shared definition of AI-infused products, by extension, we consider them as artefacts that integrate AI systems as outlined by the HLEG on AI. These might be purely digital or physical products embedding software of this kind. For the purposes of the study, we'll focus on the latter, but we'll also use examples of digital applications to delineate their characteristics in the following sections of this chapter.

An additional point of attention related to the terminology we use needs to be remarked. We employ the term AI because it is more comprehensive and resonates with the most common ways people are used to refer to the latest technological outbreaks. However, it would be more appropriate to talk about ML and deep learning (DL) capabilities, as these are the widely spread basic algorithmic approaches that materialize AI and are bringing game-changing novelties to products and services [9, 10].

2.2.2 *The Characterizing Features of AI-Infused Products*

The definitions provided above already hint at the foundational qualities that characterize and distinguish the objects of this study from traditional products, and have neat repercussions on their UX.

Addressing Uncertainty

A first peculiarity lies in the choice of words used by the HLEG on AI to describe the purposes of AI and ML systems. They stated that these systems are given *complex goals* to fulfil [8]. The concept of complexity is relative: doing 12-digit calculations in a fraction of a second is very complicated for most people, while it is an easy job for non-sophisticated machines like calculators and, analogously, grabbing an object is natural for most but a very hard task for a robot. Therefore, *complexity* cannot be intended in a qualitative sense, as it would imply highly subjective and possibly contrasting meanings.

A possible interpretation might be related to the kind of problems AI and ML systems are supposed to address. Differently from traditional programs, they tackle issues that cannot be clearly framed. As Norvig puts it, they cannot follow logical or mathematical rules, but are required to observe uncertain situations, make hypotheses and test them, using statistics instead of logic [11]. In Simon's argumentation, unfolded in the seminal book *The sciences of the artificial* [12], this way of tackling ill-defined or wicked problems [13] resembles the approach of the design discipline. AI and ML systems do not follow linear, pre-determined and universally applicable principles, like natural sciences. Instead, they commonly work through examples and the identification of patterns to uncover possible solutions, which puts them in the unique position to operate in complex, uncertain, and changing environments.

Autonomous Adaptation

As anticipated, we identify AI-infused products as primarily integrating ML systems, which are renowned for their *autonomy* and capability to *learn*. As we might associate these features to human qualities, misinterpretations can easily arise.

Indeed, AI-infused products are not autonomous in the sense that they can self-govern themselves or are free from external control or influence—as in the Oxford Dictionary definitions of autonomy. On the contrary, the concept is strictly related to the way they operate, which is not bound to step-by-step programming. Therefore, the autonomy of these systems lies in the fact that they need to derive the rules to produce their outcomes by acquiring and processing information from their environment (usually in the form of thousands of examples of inputs and expected outputs that the programmers provide to define the system’s goal). For example, for instructing a robot to walk, the developer would give the ML system a lot of examples of people or animals walking (based on the robot having two or four legs) and possibly reward or punishment functions, according to the model they chose. In any case, the ML system is not provided any precise direction about how they should achieve the requested movement, and, for this reason, it is said that they autonomously infer and put the action in place.

This property is inherently related to the *adaptability* that uniquely characterizes ML systems. In fact, their autonomous process for finding the best possible solution to fulfil their goal includes the processing of the impact that their actions are producing in their environment. This implies that they can improve their performance based on their experience, which is commonly referred to as *learning*, and produces possible adaptations of their responses over time [6, 14]. Undoubtedly, this behaviour cannot be found in any other product, as they are completely based on pre-defined rules and functioning.

Sociotechnical Systems

The misunderstandings emerging around the autonomy and adaptability of AI-infused products, ultimately resulting in sci-fi visions where machines take over the world, are often due to what Johnson and Verdicchio [15] call *sociotechnical blindness*. Basically, they note how often the discussion about AI—and, more properly, ML systems—lacks an explicit recognition of the centrality of people.

For AI-infused products, more than others, it is essential to point out their socio-technical nature. As van de Poel [16] clearly explains, AI systems are not only made of artificial agents (the algorithms), their technical norms, and the technical artifacts (actual objects that people interact with). The human agents (including programmers, users, deployers, investors, decision, and policymakers) and their social rules and cultural conventions are an integral part of these products as they determine all the steps of their development, as illustrated in Fig. 2.1 [17]. A reminder of the fact that the technical and the social aspects of AI systems cannot be separated is provided by the HLEG on AI’s definition, itself. Indeed, it has been edited to explicitly state, as an essential requirement, that these systems exist and function only because they are designed by humans who define the objectives they have to achieve [8] and, with these, they should also set the conditions to meet these goals.

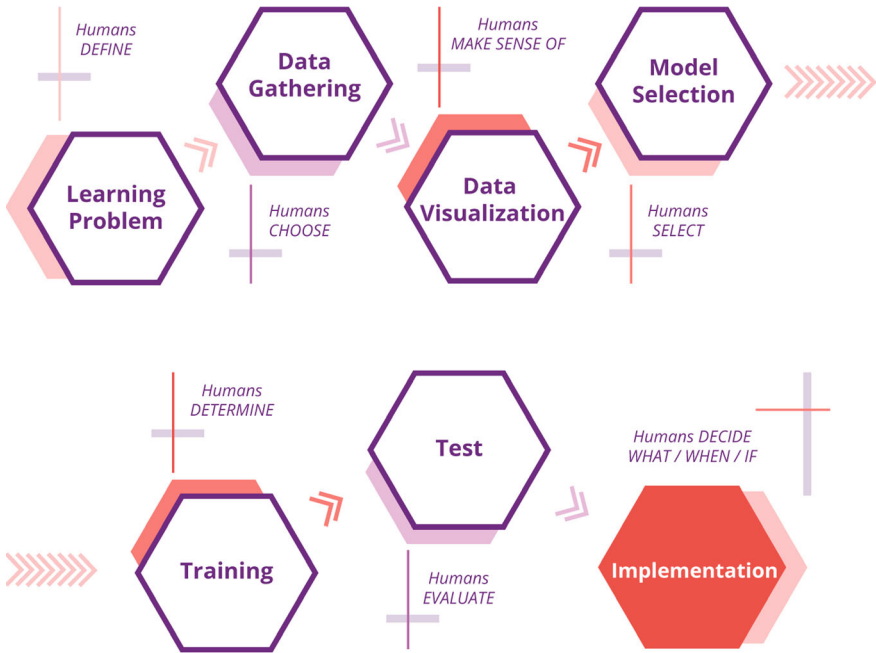


Fig. 2.1 Schema portraying how humans are involved throughout the entire AI design and development process. Original source [17]

An additional demonstration of the symbiotic relationship between people and AI-based systems is the need for the actionability of their outputs [6]. Indeed, employing AI and ML systems only makes sense if they are intended to enable humans to take action (e.g., making decisions). To provide insights on data, statistical predictions suffice.

This implies that AI-infused products are inherently designed for Human-AI cooperation. Specifically, they should be aimed at automating burdensome, repetitive, or risky human tasks; augmenting people’s capabilities; empowering them to do something very difficult or impossible for human skills alone; or inspiring further creative, critical, or insightful actions [17].

Flawed Systems

Finally, the relationship with both the design and the statistics fields makes flaws an integral part of AI-infused products. On the one hand, we have already discussed how these systems operate in uncertain and complex contexts, addressing problems that might have multiple solutions and not a uniquely correct one. This means that these systems aim at the best possible solution but this is not a guarantee that a perfect one exists.

On the other, ML systems have a probabilistic core, hence, as small as it can be, they are always subject to a certain margin of error, which is something that should be clearly and openly communicated as it inevitably affects the UX [18]. For this reason, Kozyrkov, Google's former Chief Decision Scientist, warned about considering AI systems as islands full of drunk people who will perform the required task for you [19]. Of course, the flawed nature of AI-infused products can only be worsened if we think that they feed on human data provided by people, and that unavoidably can contain bias.

Recognizing this aspect does not mean considering all AI-infused products as doomed, but it is essential to properly understand their unique position in the interaction.

2.3 UX Challenges Posed by AI-Infused Products

2.3.1 *Consequences of Their Peculiarities*

The distinguishing traits of AI-infused products do not only represent a technological breakthrough, but also new frontiers for UX. Their novelties pose challenges and unprecedented complexities for designers to handle.

As remarked, one of the main features characterizing AI-infused products is their ability to adapt over time. From a UX perspective, this translates in the possibility that the same artifact might provide different outputs to the same input, which is something that users do not expect from any other kind of product. This unpredictability, like any of the following possibilities, is not inherently negative nor positive. Thinking about smart thermostats, apprehending their users' habits and regulating the temperature accordingly can translate into energy saving. Conversely, asking an image generator such as Midjourney to create different scenes depicting the same character and seeing it change from one image to the next could be confusing and frustrating for users.

Hence, designers should put new strategies in place to make the most of it and favour the construction of new mental models. In this case, time becomes a more relevant factor in the UX. It can be exploited, in conjunction with recurring user interactions, to improve the accuracy of the outputs, to make them more personalized, or even to anticipate needs and requests.

In other words, AI-infused products can express their agency not only in technical terms, but as real agents, not just reactors, in the interaction [20]. They surpass the concept of delegated agency as framed by Kaptelinin and Nardi [21], as they not only implement the intentions delegated by humans, but might even find solutions for a problem in ways that were not anticipated by people, as this is their algorithmic purpose. This happens, for instance, with systems designed to autonomously learn a game: while we want them to play, we do not give them any rule and they will apprehend by trial-and-error (reinforcement learning). As a result, the ML systems

might uncover new winning strategies. As well, they can manifest proactivity, or the property of *sensing ahead* [22], like in the example of smart thermostats, or in Apple devices that automatically activate “full immersion” or “work” modalities, triangulating several data like geolocation, time of the day, used application and past user behaviour. As White pointed out when analysing skill discoverability in virtual assistants [23], proactivity can also ease some discoverability issues of such devices.

Indeed, one consequence of the introduction of these innovations are ergonomic challenges. While AI-infused products can perform tasks that users might not expect, these happen in a virtual space, often in the cloud. This usually means that the shape of the object or even the interface do not provide affordances about their functionalities [22] and proper communication, discussed in the next section, becomes central.

2.3.2 *Communication-based Issues*

It is no mystery that the novelty of AI-infused products is often countered by a perceived opaqueness [15, 24]. Going beyond the algorithmic aspects of this problem, which are out of the scope of this argumentation, non-intelligible functionalities or system behaviours inevitably cause frustration, uncertainty and even creepiness in users [25]. These poor outcomes in the interaction might happen because the capabilities and basic operations of these systems are still unknown to the lay public. Lacking a basic understanding of what AI is and can do in reality, people fill the gaps projecting human capabilities onto these supposedly intelligent devices or take movies as a reference. However, incomprehension also arises because of insufficient, unclear, or misleading communication.

Sometimes, companies provide disproportionate messages, provoking disillusion because they cannot fulfil their promises. An example is Google’s advertising campaign, that envisages to *make your home a nest*, with their Nest series of products for domotics. They depict a scenario of a house taking care of its inhabitants, while the reality is far from J.A.R.V.I.S., Tony Stark’s AI assistant in Iron Man movies. These systems still need human assistance and are prone to errors and misunderstandings.

Other times, a discrepancy between users’ expectations and AI capabilities is manifest. As it has been highlighted [22, 26], the concept of intelligence or smartness is not proportionate to the state-of-the-art of this technology. People tend to believe that smart devices can understand their users, in a way that is comparable to how another person might. Of course, it doesn’t match with how they actually function and their potentialities. For this reason, Shorter et al. propose building a scale that measures intelligence in a way that it’s embedded in the real world.

In general, there is uncertainty surrounding AI capabilities [24, 27]. Some patent examples can be found in people’s interactions with a system like ChatGPT, despite OpenAI’s disclaimer (written in small print). Often, users give it for granted that the responses would be indisputably true, and they are shocked when they notice that it provides incorrect calculations.

An additional consequence of interfaces that are not properly conceived for AI-infused products or of lack of information is that the actual potential of these devices is barely explored [23, 28]. If their functionalities are hidden and cannot be expected by the average user, AI-infused products like smart speakers are employed in the same way as other, better-known and less sophisticated objects, such as alarms, weather forecasters, switches, or simply speakers. The only difference would be the conversational interaction.

As one can probably sense, the problem is not only in users' non optimal interactions, but initiates at the design stage.

2.3.3 Repercussions on the Design Process

AI-based systems have been variously addressed in literature as a new material for designers [29–31]. Still, the design and HCI fields have started to give it proper attention only in the last few years and the tools and modalities for designers to deeply understand it are now emerging [4, 17, 32, 33]. Inevitably, this implies that also designers have difficulties in understanding the capabilities and potentialities of these systems, and their contribution is rarely requested from the early stages of the design process [9, 34, 35]. However, there are traces of change, and UX designers and researchers are starting to be more involved, especially in relation to Responsible AI efforts, even if with no clear purpose or guidance [34, 36].

The tangible results of this process, that is setting in motion with a considerable lag behind technological advances, are the several products that were put on the market with an evident lack of UX expertise. Some can be seen as gadgets or toys [37], appealing more to users' desire for novelty than serving any meaningful purpose in their lives. They fail to address relevant problems and are not particularly remarkable in terms of interaction quality [4, 28]. Once again, smart speakers are a clear example, but many more can be found, like smart fridges with cluttered interfaces, IoT devices that exceedingly rely on voice commands, or vacuum cleaners that do not allow their users to recover from their mapping errors and are mostly regulated by companion apps, ignoring the physical object as a suitable and often more convenient source for interaction.

Possibly the most peculiar challenge that AI-infused products pose to designers is that they offer multiple touchpoints. Frequently, a single device can be governed by different interfaces: the physical, with buttons or touch sensitive surfaces; the conversational, whether on-board or indirect, and the digital one, as these objects are usually equipped with a companion app. Moreover, AI-infused products are oftentimes part of an ecosystem, and addressing as individual items does not make much sense [38]. Considering the domestic environment as a reference, it is easy to see how smart speakers, light bulbs, vacuum cleaners and other household appliances, doorbells, security cameras, TVs, thermostats, and mobile phones themselves, operate in an integrated way. Both the interconnectedness of devices and the variety of interfaces, in which functions and feedback overlap, are new and complex realities to tackle

from a UX and interaction design standpoint and need to be unravelled also in terms of approaches and processes.

2.3.4 Ethical and Societal Impacts

Lastly but not less importantly, most of the challenges previously outlined sum up to generate ethical concerns. These can be synthesized in (i) lack of understanding and transparency in the relationship with AI-infused artifacts, and (ii) necessity for human factors in their design and development. Fundamentally, both issues substantiate how a reliable product not only needs to be robust, but it also must gain users' trust.

Shorter et al. [22] provocatively point out the apparent inconsistency between users' declared concerns about their privacy with AI systems and their actual behaviour, betraying them as careless or favourably inclined to give up their data, even with very little in return. While this might cast doubts regarding the purpose for UX to address ethical concerns at all, with their experimentations, the authors also demonstrate how effective designerly interpretations of these issues can be. A patent example is the Microphone not Speaker provotype: translating one of the basic functions of smart speakers into their actual shape dramatically help to unveil potential problematic consequences for privacy.

Undoubtedly, the matter is quite slippery and extends to several disciplinary fields. Indeed, in the last few years, a plethora of ethical guidelines for AI emerged [39], and some useful resources also arose from the design field in the form of tools [40–42]. We will not deepen all the potential issues here, and demand to the *Ethics Guidelines for Trustworthy AI* [8] as the vastest resource one can reference. However, despite the more or less punctual and applicable actions they can implement, a huge challenge for UX designers dealing with ethical concerns and societal implications will be to adapt their methods and approaches so that they account for the new centrality these problems have. One possibility might be to adopt a value-centred approach for designing AI-based systems [17, 43], but this is a very young and promising area of research.

2.4 New Interaction Paradigms

2.4.1 Design for Something Else

AI-infused products introduce new ingredients to the UX, they have unique features and present unprecedented challenges. At a higher level, they are changing the ways we interact with and design for them.

Specifically, we have already discussed how they call for novel ways of approaching the design process, widening its scope to include multiple touchpoints

and human values. Systemic and holistic methods will gain more relevance as we get closer to the materialization of ubiquitous computing, as it was envisioned by Weiser [44]. The conventional ways of thinking about interfaces will fade and the object of the design effort will be the whole environment in which people's interactions with the pervasive machines will take place, comprehending experiential, technical, and social facets.

From another angle, AI-infused products can be perceived as more than just machines. They are not just users' extensions to complete tasks, but *counterparts*, or *otherware* [45]. This shift moves away from the traditional paradigm of *embodiment* [46] and introduces yet another concept to keep in consideration when designing for AI-infused products. Hassenzahl et al. refer to it as *alterity* [45]. Hence, we will no longer be designing just interactive products, but distinct entities that we can engage in conversations with, refer to for support or inspiration, or even delegate burdensome activities.

Still, because these are closer to living beings, that can evolve and handle parts of tasks autonomously, rather than rigorous and consistent instruments, a novel foundational paradigm must be learnt: design for imperfection.

As anticipated, due to their statistical nature, AI systems cannot provide 100% accurate outputs. The advancements in the field are constantly improving them, but designers' concerns should change a little to accommodate this peculiar situation.

Instead of thinking about AI-infused products as Swiss watches that will surely meet users' expectations, designers should focus on contexts and KPIs (key performance indicators) that allow for non-optimal results. As suggested by Kozyrkov [19], along with the ideal outcome, one should define what it means for the product to behave *well enough*. What are the constraints? What the minimum requirements of acceptability? And this should be paralleled by another essential reflection: in which context and under which circumstances is it ok to have an imperfect outcome? For instance, in the entertainment sphere, if we do not get a song or a movie that we actually like, we can just move to a different one with minimal repercussions. On the contrary, having inaccurate diagnoses in medical matters or flawed assumptions in judicial cases might have very serious implications for people's life. Therefore, these situations would require considering measures that account for errors from the very beginning of the design process.

Similar expedients can also be contemplated from the user perspective, supporting their understanding and giving them agency in case things do not come out as they expected [47].

2.4.2 Reimagining Interfaces

The latest technological developments in ML systems are allowing Weiser's vision of invisible interfaces to become reality [44].

Conversational interfaces are the most evident novelty in this area. The possibility to dialogue with machines has been a sci-fi dream for a long time, and it has been

now materialized by virtual assistants and chatbots. These inherently change the way we interact with computer-based artifacts as we transition from concise directives following apprehended patterns to common chats. Natural Language Processing (NLP) allows people to talk or write to machines in a very similar way as they would to another person. Yet, this new possibility is still characterized by misunderstandings, misaligned expectations, and awkwardness, as dialoguing with objects using natural language is not at all natural. It represents a new paradigm for UX and as such it is being thoroughly addressed. Some scholars are analysing the most relevant features of conversational interfaces [48, 49], while others are going to their roots and investigating their meaning [22]. Obviously, there are several ways in which designers can envision and develop conversational interfaces [50]. Still, at the core, they need to figure out how to handle the new interface in combination or in substitution to existing ones.

Less widely spread but very much interrelated with the possibilities enabled by ML systems are the interactions based on image recognition. These allow for gestural or bodily interfaces, which have gained the attention of researchers to offer alternatives to screen-based and physical ones [51–53]. Like conversational ones, they are built from the premise of users interacting and communicating with products in a natural way. Examples of implementations include systems like Face ID, which enables users to unlock phones or even secure access to doors through facial recognition. In online meeting platforms, secondary functions such as a “thumbs up” gesture allow for non-verbal communication, highlighting user reactions seamlessly. These interfaces are also pivotal in hands-free interactions with devices like the Apple Vision Pro, where users can perform actions, such as grabbing and moving apps or scrolling through pages, with simple gestures, without requiring additional input tools. And they can pave the way for applications in new contexts, like within the car when driving [54].

While the role of intuitive interactions is growing in modern UX, experimentations and sense making are necessary. In particular, prototyping modalities and tools will require substantial research to capture these shifts [55, 56].

2.4.3 New Forms of Assistance

Finally, the different relationships with AI-infused products are also associated to new purposes of interaction. As mentioned, smart speakers, self-driving cars, smart home appliances, are not only objects we interact with, but entities with which we cooperate and co-evolve [31]. The distinctive possibilities offered by AI systems have brought the products and services that integrate them to new levels of assistance for their users. AI has become integral to three key functions: decision support, recommendation, and content generation.

To date, these advancements have primarily been realized in digital products, with their UX largely confined to computers and mobile devices. However, we believe that they are worth mentioning in the purview of future embodiment in physical products, as AI-based capabilities become more widespread.

Intelligent decision-support systems can be applied in many different contexts, from finance to air traffic management, hence encompassing high-stakes domains. Their main objective is allowing people to make informed choices and clarity is critically essential. The reason why this is a UX matter is because the interface, of whatever kind, assumes a primary role to ensure that information is properly presented and adequate to such a role of responsibility [57].

Analogously, recommendation systems are very powerful in nudging people's behaviour and their implications may span from trivial to serious societal ones. Also in this case, the UX needs to incorporate the awareness of ergonomic and ethical challenges to live up to these systems' capacity as fair advisors.

To conclude, the popular enthusiasm toward generative systems also brings to the foreground of the UX field points of reflection about their contribution in human-AI interaction. They can easily be seen as substitutes for people's work, with consequent negative implications. Nonetheless, they can also be used and designed to enhance people creativity and reflexivity. For sure, generative systems are an intriguing topic of debate and research in the UX and HCI communities, which are already looking principles to orient themselves in this challenging space [47].

2.5 Conclusion

The chapter portrays the background to the research project presented in this book. It offers an overview of the peculiar features, challenges, and interaction paradigms of AI-infused products. For the scope of the investigation, these are intended as physical objects integrating AI or—more precisely—ML systems and can thus express their capabilities. However, examples from both the tangible and the digital world are referenced for a more comprehensive overview.

AI-infused products bring a set of unique qualities that distinguish them from traditional products. Their ability to handle uncertainty, learn from experience, and adapt accordingly by autonomously inferring how to achieve their goals makes them powerful but also complex. Unlike static systems, these products can evolve over time, adjusting their behaviour based on new data and patterns. This adaptability is central to their function, but it also creates unpredictability and a certain opaqueness, especially to inexperienced users, who are still lacking basic knowledge about AI systems. The absence of clear affordances and transparency often leaves users uncertain about how these systems operate. Inevitably, this affects the UX, possibly causing confusion, frustration, or mistrust, which might be increased by their flawed nature. Indeed, AI systems are constructed on probabilistic premises and their outcomes can present errors.

Already, these characteristics introduce significant challenges to the UX design process, as it must account for the consequences of the agency that AI system have, spanning from their unpredictability to their possible proactivity.

Moreover, AI-based systems, as sociotechnical systems, require a special attention for ensuring a balance between the focus on their technical possibilities and the

role and implications they have for people. Their usefulness and meaning, as well as ethical concerns around their reliability also require careful consideration, starting from the early stages of the design process. This reflects on the job of designers, who need to be prepared to tackle these novelties. Several UX aspects need further exploration, from ergonomic issues to the ecosystems of people and artificial agents that AI-infused products generate, and the introduction of human factors and principles in the development of this new type of artifacts can positively impact their outcomes.

In addition to these challenges, AI-infused products are reshaping interaction paradigms. They demand designers to move beyond traditional tool-based interactions and instead consider systems that engage dynamically with users. Interfaces now need to account for imperfection, the distribution across multiple touchpoints, and even new formats, like the conversational and gestural ones. The purposes of products themselves can shift, introducing new possibilities to assist people: supporting their decision-making, providing customized recommendations, or even generating content for them to ease their tasks or spark creativity.

Given these distinctive qualities, challenges, and evolving interaction paradigms, a significant distance from traditional UX principles becomes quite manifest. The picture provided in this chapter, combined with similar concerns found in literature [58, 59], raises a reasonable doubt about the capability of traditional UX evaluation methods to encompass the peculiarities of AI-infused products. It is plausible that the dynamic, adaptive nature of these systems, their unfulfilled potentialities, flaws, ethical concerns, and their new roles in supporting people, demand a new approach to evaluation.

For these reasons, the suitability of current assessment methods is further deepened in the following chapters, starting from the analysis of the dimensions measured to determine the quality of the UX.

References

1. Burr C, Taddeo M, Floridi L (2020) The ethics of digital well-being: a thematic review. *Sci Eng Ethics*. <https://doi.org/10.1007/s11948-020-00175-8>
2. Nishant R, Kennedy M, Corbett J (2020) Artificial intelligence for sustainability: challenges, opportunities, and a research agenda. *Int J Inf Manag* 53:102104. <https://doi.org/10.1016/j.ijinfomgt.2020.102104>
3. Maslej N, Fattorini L, Perrault R et al (2024) The AI index 2024 annual report AI. Index Steering Committee, Institute for Human-Centered AI, University of Stanford, Stanford, CA
4. Yildirim N, Oh C, Sayar D, et al (2023) Creating design resources to scaffold the ideation of AI concepts. In: *Proceedings of the 2023 ACM designing interactive systems conference*. Association for Computing Machinery, New York, NY, USA, pp 2326–2346
5. McCarthy J, Minsky ML, Rochester N, Shannon CE (2006) A proposal for the Dartmouth summer research project on artificial intelligence, 1955. *AI Mag* 27:12–12. <https://doi.org/10.1609/aimag.v27i4.1904>
6. Russell S, Norvig P (2020) *Artificial intelligence: a modern approach*, 4th edn. Pearson, Hoboken, NJ
7. Turing AM (1948) *Intelligent machinery*. National Physics Laboratory

8. High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI
9. Dove G, Halskov K, Forlizzi J, Zimmerman J (2017) UX design innovation: challenges for working with machine learning as a design material. In: Proceedings of the 2017 CHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 278–288
10. Zhang D, Maslej N, Brynjolfsson E et al (2022) The AI index 2022 annual report. AI Index Steering, Stanford Institute for Human-Centered AI, Stanford University, Stanford
11. Norvig P (2018) Introduction to machine learning | Google developers. <https://developers.google.com/machine-learning/crash-course/ml-intro>. Accessed 7 Dec 2022
12. Simon HA (1969) The sciences of the artificial, 3rd edn. MIT Press, Cambridge, Mass
13. Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. *Policy Sci* 4:155–169. <https://doi.org/10.1007/BF01405730>
14. Mitchell TM (1997) Machine learning. McGraw-Hill, New York
15. Johnson DG, Verdicchio M (2017) Reframing AI discourse. *Minds Mach*. <https://doi.org/10.1007/s11023-017-9417-6>
16. van de Poel I (2020) Embedding values in artificial intelligence (AI) systems. *Minds Mach* 30:385–409. <https://doi.org/10.1007/s11023-020-09537-4>
17. Sciannamè M (2023) Machine learning (for) design. Towards designerly ways to translate ML for design education. Phd Thesis, Department of Design, Politecnico di Milano
18. Feng KJK, McDonald DW (2023) Addressing UX practitioners’ challenges in designing ML applications: an interactive machine learning approach. In: Proceedings of the 28th international conference on intelligent user interfaces. Association for Computing Machinery, New York, NY, USA, pp 337–352
19. Kozyrkov C (2022) Advice for finding AI use cases. In: Medium. <https://kozyrkov.medium.com/imagine-a-drunk-island-advice-for-finding-ai-use-cases-8d47495d4c3f>. Accessed 12 Mar 2023
20. Giaccardi E, Redström J (2020) Technology and more-than-human design. *Des Issues* 36:33–44. https://doi.org/10.1162/desi_a_00612
21. Kaptelinin V, Nardi BA (2009) Acting with technology: activity theory and interaction design. MIT Press, Cambridge, London
22. Shorter M, Minder B, Rogers J, et al (2022) Materialising the immaterial: prototyping to explore voice assistant complexities. In: Proceedings of the 2022 ACM designing interactive systems conference. Association for Computing Machinery, New York, NY, USA, pp 1512–1524
23. White RW (2018) Skill discovery in virtual assistants. *Commun ACM* 61:106–113
24. Dove G, Fayard A-L (2020) Monsters, metaphors, and machine learning. <https://www.semanticscholar.org/paper/Monsters%2C-Metaphors%2C-and-Machine-Learning-Dove/4b1ea05c83d44b984db8b7d5764ef306d602dc35>. Accessed 18 Apr 2020
25. Fruchter N, Liccardi I (2018) Consumer attitudes towards privacy and security in home assistants. In: Extended abstracts of the 2018 CHI conference on human factors in computing systems. Association for Computing Machinery, Montreal QC, Canada, pp 1–6
26. Kulesz O (2018) Culture, platforms and machines: the impact of artificial intelligence on the diversity of cultural expressions. UNESCO, Paris
27. Yang Q, Steinfeld A, Rosé C, Zimmerman J (2020) Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In: Proceedings of the 2020 CHI conference on human factors in computing systems. Association for Computing Machinery, Honolulu, HI, USA, pp 1–13
28. Sciuto A, Saini A, Forlizzi J, Hong JI (2018) “Hey Alexa, what’s up?”: a mixed-methods studies of in-home conversational agent usage. In: Proceedings of the 2018 designing interactive systems conference. ACM, New York, pp 857–868
29. Antonelli P (2011) Talk to me: design and communication between people and objects. MoMa, New York
30. Stoimenova N, Price R (2020) Exploring the nuances of designing (with/for) artificial intelligence. *Des Issues* 36:45–55. https://doi.org/10.1162/desi_a_00613

31. Yang Q (2020) Profiling artificial intelligence as a material for user experience design. Carnegie Mellon University
32. Ghajargar M, Bardzell J (2023) Making AI understandable by making it tangible: exploring the design space with ten concept cards. In: Proceedings of the 34th Australian conference on human-computer interaction. Association for Computing Machinery, New York, NY, USA, pp 74–80
33. Jansen A, Colombo S (2023) Mix & match machine learning: an ideation toolkit to design machine learning-enabled solutions. In: Proceedings of the seventeenth international conference on tangible, embedded, and embodied interaction. Association for Computing Machinery, New York, NY, USA, pp 1–18
34. Liao QV, Subramonyam H, Wang J, Wortman Vaughan J (2023) Designerly understanding: information needs for model transparency to support design ideation for AI-powered user experience. In: Proceedings of the 2023 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 1–21
35. Meyer MW, Norman D (2020) Changing design education for the 21st century. *She Ji J Des Econ Innov* 6:13–49. <https://doi.org/10.1016/j.sheji.2019.12.002>
36. Varanasi RA, Goyal N (2023) “It is currently hodgepodge”: examining AI/ML practitioners’ challenges during co-production of responsible AI values. In: Proceedings of the 2023 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 1–17
37. Levinson P (1977) Toy, mirror, and art: the metamorphosis of technological culture. *ETC Rev Gen Semant* 34:151–167
38. Zhang R, Shi Y, Schuller B, et al (2021) User experience for multi-device ecosystems: challenges and opportunities. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 1–5
39. Algorithmic Watch (2020) AI ethics guidelines global inventory. <https://inventory.algorithmwatch.org/>. Accessed 27 Jul 2021
40. Calderon A, Taber D, Qu H, Wen J (2019) AI blindspot: a discovery process for preventing, detecting, and mitigating bias in AI systems
41. Futurice (2017) The intelligence augmentation design toolkit. <https://futurice.com/ia-design-kit>. Accessed 1 Dec 2021
42. IDEO (2019) AI & ethics: collaborative activities for designers. <https://www.ideo.com/post/ai-ethics-collaborative-activities-for-designers>. Accessed 5 Jul 2021
43. Umbrello S, van de Poel I (2021) Mapping value sensitive design onto AI for social good principles. *AI Ethics* 1:283–296. <https://doi.org/10.1007/s43681-021-00038-3>
44. Weiser M (1994) Creating the invisible interface. In: ACM Symposium on user interface software and technology UIST’94. Marina del Rey, California
45. Hassenzahl M, Borchers J, Boll S, et al (2020) Otherware: how to best interact with autonomous systems. *Interactions* 28:54–57. <https://doi.org/10.1145/3436942>
46. Dourish P (2001) Where the action is: the foundations of embodied interaction. The MIT Press, Cambridge, Mass
47. Weisz JD, et al (2024) Design principles for generative AI applications. In: Proceedings of the CHI conference on human factors in computing systems, pp 1–22. <https://doi.org/10.1145/3613904.3642466>
48. AI is your new design material at amuse UX conference (2019)
49. Maguire M (2019) Development of a heuristic evaluation tool for voice user interfaces. In: Marcus A, Wang W (eds) *Design, user experience, and usability. Practice and case studies*. Springer International Publishing, Cham, pp 212–225
50. Vitali I, Paracolle A, Arquilla V (2023) The role of design in the era of conversational interfaces. In: Sciannamé DS Martina (ed) *Embedding intelligence. designerly reflections on AI-infused products*, First edition. Franco Angeli
51. Grandhi SA, Joue G, Mittelberg I (2011) Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. In: Proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 821–824

52. Kadaskar HR (2024) Enhancing user experience in mobile application design through gestural interaction: a human-computer interaction perspective. *Int J Sci Res Mod Sci Technol* 3:01–06. <https://doi.org/10.59828/ijstrmst.v3i8.239>
53. Pomboza-Junez G, Holgado-Terriza JA, Medina-Medina N (2019) Toward the gestural interface: comparative analysis between touch user interfaces versus gesture-based user interfaces on mobile devices. *Univ Access Inf Soc* 18:107–126. <https://doi.org/10.1007/s10209-017-0580-6>
54. Jahani H, Alyamani HJ, Kavakli M et al (2017) User evaluation of hand gestures for designing an intelligent in-vehicle interface. In: Maedche A, vom Brocke J, Hevner A (eds) *Designing the digital transformation*. Springer International Publishing, Cham, pp 104–121
55. Subramonyam H, Seifert C, Adar E (2021) ProtoAI: model-informed prototyping for AI-powered interfaces. In: *Proceedings of the 26th international conference on intelligent user interfaces*. Association for Computing Machinery, New York, NY, USA, pp 48–58
56. Yang Q (2018) Machine learning as a UX design material: how can we imagine beyond automation, recommenders, and reminders?
57. Karahasanovic A, Gausta Nilsson E, Grani G, et al (2021) User involvement in the design of ML-infused systems. In: *CHI Greece 2021: 1st international conference of the ACM Greek SIGCHI chapter*. Association for Computing Machinery, New York, NY, USA, pp 1–5
58. Pavlovic M, Kotsopoulos S, Lim Y, et al (2020) Determining a framework for the generation and evaluation of ambient intelligent agent system designs. In: Arai K, Bhatia R, Kapoor S (eds) *Proceedings of the future technologies conference (FTC) 2019*. Springer International Publishing, Cham, pp 318–333
59. Pettersson I, Lachner F, Frison A-K, et al (2018) A Bermuda triangle? A review of method application and triangulation in user experience evaluation. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. Association for Computing Machinery, Montreal QC, Canada, pp 1–16

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 3

UX Dimensions for AI. Past and Future Perspectives



Abstract This chapter offers a comprehensive overview and systematic analysis of current UX evaluation methods with the objective to identify relevant dimensions to describe the qualities of AI-infused products and possible gaps. The state of the art is grounded on an umbrella review of previous consistent studies, a scoping review of about 130 UX evaluation methods, and a systematic literature review on the assessment of AI-based systems. This preliminary investigation substantiated the need of new, specific qualities to properly assess the UX of AI-infused products, and eight primary UX dimensions have been identified as a starting point for more experimental studies.

3.1 Introduction

The advent of AI in consumer products has introduced new complexities and challenges that necessitate a re-evaluation of existing UX methods. AI-infused products, such as smart speakers and autonomous vehicles, operate on principles of adaptive behaviour and autonomous achievement of results. The dynamic nature of these products, not only responding but also evolving based on people's behaviour, poses unique challenges for UX evaluation, which must account for such unpredictability, among other things.

The integration of AI into products significantly impacts the UX, as more extensively explored in Chap. 2. These might raise important questions about trust, transparency, and user control [1]. Users need to understand how AI systems get to the presented outcomes, trust that these results are reliable and in their best interest, and feel in control of their interactions with the system. Traditional UX methods, which were developed for static and predictable systems, may fall short in addressing these new qualities.

Special attention also needs to be reserved to the new relational dynamics, interfaces, interaction paradigms and implications that dealing with these systems generate [2].

Recognizing these challenges, the chapter delves into the past perspectives and foreseeable tendencies in UX evaluation, specifically focusing on AI-infused products. It aims to respond to three main research questions to get a better understanding of UX evaluation when AI systems are involved: RQ1—Are current UX assessment methods enough for AI-infused products? RQ2—Are new UX dimensions needed for these products? RQ3—What characteristics should a new method have?

To this end, the researchers explored the adequacy of current UX evaluation methods for AI-infused products through an umbrella and a scoping review of general UX assessment methods and identified potential areas for improvement to effectively capture the complexities of AI interactions. This has been achieved through a systematic review targeting AI-based systems.

By addressing these objectives, we aim to set the stage for a thorough analysis of UX dimensions in AI, paving the way for the development of a tailored UX evaluation method capable of meeting the needs raised by AI-infused products.

Thus, the chapter first provides the background and context for the study, reviewing the evolution of UX evaluation research and methods, and highlighting key developments and trends. Subsequently, it describes the review methods adopted in this study, including the umbrella review, scoping review of current UX methods, and systematic review of AI-related UX. Then, it describes the results of the investigation and discusses the findings in relation to the research questions, focusing on the most relevant dimensions.

Finally, the chapter summarizes the key points, suggesting future research directions.

3.2 The Evolution of UX Assessment

The field of UX has evolved significantly over the past few decades, transitioning from basic usability studies to comprehensive evaluations that encompass emotional and experiential aspects of interaction. As digital products have become more complex and integral to everyday life, the methods for evaluating UX have similarly advanced, adapting to technological advancements and shifts in user expectations.

This section pinpoints the key stages in the evolution of UX evaluation to illustrate the context in which the study arises.

In the early days of computing, UX evaluation primarily focused on usability. This period, dating back to the 1980s, was characterized by lab-based studies where users interacted with software or hardware under controlled conditions. The emphasis was on identifying and mitigating usability issues to improve efficiency, effectiveness, and satisfaction. Researchers like Nielsen and Molich [3] introduced heuristic evaluation, a method that involved experts reviewing interfaces against ten established usability principles. The technological landscape at this time was marked by the proliferation of personal computers and the rise of graphical user interfaces (GUIs). Usability testing became crucial as software applications moved from command-line interfaces to more complex GUIs, requiring more intuitive and user-friendly designs.

The 1990s saw the emergence of cognitive and contextual approaches to UX evaluation. Researchers began to consider the cognitive processes of users, leading to methods like cognitive walkthroughs [4], which focused on the users' thought processes and decision-making during interaction. On a similar note, contextual design, introduced by Beyer and Holtzblatt [5], emphasized understanding users' needs in their natural environments, leading to more ethnographic methods of evaluation. This period was characterized by the rise of the internet, with web usability becoming a critical focus. Indeed, the internet introduced a vast, interconnected network of information and services, which significantly increased the complexity of the interfaces. Early websites varied greatly in design and usability, often leading to user frustration and difficulty in finding information [6]. Hence, the HCI community started to explore how users navigated and interacted with web-based applications, pushing for designs that accommodated real-world contexts and cognitive constraints.

As a further evolution, User-Centred Design (UCD) gained prominence in the late 1990s and early 2000s, advocating for iterative design processes that involve users at every stage. This period saw the development of various user-centred evaluation techniques, also integrated in participatory design activities and agile methodologies, linked to the rapid evolution of mobile technology. In fact, the widespread adoption of smartphones brought new challenges and opportunities for UX evaluation, as designers needed to account for diverse usage contexts, touch interfaces, and varying screen sizes, reinforcing the importance of focusing on how people actually interact with these new products.

As digital products became more and more embedded in daily life, competition in the market increased, with a growing recognition that usability alone was insufficient to ensure a positive user experience. The rise of experience design emphasized creating meaningful, engaging, and memorable experiences rather than just functional interfaces.

Supporting this stream of research, Jordan [7], Tractinsky et al. [8], Norman [9], Desmet and Hekkert [10]—just to name a few—emphasized the importance of emotional design, highlighting the importance of aesthetics and emotional responses in user experience. Indeed, they posited that attractive products are perceived to work better due to positive affective responses, significantly impacting the UX.

This shift from functionality and efficiency to recognizing users' emotional and experiential responses spread in Human–Computer Interaction (HCI) and interaction design. Recently, Hassenzahl et al. [11] emphasised the significance of this transition in a review of the past two decades of UX research, starting from their influential paper published in 2002 [12]. It challenged the prevailing notion that usability was the sole determinant of a good UX, arguing instead that experience is a multifaceted construct and requires an expansion from functionality and efficiency to include hedonic qualities, considering how users feel and the significance they derive from their interactions with products and systems [13].

In recent years, as digital technologies have become more pervasive, encompassing ubiquitous IOT, wearable devices, and smart environments, the need to account for complex and dynamic user interactions emerged. Thus, we witnessed

an increasingly holistic approach to UX assessment to ensure a more comprehensive understanding of users' needs and behaviours. The UX evaluation methods spanned from qualitative to quantitative, also leveraging the widespread availability of user data related to their activities on web and mobile applications. Techniques like A/B testing, usability testing (also in a remote format), and analytics have become integral to UX evaluation, allowing for data-driven decision-making that resonate to additional fields, like marketing and management. Quantitative metrics, in particular, allow for the inclusion of a diverse user base in real-world contexts and provide a sense of objectivity that is effective in communicating with a variety of stakeholders. Nonetheless, qualitative methods remain essential to inform the design process.

As this brief synthesis shows, the evolution of UX evaluation reflects the dynamic nature of this field. Following technological developments and playing a critical role in responding to users' needs, UX methods have consistently advanced to allow for effective and enjoyable experiences. Hence, it is time to explore how to respond to the challenges that AI-infused items are bringing.

3.3 Methodology

Based on the hypotheses that: (i) AI-infused products present unique challenges to user experience; (ii) existing UX evaluation methods might be insufficient in addressing these challenges; and (iii) novel qualities of user experience should be considered, potentially establishing the foundation for a new assessment method, this study is structured in three distinct review phases followed by a thorough reflective analysis, mainly conduct to infer the characteristics of the new method (RQ3).

Phase 1—Umbrella Review

To answer RQ1—*Are current UX assessment methods enough for AI-infused products?*—the introductory stage consisted of an umbrella review.

An initial query focused on works published between 2000 and 2020, sourced from the ACM Digital Library and Springer Link using the query strings “UX AND evaluation” and “UX AND assessment.” This analysis concentrates on the limited number of articles that trace the evolution of UX and its assessment over time. These are depicted in Table 3.1.

Phase 2—Scoping Review of UX Evaluation Methods

To get a comprehensive picture of the state of the art of UX evaluation and understand the most relevant qualities describing interactions between people and various artefacts as well as their adequacy to address AI-infused products, the investigation progressed into an extensive and thorough review and critical analysis of current UX evaluation methods.

The research team, comprising five researchers, independently identified and examined relevant scales and methods for assessing UX, both within the field of design and related social sciences. The review was confined to the articles emerged

Table 3.1 Most relevant articles tracing the evolution of UX evaluation methods and related UX studies samples

References	Sample of UX evaluation methods
Vermeeren et al. [16]	96
Bargas-Avila and Hornbæk [17]	66
Lachner et al. [45]	84
Rivero and Conte [18]	227
Pettersson et al. [19]	100

in the previously described query. To ensure comprehensive coverage, the All About UX website—the largest repository of UX evaluation methods available at the time—was also consulted. This process yielded a list of 129 UX evaluation methods, which were then analysed according to various criteria [14].

Analysis Criteria The analysis had a twofold objective. Primarily, it focused on identifying the UX dimensions and descriptors in each method. With the term dimensions, we refer to the general qualities that significantly describe people’s experiences of products. Descriptors, instead, explain the nuances of these overarching qualities.

Additionally, considering the research project goal to build a UX evaluation method for AI-infused products, the study also examined the operationalisation of current ones. Specifically, also based on the points of attention raised in the studies analysed in the umbrella review, the following aspects were considered:

- **Collection Method(s):** the modalities used to gather UX evaluations (e.g., questionnaires, interviews, physiological measurements, etc.).
- **Triangulation:** whether multiple methods were employed and cross-examined.
- **Lab/Field:** the context in which the evaluation took place.
- **Support Materials:** including all the physical and digital tools used for the collection of the assessments.
- **Nature of Investigation:** whether the study was qualitative, quantitative, or both.
- **Development Phase:** the development stage of the artefact to assess (concept, early prototype, functional prototype, or market level).
- **Period of Experience:** when the tester compiled the evaluation (before use, after an episodic interaction, post-task completion, or long-term use).
- **Object(s) of Study:** the type of product(s) evaluated.
- **Evaluators:** the profile of people testing and evaluating (single users, groups, expert users).

For each method, sources, links, and notes were included for easy reference, and researchers rated their level of consistency with the purposes of the investigation, evaluating how coherent each method is with the assessment of AI-infused products.

This multifaceted approach was crucial for identifying tendencies and gaps in current evaluation methods.

Phase 3—Systematic Review of UX Evaluation for AI-Related Artefacts

Moving to RQ2—Are new UX dimensions needed for these products?—a closer look into scientific papers specifically addressing AI-infused artefacts was necessary. Even after a recent update, we could observe that there is not an abundance of vertical studies on the subject, and a contribution to this field is timely.

The systematic review was conducted using Scopus with the aim of capturing relevant literature on the UX of AI-infused artefacts. The query used was:

TITLE-ABS-KEY (“artificial intelligence” OR ai) AND (“user experience” OR ux OR “user interaction” OR “user interface” OR ui) AND ((assessment OR evaluation) W/0 (method OR framework OR approach OR heuristic OR system)).

This search produced 171 entries. To refine the results, we applied several inclusion criteria. We considered scientific papers published as conference proceedings, journal articles, or book chapters, excluding conference reviews. The publications needed to be in English and describe a framework or evaluation system related to user experience, even if not explicitly coded as such. Furthermore, the focus had to be on evaluations of AI-infused artefacts, not on AI systems supporting evaluation or on assessments of AI-generated products. We included any form of AI systems, such as natural language processing, computational creativity, or recommender systems, with the condition they were integrated into wider systems with which users can interact.

The selection process involved an initial screening of titles and abstracts, which led to the exclusion of 129 documents. Additionally, two entries were previous publications by the authors and were not considered further. This left 40 documents, which were subjected to a full-text review to ensure their relevance and alignment with the study objectives. Of these, 24 qualified for the research purposes. Indeed, eight targeted developers and focused on technical aspects of AI systems, two were out of scope even if marginally including AI, two did not include any attribute (only estimations or argumentations about other methods, which we had already assessed), and four did not have the full text available.

The selected papers were coded by the researchers, following two cycles. The first included Initial and In Vivo Coding [15] to highlight all the emerging qualities to evaluate AI-infused artefacts. The second cycle of Axial Coding [15] was instead meant to categorize the previously identified attributes into overarching dimensions according to both the authors of the papers and the researchers of the present study.

3.4 Results

3.4.1 Umbrella Review

This section aims to provide an overview of the state-of-the-art in UX evaluation methods, highlighting key findings and insights from seminal literature reviews, encompassing the period from 2000 to 2020. By comparing the results of these

studies, we can better understand the possible needs and future directions for the development of UX evaluation toward the necessities presented by AI systems.

Overview of the Literature Review Papers

Vermeeren et al. [16] gathered a wide array of 96 UX assessment techniques from both academia and industry. They classified these techniques according to various criteria, including product development stages and the periods of user experience they evaluate. The study highlights that most methods were applied to digital interfaces at advanced stages of development involving fully functional systems or prototypes. Indeed, among the areas that would need further development, the authors identified early-stage evaluation, as well as collaborative UX assessment that can guarantee practical applicability and scientific rigor.

Bargas-Avila and Hornbæk [17], reviewing 51 publications from 2005 to 2009 and encompassing 66 empirical studies, also highlighted a strong focus on digital interfaces. Their research revealed that situated data about people's context of use and expectations are rarely considered, while emotions, enjoyment, and aesthetics are the most commonly assessed dimensions. Besides the dominance of familiar qualitative approaches derived from usability research tradition, the authors interestingly observe the growing use of constructive methods, usually self-developed and with unclear validity because they lack item transparency and statistical validation.

Rivero and Conte [18] pushed this exploration further, focusing their thorough review on UX evaluation technologies mostly to assess digital interfaces. From the 227 papers analyzed, they confirmed a reliance on traditional methods such as questionnaires, observations, and physiological measurements. While these might provide accurate information during the user experience, the authors noted that they were typically used in controlled settings, either during or after the interaction. Hence, they pointed to the need for technologies that allow the collection of valuable quantitative information in user-friendly ways, directly in their real context of use.

In a subsequent study, Pettersson et al. [19] reviewed 100 academic articles published between 2010 and 2016, reserving particular attention to the triangulation of UX evaluation methods. They highlighted an increase in triangulated techniques to enhance the robustness of UX assessments, which is probably related to the higher proportion of field studies over lab studies, granting a more comprehensive and contextualized understanding of the UX results. For the rest, their findings are consistent with the previous studies. Pragmatic features, such as usability, are still the most frequently considered, and questionnaires and interviews are the most common formats, and often combined. Additionally, outlining critical future UX research topics, Pettersson et al. [19] identified the need to adapt UX methodologies to evolving technologies and non-human intelligence.

Emerging Trends

The analysis of the presented studies revealed recurring features of UX evaluation methods, from which some tendencies can be outlined.

The most patent one regards the preferred format for collecting evaluations. Questionnaires are employed by the vast majority of studies, being either self-developed

or standardised [19]. The investigation by Vermeeren et al. [16] reports 42 out of 96 methods collecting data in this way, representing the most frequent modality at all development stages. The authors further state that this is the most versatile and often misused evaluation means. Rivero and Conte [18] also underline that questionnaires are the most familiar tool for users, being able to proceed with the evaluation without requiring the researcher's intervention. As well, they easily allow for quantifiable and comparable results.

The stage of development of the artefact under investigation and the temporal scope of the evaluation show additional dominances in the current panorama. Vermeeren et al. [16] specifically observed that UX evaluation methods predominantly focus on advanced stages of product development, often overlooking critical issues that arise earlier in the design process. Only a small percentage of methods (25%) are also suitable for early-stage product development.

They also provided a broad overview of UX assessment methods, ranging from unique snapshots or episodic activities to long-term usage through specific tasks. Their analysis found that only 36% of methods assessed systems' performance in everyday life over long-term interactions. However, Rivero and Conte [18] highlighted a much lower percentage (6.6%) of longitudinal evaluation methods. Indeed, while remaining a standard practice in UX evaluation, task-based experimentations are more common.

Also from the *object of study* perspective, the arising portrait is quite traditional, addressing the sphere of digital products and emphasizing their HCI origins. Vermeeren et al. [16]'s investigation underlined a prevalence of web services (81%), mobile software (77%), and PC software (76%). Bargas-Avila and Hornbæk [17] similarly highlighted a focus on mobile applications, phones, audio, video, and TV applications, noting that 21% of the assessed methods involved art forms such as audio photography and interactive canvases. These studies predate the widespread adoption of AI-infused products and, consequently, could not address methods for evaluating such systems.

Rivero and Conte [18] identified a significant percentage of methods broad enough to assess any type of interface (33.9%), with specific attention to web (13.7%) and mobile (8.8%) applications. Pettersson et al. [19], instead, provided a more fragmented picture, ranking mobile apps (15%), interactive games (13%), web tools (12%), and websites (10%) as the most assessed systems, while only 4% of methods addressed connected/IoT devices.

Finally, specifically looking at the assessed qualities, generic UX is the most researched dimension reported by Bargas-Avila and Hornbæk [17], representing 41% of the reviewed techniques. The number even increases to 56% in the work of Pettersson et al. [19]. Both studies recognize that generic UX is a broad and overarching concept that is usually assessed in a qualitative and holistic way.

Pragmatic aspects, like usability, rank second in importance in Pettersson et al. [19], while they were not even considered as UX dimensions in Bargas-Avila and Hornbæk [17], possibly to reinforce the intended dissociation from usability studies at that time.

Emotion or affect, and enjoyment are also crucial dimensions of UX, frequently examined after general UX [17], and usability [19].

Challenges and Development Needs

While UX research and evaluation demonstrate their capability to evolve over time, there are a few directions for improvement to address the latest challenges.

Among the gaps highlighted by the analysed studies and relevant for the present investigation, UX evaluation in the early phases of product development is a patent one [16, 19]. Both research works point out the necessity for innovative approaches that can reliably assess user experience during these initial stages as traditional UX evaluation methods often fall short in providing actionable insights at this phase.

Unexpectedly, also the lack of rich qualitative data, offering a comprehensive multidimensionality of UX qualities has been emphasized [17, 18]. The authors underline the importance of going beyond the assessment of the general UX to acquire a deeper understanding and purposeful data to inform design interventions and iterations.

Another critical area that requires attention is the validation of measures for UX constructs. The reliability and validity of measure-based methods are frequently questioned, especially considering the flourishing of self-developed evaluation methods tailored to specific purposes [16, 19]. This leads to the spreading of non-reusable methods that might be a necessary quality for UX assessment, but still produce non-comparable or generalizable results.

As anticipated, longitudinal studies may benefit from further exploration [17, 18]. Indeed designers and developers might infer valuable information from the assessment of products and services over a long time and, eventually, tracking how the quality of the experience may evolve [20].

An effective multi-method approach is another area that requires further development. Combining different UX evaluation methods can provide a more comprehensive understanding of user experiences, but identifying which methods work well together and how to effectively integrate and analyze data from multiple sources remains challenging. Again, both Vermeeren et al. [16] and Pettersson et al. [19] suggest that further research is needed to identify guidelines and practical examples for their implementation.

Furthermore, the latest study [19] stresses the increasing importance of multi-device and multi-user experiences which necessitate the development of evaluation strategies that encompass their inherent complexities. They highlight that emerging technologies in people's daily life—such as IoT and machine learning—present new opportunities and challenges for UX evaluation, requiring methods that can adapt to these evolving contexts.

3.4.2 UX Evaluation Methods Scoping Review

Depicting Common Formats and Approaches

The systematic review of current UX evaluation methods culminated in the creation of a detailed table, collecting all the information resulting from the critical analysis of the 129 UX evaluation methods. The complete table is freely accessible on Figshare [21], while Fig. 3.1 synthesizes the most relevant insights.

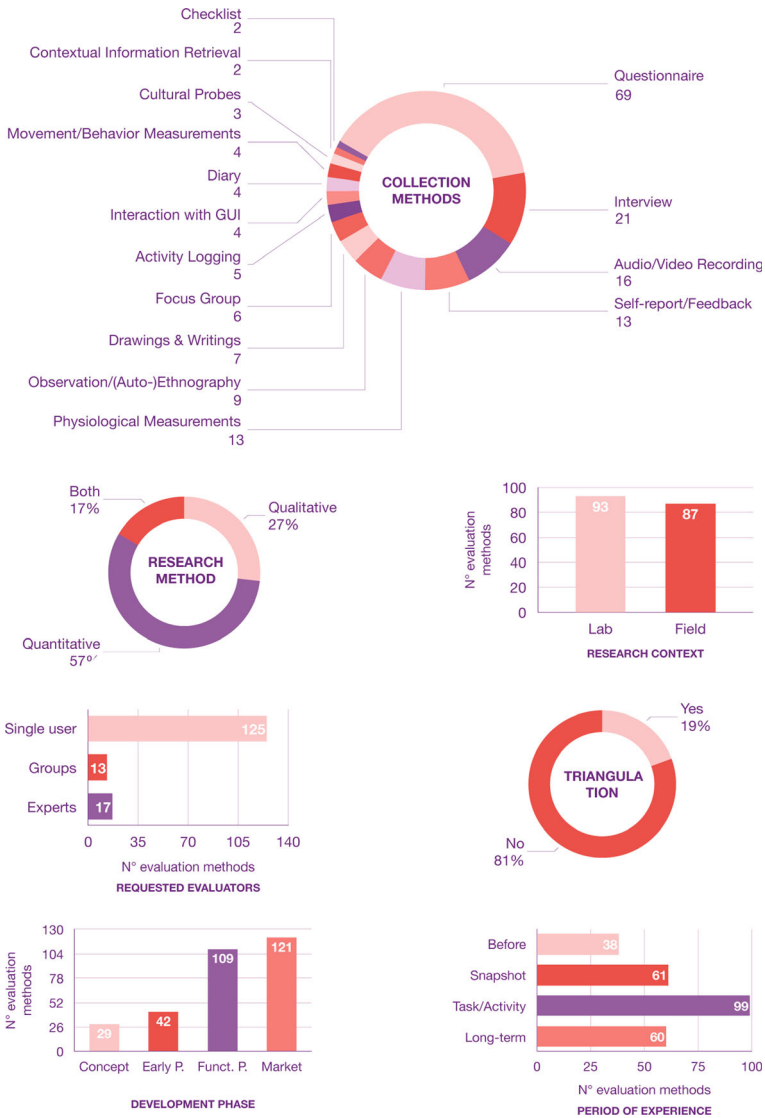


Fig. 3.1 Synthetic overview of the UX evaluation methods analysis

The most readily quantifiable findings pertain to the formats used for assessment methods. Unsurprisingly, the most common is the questionnaire, being employed in 69 different methods. These questionnaires have been implemented in both conventional ways, such as questions and scales, and with original variations, including graphic, pictorial, and auditory versions, and even taking advantage of modern possibilities and human habits, like randomly appearing when people unlock their phones. Interviews (21 methods), video/audio recordings (16 methods), physiological measurements, and self-reports/feedback (13 methods) are also prominent. These methods generally reflect scientific evaluation approaches, whereas those rooted in the design and social sciences, such as diaries (4 methods) and cultural probes (3 methods), are less frequently used.

It is observable that the total number of collection methods surpasses the number of UX evaluation methods analysed. This is because 19% of the methods employ triangulation, gathering data through multiple means to bolster the reliability of experimental or qualitative studies. Qualitative approaches are a minority (27.6%) compared to quantitative methods (57%). However, mixed methods are interestingly spreading, and they represent 16.4%.

On more practical matters, among the support materials used to submit UX evaluation methods, computers, and mobile phones (and digital devices in general) are the preferred ones. They can reach greater capillarity and allow more straightforward and quicker processing of data collected through questionnaires, sensors, activity logs, and video/audio recordings. Sometimes, even custom software or apps are developed. The environment in which UX evaluation methods are submitted, instead, is relatively equally divided among the lab (93 cases) and in the actual contexts of use (87 occurrences).

Significantly, the majority does not have one specific object of study. Most methods are versatile, capable of being applied to various industrial products, systems, environments, and events. Only few methods are specifically tailored to niche interactive content, such as visual interfaces or video games.

Another relevant information is that the cases analysed are typically intended for individual non-expert users (125) who evaluate the artefacts under investigation after performing some tasks or activities (99 cases) when they are already at an advanced design level. Namely, 109 methods can be applied to functional prototypes, 121 to products already on the market. Early design phases seem to be underrepresented, confirming a lack of emphasis on early evaluations that could facilitate rapid and less expensive product iteration. Only 29 and 42 methods can be employed respectively at the concept and early prototyping phase. Again, the sum is higher than the total of examples because the same evaluation method can be submitted at different design stages.

Dimensions and Descriptors

The core ingredients of a UX evaluation method are the qualities of products and services that they aim to assess. It is important to note that the literature reveals no consensus on terminology, with terms often used interchangeably or without

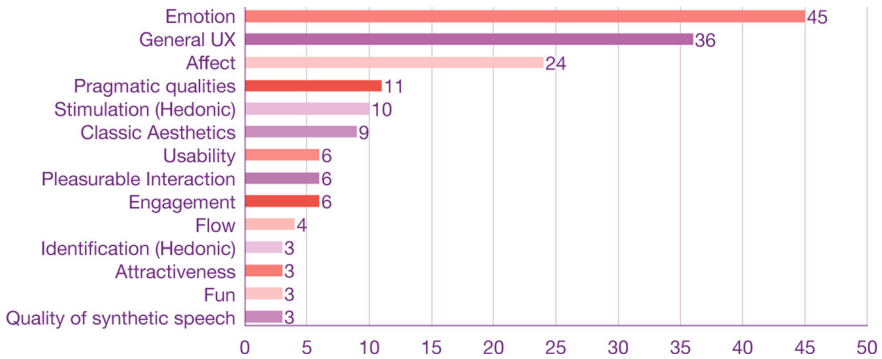


Fig. 3.2 Prevailing dimensions in the mapping of UX evaluation methods

regard for semantic nuances. Therefore, the same terms can be found indistinctly as dimensions and descriptors.

57 different UX evaluation dimensions emerged from the analysis of the 129 methods. A significant shift from previous studies emerged, with *emotion* and *affect* playing prominent roles as UX evaluation dimensions, appearing in 45 and 24 cases, respectively (Fig. 3.2). In consonance with the shift of the design field from the dominance of functionality and usability toward softer considerations such as pleasure [22], positive emotions [9], and aesthetics [8], this result underlines the importance that the intimate sphere and subjective perceptions play in every moment of people lives. Even if it contradicts the findings that previously mentioned studies have delivered, the prevalence of emotion over general UX (demoted to second place with 36 occurrences) may depend on the substantial number of methods that come from social sciences and psychology.

In third place, we find the *pragmatic qualities* that can be merged with *usability* and reach 17 occurrences. Another joint consideration can be made for *stimulation* (10) and *identification* (3), both expressions of the *hedonic* dimension, which, along with *aesthetic qualities* (9), complements the practical facets of an experience.

Considering the outcomes of the comprehensive work done, these four dimensions (*affective*, *pragmatic*, *hedonic*, and *aesthetic*) have been retained by the researchers as the most significant for building a holistic UX evaluation method that synthesises the legacy of UX evaluation. Indeed, *general UX* has been discarded because its excessively broad meaning entails a great hindrance to unambiguous measurability, while, to a critical examination, the following dimensions (*pleasurable interaction* (6), *engagement* (6), *flow* (4), *attractiveness* (3), *fun* (3), etc.) appear as subsets of the previous ones. Instead, the novelty of the *quality of synthetic speech* (3) among the dimensions deserves a peculiar remark. It demonstrates that methods dealing with the emergence of novel types of interaction modalities linked to AI systems were starting to be developed and that they need to introduce more precise attributes.

To deduce information from the multitude of descriptors collected, they have been firstly systematized according to their related dimensions. In this way, the most

recurrent ones could be easily identified and used to depict a nuanced portrait of their overarching dimension.

Once again, the affective component reveals the influence of psychology, mainly determined by *valence* and *arousal* [23] and further described by the most recognized basic emotions: *pleasure, fear, sadness, happiness, disgust, anger, and surprise*. For this reason, it risks overlapping with the hedonic dimension, primarily characterized by *enjoyability* and *excitement* as indices of pleasure of use. Additionally, *creativity, inventiveness, and innovativity* seem to have some relevance. Aesthetics presents a dual interpretation. On the one side, it can be referred to in terms of *appearance* (*clarity* and *sophistication* being two recurring themes). On the other, it is considered as the *attractiveness* of a product or service (frequently assessed as *good* and *pleasant*). Of course, this highlights the subjective nature of the concept. Finally, the pragmatic dimension involves *helpfulness, efficiency, and functionality*, as well as more user-friendly aspects like *easiness, simplicity, clearness, navigation, learnability, reliability, and convenience* to qualify the use of an artefact (Fig. 3.3).

Another relevant insight is provided by the number of UX dimensions covered by each method, that varies, with most methods addressing an average of 1.7 dimensions. This highlights their limitations in handling complex, multifaceted products. However, a few outliers, demonstrated a more holistic perspective by exploring multiple dimensions. Two of the most distant methods from the mean address conversational interfaces. Specifically, these are SASSI—Subjective Assessment of Speech System Interfaces, including 12 dimensions; the SUISQ—Speech User Interface Service Quality [24], considering 8 dimensions; the UEQ [25], with 6 dimensions; and the AttrakDiff [26], based on 4 dimensions.

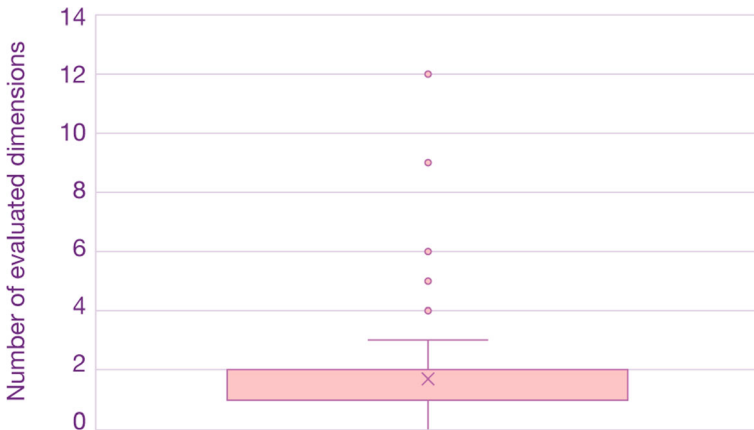


Fig. 3.3 Variability of UX dimensions assessed by each method

3.4.3 AI-Infused Systems Systematic Review

In the analysis of the 24 documents examined, 355 codes emerged. They were categorized in 10 dimensions (as shown in Table 3.2) and the same code might be included in multiple of them, based on their meaning as intended by the researchers or the interpretation of the authors.

Some reflect traditional UX qualities, while others are specific to AI-infused products. Instead, two categories collected attributes that can arguably be considered as descriptors for the scope of the project.

Of these, 11 were classified as not acceptable (NA), as they relate to aspects that go beyond UX (e.g., education, health), are too generic (e.g., user experience, user characteristics for implementation), or simply cannot be qualified as descriptors (e.g., abilities, instructions, evaluation framework) but were useful to determine the relevance of the document.

The second category, which elements were not considered suitable for the purposes of the research, focuses more on the users than on the evaluated products. Defined as “Users’ interaction/disposition”, it includes 37 properties that describe people’s expectations, intentions, beliefs and behaviour, or the context in which the interaction takes place (e.g., environmental factors, facilitating conditions). Of course, this kind of information is very helpful to obtain a comprehensive picture of the UX, including factors that are external to the product but can still impact the way people perceive and interact with it. However, considering the aim to identify qualities specifically relevant for AI-infused products, the researchers deemed these aspects to be out of scope. Nonetheless, they can be part of a larger research, triangulating different methods for deeper understanding.

Among the pertinent dimensions, four are recurring from the scoping review previously performed. Specifically, qualities in the *pragmatic* dimension are the most frequently mentioned (23 out of 24 documents) and reported the largest number of codes (149). Classic qualities, declined in different ways, are obviously present and

Table 3.2 Synthesis of the results from the systematic review of AI-infused systems evaluation

Dimensions	N. codes	Frequency in the documents
Pragmatic dimension	149	23/24
Trustworthiness	55	21/24
Meaningfulness	37	20/24
Users’ interaction/disposition	37	19/24
Affective dimension	36	16/24
Conversational dimension	30	16/24
Intelligence	22	15/24
Hedonic dimension	14	14/24
Non-acceptable (NA)	11	14/24
Aesthetic dimension	10	6/24

among the most recurrent—i.e., *usability* (in 16 papers), *accuracy* and *ease of use* (11), *efficiency* (10), *effectiveness* and *usefulness* (9), *learnability* (8), *control* and *understandability* (7). Yet, the attributes more closely addressing AI-based systems tend to be rather technical, unveiling that the major interest in the evaluation of these artefacts primarily comes from the fields of HCI and computer science. 57 out of 149 qualities are in this technical realm, which is arguably relevant for UX designers. They revolve around five main aspects: (i) *precision and recall*—including success and error rates, (ii) *information properties*—encompassing quality, quantity, coherence, variety and novelty, (iii) *processing and functionality*, (iv) *iteration counts*, and (v) *speed*.

The *affective* and *hedonic* dimensions are also present (respectively in 20 and 16 documents) with a relatively small amount of detected codes. Because of the wide spectrum of possible emotions, 36 descriptors have been identified in the *affective* dimension, and only 14 in the *hedonic* one. Yet, *satisfaction* is the most recurrent quality (13 documents) after *usability*.

Curiously, the *aesthetic* values of these artefacts are rarely considered (6 documents) with a total of 10 codes. When mentioned, they mostly refer to the digital user interface (UI) or relate to *human likeness*.

From the analysis, four new dimensions emerge, namely: trustworthiness, meaningfulness, conversational and intelligence.

Trustworthiness represents a major concern, second only to the *pragmatic* dimension. It recurs in 21 documents with a total of 55 descriptors. This category includes all aspects of trust in the system, primarily focusing on the ethical implications of being based on the employment of a significant amount of personal data. The importance of *trustworthiness* is also evidenced by the flourishing of ethical guidelines, of which those developed by the European Commission [27] are a primary reference for their completeness and articulation in seven essential requirements: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, (7) accountability. Thus, a product can be defined as trustworthy when acceptable and reliable from an individual and social perspective, representing a well-balanced compromise between human principles, practical needs, benefits and risks.

A similar picture emerges from the systematic analysis. The main clusters that can be observed—in order of recurrency—are *transparency* (9 documents), *acceptability* (6), *privacy and safety* (5), *legality* (4), *data concerns* (3), while *fairness*, *societal impact*, *risks*, and *faith* only appear in one document each.

Meaningfulness—or the property of products of showing a precise purpose, personal or shared significance, the generation of an experience, a symbol or a temporal quality—also stands out, appearing in 19 documents and with 37 related codes. It encompasses aspects related to the system's *helpfulness* or *usefulness* (8 documents), *value*—both in a broad sense (7) and in terms of business value (3)—*triggered interest* (5), *actual use* (4), and *sense-making* (1). While the topic of meaning in UX has been discussed at length, traditionally divided into two main strands, the cognitive [28] and the psychologically fulfilling [29] ones, recent frameworks

synthesize psychological meaning [30] and product properties [31], as it can also be observed in the presented results. Possibly, the manifested relevance of this dimension is connected to the necessity for AI-infused products to avoid being relegated to the dimension of gadget [32], as it is often happening.

Hints to the *conversational* dimension already appeared in the scoping review of UX evaluation methods, and it has seen an increase of studies on the subject. As not all AI-infused systems exploit NLP, a more limited number of documents, 15, includes this dimension, articulated in 30 codes. *Understanding* and *understandability* are among the most mentioned (12 documents), also because they are linked to multiple dimensions. Additionally, qualities related to the *naturalness* or *human likeness* (6), *dialogue* (5), *linguistic properties* (4), and *voice* (2) can be encountered.

Finally, a set of qualities were found to be uniquely referring to AI-based artefacts, and for this, they were clustered in the *intelligence* dimension. It counts 22 codes spread over 14 documents. Considering Russel and Norvig's [33] definition of AI and ML systems as agents perceiving and processing information from their environment and accordingly responding with the possibility—of ML agents—to improve over time, the attributes that more closely relate to this conception are: *context awareness* (recurring in 7 papers), *adaptability* (5), and *autonomy* (2), which lead to *proactivity* (2), *personalization* (4), *unpredictability* (1), and the manifestation of *reasoning* (2) and *social abilities* (2), inevitably intertwined with *human likeness* and *understanding*, possibly interpreted as *empathy* (2). Similar findings can be encountered in Amershi's argumentations [34], which can serve as a useful reference to better understand this dimension.

3.5 Discussion and Findings

3.5.1 RQ1—Are Current UX Assessment Methods Enough for AI-Infused Products?

All three research phases point at the insufficiency of current UX evaluation methods to properly tackle AI-infused products.

As the evolution of UX assessment methods illustrates, the field has shifted focus to embrace market and users' needs as they evolved over time. Undoubtedly, the new possibilities offered by AI systems mark a neat change in the way we can interact with the objects they empower, and this should have an impact on how we perceive, understand, and judge these artefacts.

As Pettersson et al. [19] highlighted, currently spreading technologies are bringing different challenges to the forefront, like multi-device and multi-user experiences. For this reason, they argue that existing methods are not equipped to handle the evolving contexts shaped by these technologies, necessitating the development of new evaluation strategies that can encompass the complexities of AI and ML among others.

Moreover, as final task in the researchers' analysis of the UX assessment methods collected in the scoping review, they rated the level of consistency of each method for the purposes of the study. This confirmed that, in their opinion, none of the evaluated methods is comprehensive enough or not too broad to adequately address AI-infused products. Indeed, most of them received an average score of 3 (on a scale from 1 to 5), with 31 methods scoring 4, but none was deemed fully suitable to assess products integrating AI systems. Yet, it is possible to consider the triangulation of a novel method specifically targeting AI-infused products with other consolidated ones, focusing on specific aspects that would still be relevant even if not peculiar or comprehensive enough for AI-enhanced possibilities. For instance, the SUS [35], AttrakDiff [36], Kansei [37] can be precious resources for pragmatic, hedonic, and affective aspects of the physical products, while Nielsen and Molich heuristics [3], Pyae and Joelsson [38] and Maguire [39] may respectively support the development of digital and conversational interfaces.

Finally, also the systematic review on the evaluation of AI-based systems reinforced the need for widening the borders of a method specifically designed to include AI-related features. Four dimensions emerged that are not covered by traditional UX methods, which directly leads to the next research question.

3.5.2 *RQ2—Are New UX Dimensions Needed for These Products?*

While the whole research hinted at confirming the need for additional UX dimensions to meet the purpose of the study, the systematic review of AI-based systems eventually confirmed this hypothesis. It resulted in almost 150 descriptors and four dimensions pointing at issues that are uniquely posed by AI-infused products and central in their characterisation, which is a sound reason to consider them for a proper evaluation.

Indeed, *trustworthiness* is related to all the concerns raised by how AI and ML systems are built and operate, generating, for example, unintelligible and unpredictable systems, deepfakes, or unfair outcomes. *Meaningfulness* becomes particularly relevant to AI-based artefacts as a consequence of the many products that clearly showcase the novel technical possibilities, without bringing any value or usefulness to their users. Additionally, understanding the meaning of engaging with other intelligences [40] is also an intriguing investigation. The *conversational* dimension encompasses all the aspects that the new way of interacting with machines, by just dialoguing with them in natural language, implies. *Intelligence*, instead, includes all the qualities that can be embedded and exploited in the UX that are only possible thanks to AI and ML systems, like making products aware of the context they are inserted in and adapting over time.

However, expanding UX research to new qualities and considerations does not imply that what already exists has to be neglected. On the contrary, well-affirmed

methods still prove their value in addressing their specific areas. Indeed, the systematic review demonstrated how pragmatic qualities still play a prominent role. Yet, they might have difficulties in accurately tackling AI-infused products as they cannot capture the core nature of these devices and their implications on the UX. An example might be considering the most frequent uses that people make of smart speakers, like Amazon Echo, and their actual potentialities [41]. While they are often employed for weather forecasting, listening to music, or setting alarms, like other more common products, they actually offer several features that users are not even aware of. This problem relates to both a traditional UX concern, discoverability, and proactivity, a feature that has become more easily implementable thanks to ML systems and is rarely or never considered when evaluating a product. If taken into consideration during the development and testing of this kind of artefacts, however, it can dramatically change the effectiveness of the interaction.

3.5.3 *RQ3—What Characteristics Should the New Method Have?*

In the light of the thorough research conducted, some considerations about the desirable or possible traits of the evaluation method to be developed can be advanced to guide the following steps of the project.

In accordance with Bargas-Avila and Hornbæk's indications [17], we articulate here the methodology to be used and the dimensions of the user experience, along with the object of the study.

Being the foundation of the entire research, the latter can be identified as AI-infused products, primarily considering physical artefacts in which AI or ML systems are integrated to improve or introduce new prospects in the UX. Of course, the threshold with digital products implementing similar features is blurry. However, for the sake of testing and validation purposes, we prefer to narrow this down to the less represented category of AI-infused products, aiming to capture specific characteristics as long as more generalisable ones.

This links to the first of the desired requirements we would like to pursue in our methodological approach. In the attempt to tackle the complexity and peculiarity of such systems, we can state three high-level traits of the evaluation method to be designed:

- **Flexible and comprehensive:** as AI systems can be integrated in a variety of contexts and with multiple possible interfaces, the evaluation method should adapt to different situations, possibly in a modular manner.
- **Prone to evolution:** operating in an ever-changing domain, it should be precise to depict the peculiarities of the object of the study as well as open enough to embrace future developments. To this end, it should also be able to cover the evolution of the experience with a product over time, to detect possible modifications.

- Capable to capture the essence of AI-infused products: through several stages of research, the method should pinpoint the core qualities that characterize and uniquely affect the UX with such artefacts.

Building on the points of attention identified and analysed in the scoping review of current UX evaluation methods, further details can be unfolded.

Collection Method

How users assess the product under study and the researchers can process the data is a crucial aspect of an evaluation method as it can determine its efficacy and actual adoption.

Our targeted recipients are UX researchers, designers or developers working for technology-driven companies producing artefacts that integrate AI, and we need to take into consideration their needs and expectations. As our research and the previously analysed studies testify, questionnaires are largely predominant in this field, implying their wide acceptance and recognition as a valid instrument. Considering that our aim is to propose new UX dimensions to assess, the familiarity of the collection method assumes an increased value. Moreover, questionnaires are usually straightforward for the respondents, have a more extensive range of possibilities to acquire a significant amount of answers for quantitative analysis and are easier to introduce in the design process as they do not take much time to obtain results. For these reasons, despite the collection formats derived by design and social sciences are usually more engaging and qualitatively rich ways for reporting on experiences, we would opt for the most common format, which also provides the opportunity for statistical validation to increase reliability.

Nature of Investigation

Being oriented toward a questionnaire as a collection method inevitably positions the evaluation method within the long-standing debate between quantitative and qualitative approaches. Indeed, it aligns with a quantitative one, which suits our target audience, particularly because those in business or engineering environments tend to associate numbers with objective truth, as noted by Cooper et al. [44]. However, while quantitative data provide measurable insights, they often fall short in delivering the deeper understanding needed to inform the design process effectively, a critical outcome we aim to achieve.

When considering the methods employed in other studies, it becomes clear that a mixed approach might be the preferable way. Although our investigation found a predominance of quantitative methods, the broader umbrella review presents contrasting findings. For instance, Vermeeren et al. [16] observed a relatively balanced distribution among methods that use purely quantitative, qualitative, or a combination of both data types. In contrast,argas-Avila and Hornbæk [17] identified a majority of qualitative approaches (50%) over quantitative ones (33%), with the remaining 17% using a mixed-methods approach. Meanwhile, Rivero and Conte [18] reported that 58% of techniques collected quantitative data, 14% focused only on qualitative data, and 28% combined both. The variation in these results could

be attributed to the different methodologies researchers used to select and analyse the available studies. However, these inconsistencies suggest that a single approach may not suffice, reinforcing the importance of triangulating different methods to enhance the reliability of the results, a trend highlighted by Vermeeren et al. [16] and further explored by Pettersson et al. [19].

Given these considerations, a potential direction for understanding the complex nature of the user experience generated by AI-enhanced artefacts is to complement a quantitative questionnaire with richer qualitative studies tailored to the specific needs of each situation. This approach aligns with findings from research focused on AI-infused devices, such as the study by Sciuto et al. [42] on Amazon Echo, which effectively employed a mixed-methods approach combining data logs with qualitative interviews.

Development Phase

Similarly to most evaluation methods, we expect ours to be applicable to functional prototypes and for products already on the market, with which users can easily interact and form a judgment.

However, as both Vermeeren et al.'s [16] and our investigation portrayed a lack of means to assess the projects at their conceptual or early stages of development, we would like to address this gap. Indeed, we aim to enable designers and developers to employ the information collected to improve the UX of their AI-infused products. Still, we acknowledge the particular difficulty represented by AI-infused products, as their actual UX is hard to prototype quickly, and the final versions might present unanticipated behaviours.

Evaluators and Period of Experience

At this point, providing details in this regard might be premature. Instead, we can outline some guidelines. Specifically, we will not narrow down the respondents to a set typology of users or a predetermined expertise level, as it might change based on the purposes of the study. Instead, we can envision that a little experience with the kind of product under evaluation is essential to being able to grasp their qualities. One of the reasons being the possible evolution of these systems' behaviours over a period of time.

Because of this, we would suggest that the evaluation of AI-infused products happens after a relatively long-term use, so that people's multiple and prolonged interactions can lead to reliable results. With respect to previous studies [18, 43], we can remark a positive increment in the number of evaluation methods covering larger periods of time, though they are not the majority. Yet, for the very nature of systems integrating ML algorithms—the most spread algorithmic approach to AI at the moment—we believe that a longer period of time is the only viable option to appreciate their capabilities of improving from experience.

Finally, the founding dimensions for the evaluation method—as derived by the presented research phases and extensively discussed—are eight. Four are the most frequent in current UX evaluation methods (*pragmatic, hedonic, affective,*

aesthetic), and the other four were found to be significant for AI-infused products (*trustworthiness, meaningfulness, conversational, intelligence*).

3.6 Conclusion

In exploring the current UX dimensions and their application to AI-infused products, this chapter has underscored the unsuitability of existing evaluation methods. The UX dimensions traditionally assessed remain relevant but insufficient to evaluate the emergent qualities introduced by AI. In fact, while current methods like SUS, AttrakDiff, and Nielsen's heuristics remain invaluable for evaluating certain aspects of the UX, they need to be supplemented or redefined to reflect the intricacies of AI. New, complementary dimensions are needed, and *trustworthiness, meaningfulness, conversational, and intelligence*, appear to be central to the characterisation of AI-infused products. However, these are either underrepresented or absent in existing UX methods, further emphasising the limitations of current evaluation practices.

Therefore, the findings presented in this chapter pave the way for developing a more comprehensive UX evaluation method that is specifically tailored to the unique challenges and qualities of AI-infused products, and should also be flexible and apt to change. This marks a necessary step forward in the evolution of UX assessment and sets the stage for further, more experimental, research into the development of this new method.

References

1. Sundar SS (2020) Rise of machine agency: a framework for studying the psychology of human–AI interaction (HAI). *J Comput-Mediat Commun* 25:74–88. <https://doi.org/10.1093/jcmc/zmz026>
2. Guzman AL, Lewis SC (2020) Artificial intelligence and communication: a human-machine communication research agenda. *New Media Soc* 22:70–86. <https://doi.org/10.1177/1461444819858691>
3. Nielsen J, Molich R (1990) Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 249–256
4. Polson PG, Lewis C, Rieman J, Wharton C (1992) Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *Int J Man-Mach Stud* 36:741–773. [https://doi.org/10.1016/0020-7373\(92\)90039-N](https://doi.org/10.1016/0020-7373(92)90039-N)
5. Beyer H, Holtzblatt K (1997) Contextual design: defining customer-centered systems. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
6. Spool JM (1999) Web site usability: a designer's guide. Morgan Kaufmann
7. Jordan P (2000) Designing pleasurable products: an introduction to the new human factors. Tayr and Francis, London
8. Tractinsky N, Katz AS, Ikar D (2000) What is beautiful is usable. *Interact Comput* 13:127–145. [https://doi.org/10.1016/S0953-5438\(00\)00031-X](https://doi.org/10.1016/S0953-5438(00)00031-X)
9. Norman DA (2004) Emotional design: why we love (or hate) everyday things. Basic Books, New York

10. Desmet PMA, Hekkert P (2007) Framework of product experience
11. Hassenzahl M, Burmester M, Koller F (2021) User experience is all there is: twenty years of designing positive experiences and meaningful technology. *i-Com* 20:197–213. <https://doi.org/10.1515/icom-2021-0034>
12. Burmester M, Hassenzahl M, Koller F (2002) Usability ist nicht alles – Wege zu attraktiven Produkten (Beyond usability – appeal of interactive products). *i-Com* 1:32–40. <https://doi.org/10.1524/icom.2002.1.1.032>
13. Hassenzahl M, Tractinsky N (2006) User experience—a research agenda. *Behav Inf Technol* 25:91–97
14. Spallazzo D, Ajovalasit M, Ceconello M, et al (2021) Assessment of descriptors for UX evaluation of AI-infused products. 124653 Bytes
15. Saldaña J (2009) *The coding manual for qualitative researchers*. Sage, Los Angeles, Calif
16. Vermeeren APOS, Law EL-C, Roto V, et al (2010) User experience evaluation methods: current state and development needs. In: *Proceedings of the 6th Nordic conference on human-computer interaction: extending boundaries*. Association for Computing Machinery, Reykjavik, Iceland, pp 521–530
17. Bargas-Avila JA, Hornbæk K (2011) Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, USA, pp 2689–2698
18. Rivero L, Conte T (2017) A systematic mapping study on research contributions on UX evaluation technologies. In: *Proceedings of the XVI Brazilian symposium on human factors in computing systems*. Association for Computing Machinery, Joinville, Brazil, pp 1–10
19. Pettersson I, Lachner F, Frison A-K, et al (2018) A Bermuda triangle? A review of method application and triangulation in user experience evaluation. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. Association for Computing Machinery, Montreal QC, Canada, pp 1–16
20. Karapanos E, Zimmerman J, Forlizzi J, Martens J-B (2009) User experience over time: an initial framework. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, Boston, MA, USA, pp 729–738
21. Spallazzo D, Sciannamè M, Ajovalasit M, et al (2021) UX evaluation methods mapping. 100905 Bytes
22. Jordan PW (2000) *Designing pleasurable products: an introduction to the new human factors*. CRC Press
23. Russell JA (1980) A circumplex model of affect. *J Pers Soc Psychol* 39:1161–1178. <https://doi.org/10.1037/h0077714>
24. Polkosky MD, Lewis JR (2003) Expanding the MOS: development and psychometric evaluation of the MOS-R and MOS-X. *Int J Speech Technol* 6:161–182. <https://doi.org/10.1023/A:1022390615396>
25. Polkosky MD (2005) *Toward a social-cognitive psychology of speech technology: affective responses to speech-based e-Service*, PhD thesis
26. Laugwitz B, Held T, Schrepp M (2008) Construction and evaluation of a user experience questionnaire. In: *Holzinger A (ed) HCI and usability for education and work*. Springer, Berlin, Heidelberg, pp 63–76
27. Hleg AI (2019) *Ethics guidelines for trustworthy AI*. European Commission, Brussels
28. Dourish P (2001) *Where the action is: the foundations of embodied interaction*. The MIT Press, Cambridge, Mass
29. Hassenzahl M, Eckoldt K, Diefenbach S et al (2013) Designing moments of meaning and pleasure. *Experience design and happiness*. *Int J Des* 7:21–31
30. Mekler ED, Hornbæk K (2019) A framework for the experience of meaning in human-computer interaction. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. Association for Computing Machinery, Glasgow, Scotland Uk, pp 1–15
31. Ajovalasit M, Giacomini J (2019) Meaning of artefacts: interpretations can differ between designers and consumers. In: *Conference proceedings of the academy for design innovation management*, pp 1178–1188

32. Levinson P (1977) Toy, mirror, and art: the metamorphosis of technological culture. *ETC Rev Gen Semant* 34:151–167
33. Russell S, Norvig P (2020) *Artificial intelligence: a modern approach*, 4th edn. Pearson, Hoboken, N.J.
34. Amershi S, Weld D, Vorvoreanu M, et al (2019) Guidelines for human-AI interaction. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. Association for Computing Machinery, Glasgow, Scotland Uk, pp 1–13
35. Brooke J (1995) SUS: a quick and dirty usability scale. *Usability Eval Ind* 189:
36. Hassenzahl M, Marc, Burmester M, et al (2003) *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität*
37. Schütte S, Ayas E, Schütte R, et al (2006) Developing software tools for Kansei engineering processes: Kansei engineering software (KESo) and a design support system based on genetic algorithm. In: *9th International quality management for organizational development (QMOD) conference*, August 9–11, Liverpool, England
38. Pyae A, Joelsson TN (2018) Investigating the usability and user experiences of voice user interface: a case of Google home smart speaker. In: *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services adjunct*. Association for Computing Machinery, Barcelona, Spain, pp 127–131
39. Maguire M (2019) Development of a heuristic evaluation tool for voice user interfaces. In: Marcus A, Wang W (eds) *Design, user experience, and usability. Practice and case studies*. Springer International Publishing, Cham, pp 212–225
40. Hassenzahl M, Borchers J, Boll S, et al (2020) Otherware: how to best interact with autonomous systems. *Interactions* 28:54–57. <https://doi.org/10.1145/3436942>
41. White RW (2018) Skill discovery in virtual assistants. *Commun ACM* 61:106–113
42. Sciuto A, Saini A, Forlizzi J, Hong JI (2018) “Hey Alexa, what’s up?”: a mixed-methods studies of in-home conversational agent usage. In: *Proceedings of the 2018 designing interactive systems conference*. ACM, New York, pp 857–868
43. Vermeeren A, Roto V, Väänänen K (2016) Design inclusive UX research: design as a part of doing user experience research. *Behav Inf Technol* 35:21–37. <https://doi.org/10.1080/0144929X.2015.1081292>
44. Cooper A, Reimann R, Cronin D, Cooper A (2014) *About face: the essentials of interaction design*, Fourth edition. John Wiley and Sons, Indianapolis, IN
45. Lachner F, Naegelein P, Kowalski R, et al (2016) Quantified UX: towards a common organizational understanding of user experience. In: *Proceedings of the 9th nordic conference on human-computer interaction*. ACM, Gothenburg Sweden, pp 1–10

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4

Unpacking AI-Infused Systems Qualities. Building a UX Evaluation Method



Abstract The chapter illustrates the results of an enquiry aiming to systematize the qualities of AI-infused devices and to propose a specific method to assess the UX they entail. For the purpose, advanced users were involved in a semi structured questionnaire aimed to understand their perception of AI-infused products and the related UX qualities they care the most. Pragmatic, aesthetic, hedonic, affective, intelligence, trustworthiness, conversational, and meaningfulness dimensions have been proposed as a starting point and the less traditional have garnered particular consensus.

4.1 Actively Exploring the Qualities of AI-Infused Products. Overview of the Research Methods

As Chap. 2 highlights, being able to properly describe AI-infused products, in a way that embraces their complexity and unique characteristics, is an indispensable factor to making sense of this kind of objects and services.

This is the reason why Meet-AI researchers drew the state of the art of the dimensions and descriptors employed by current qualitative and quantitative UX evaluation methods through a wide-ranged critical analysis and started to identify the peculiar features of AI-infused systems through systematic literature review on the topic (as presented in Chap. 3). What emerged is a set of eight overarching dimensions, informed both by the parameters traditionally used in UX assessment practices and by the rising behaviours and gestures enabled by AI capabilities. These are *pragmatic, aesthetic, hedonic, affective, intelligence, trustworthiness, conversational, and meaningfulness* dimensions.

This chapter aims to further the research on determining the focal qualities to assess products embedding AI, to include multiple perspectives in the process of defining the building blocks of a UX evaluation method specifically intended for this type of artefacts. In particular, it introduces expert users' involvement as a co-creation approach to guide the selection of dimensions and descriptors suitable for AI-infused systems. This method also allowed to keep as much objectivity as possible,

by complementing the researchers' inferences from the comparative analysis of UX evaluation methods and the systematic review.

To fulfil this requirement, a protocol integrating mixed methodologies and multiple phases was developed.

In Phase 0, the study is driven by a survey designed to test the hypotheses regarding the dimensions to describe AI-infused products and to solicit novel contributions on descriptors to expand the non-comprehensive list gathered from the literature review. While the survey yielded immediate findings to frame the most relevant dimensions (Phase 1), Phase 2 required additional analysis to derive useful descriptors from the responses to the survey. Two of the researchers independently performed a preliminary homologation of the suggested features of AI-infused artefacts, which they then compared to produce a shared list. Then, an affinity map was used to synthesise repetitions and filter out out-of-context responses from the list of descriptors. The resulting set of descriptors was shared with the other Meet-AI researchers for an intercoder evaluation, with the goal of determining the descriptors' consistency with the corresponding dimension, their relevance for AI-infused products, and lastly detecting the most significant ones. Summing up the results of the entire investigation as a necessary step to establish the foundations of the evaluation scale, Phase 3 consisted of a workshop internal to the Meet-AI research group to outline the elements (dimensions and descriptors) from which to build its structure.

In the following, the different phases are presented with more thorough methodological details and along with the outcomes that each activity brought to light.

4.2 Phase 0: Broadening the Boundaries of AI-Related Qualities Through a Survey

Building on the findings of the previous investigations—the set of eight AI-related dimensions—a digital survey was developed and submitted to a group of advanced users to further broaden the range of traits that may be used to describe the target items.

The study aimed to include a population of 110 students from the MSc in Digital and Interaction Design programme and 47 young researchers from Politecnico di Milano—Design Department, all of whom shared two essential characteristics: they had to be familiar with the products being observed and to have a developed sensitivity and understanding of the design of interactive objects. A total of 42 people responded, for a response rate of 26.75%.

The survey was particularly intended to be as much straightforward and transparent as possible, as explicitly mentioned in the beginning of the created Google form. For this reason, it opened with the description of some examples of AI-based devices (namely, smart speakers, learning thermostats, and smart cams) both to provide examples of the artefacts to be addressed and to measure respondents' level of familiarity with the illustrated products.

The core exploratory section followed, and it developed in a twofold direction. First, it aimed to collect a new set of descriptors based on the predefined dimensions to describe AI-infused systems. These were adequately explained to establish a common understanding, and then respondents were prompted to suggest new UX attributes as portrayed in Table 4.1. Some questions demanded the indication of at least three attributes (with no further indications), the more challenging ones required a minimum of two positive and two negative features to encourage more varied answers.

The second objective shifted the focus on the UX dimensions themselves, pointing at obtaining feedback about their relevance. In particular, the researchers defined three main enquiries that engaged respondents in a critical analysis: (i) assess how well they perform in the evaluation of AI-infused products, (ii) understand which are considered the most relevant, and (iii) identify if some essential ones are missing. Direct questions were used to acquire these pieces of information and generate clear data.

Finalising the path, the questionnaire closed with a profiling section collecting basic information on the respondents for statistical purposes.

4.3 Phase 1: UX Dimensions of AI-Infused Products According to Advanced Users

As a first step in outlining the structure of the evaluation method, this paragraph reasons about the inferences that can be drawn from the survey responses about the proposed UX dimensions for AI-infused systems. Subsequently, the discourse will deal with the finer grained level of the descriptors that the respondents attributed to each of them.

Overall, the encompassing qualities proposed in the survey gained the favour of the respondents, as the positive evaluations depicted in Fig. 4.1 demonstrate.

An encouraging result confirmed the researchers' assumptions derived from the previous reviews: the advanced users, in fact, underlined the consistency that *trustworthiness*, *intelligence*, *conversational*, and *meaningfulness* have with the target products, by expressing a solid consensus on their relevance.

While remaining pertinent, instead, *pragmatic*, *aesthetic*, *hedonic* and *affective* dimensions have been mainly marked as "important". This corroborates the fact that current UX evaluation methods (from which those are attained) are not specifically equipped to handle artefacts embedding AI capabilities, as the Meet-AI project states in its premises.

In line with these findings, the second question (Fig. 4.2), encouraging the respondents to select what they felt were the most relevant UX dimensions, revealed an analogous preference for *trustworthiness* (identified among the most appropriate by 76% of respondents), followed by *conversational* (59.5%), *intelligence* (50%), and *meaningfulness* (40.5%).

Table 4.1 Synthesis of the descriptors requests as they appeared in the survey

Dimension	Description	Question
Pragmatic dimension	Some qualities of products support users in achieving their concrete goals, such as performing specific tasks. They may include (but are not limited to) usability, intelligibility, efficacy issues	Please write at least three attributes (adjectives, nouns, verbs) you consider peculiar and relevant to describe the quality of use of AI-infused products
Aesthetic dimension	The aesthetic appearance of industrial products plays an essential role in our relationship with them. Despite being subjective, the appreciation of beauty may be affected by different aspects (e.g., shape, colour, material, finishing, behaviour, etc.)	Please write at least three attributes (adjectives, nouns, verbs) you consider the most relevant and unique to describe the aesthetic qualities of AI-infused products
Hedonic dimension	Some qualities of products can make them attractive and engaging, and arise pleasant and satisfying sensations during use	Thinking specifically of AI-infused products, please write at least three essential qualities (adjectives, nouns, verbs) that characterize them as pleasurable and attractive
Affective dimension	While interacting with products, they often influence our emotional state by inducing subjective feelings. This can be particularly relevant with AI-infused products	List a minimum of 2 positive and 2 negative affective responses you consider typically caused by AI-infused products
Trustworthiness	A product can be defined as trustworthy when it is individually and socially acceptable and reliable, and it represents a well balanced trade-off between human principles and practical needs, benefits and risks	Envisioning the possible positive and negative impacts of AI-infused products, write at least 2 essential features for them to be trustworthy and at least 2 unreliable
Conversational dimension	Some AI-infused products like smart speakers (Amazon Echo, Google Home...) can use voice and text to interact with users. Voice can be used to do tasks, answer questions, control other products, and engage in conversation. A “conversational” product or system is able to use natural language in an interaction that lasts multiple turns of dialogue	Reflecting on the most impactful features in the design and use of conversational systems, write at least 2 features (adjectives, nouns, verbs) that contribute to creating a positive and efficient interaction, and at least 2 features that may ruin the experience

(continued)

Table 4.1 (continued)

Dimension	Description	Question
Intelligence	AI-infused products can autonomously learn to adapt their behaviour over time, and can proactively take action or propose suggestions to their users	Write at least 2 relevant features (adjectives, nouns, verbs) an AI-infused product should have to be considered intelligent, and at least 2 features that lessen the perception of intelligence
Meaningfulness	Some aspects of products can make them meaningful to their users in the sense that they may manifest a tangible purpose, a personal significance, a shared/cultural significance, generate past experience, communicate a symbol or exhibit a temporal quality	Thinking specifically to AI-infused products, please write at least three attributes (adjectives, nouns, verbs) that make you perceive AI-infused products as meaningful

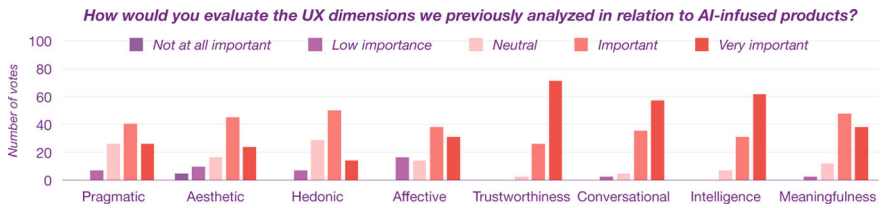


Fig. 4.1 Survey results on the evaluation of the proposed UX dimensions for AI-infused products

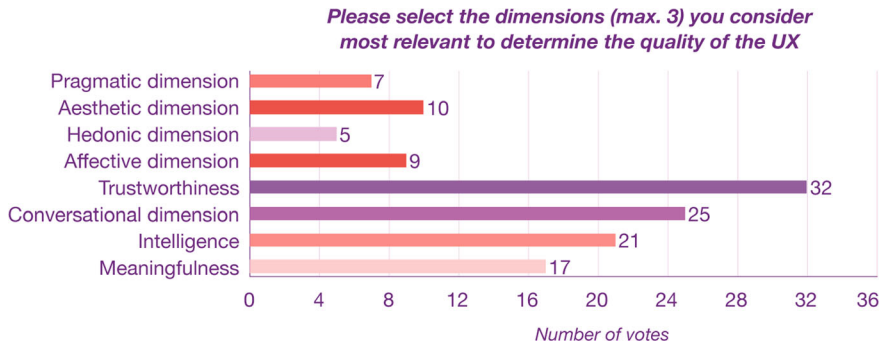


Fig. 4.2 Survey results highlighting the most relevant UX dimensions for AI-infused products

The third request, on the other hand, received no helpful responses. It directly solicited the respondents to provide additional dimensions that could be significant to achieve a more precise UX evaluation scale for AI-infused systems, but the few comments received either reinforced the previous selections, contained qualities that would be better classified as descriptors, or were off-topic.

However, the above outcomes and the researchers' initial hypothesis may be confirmed by examining additional parameters of the survey. Certainly significant is the content and style of responses providing descriptors for each of the dimensions. Indeed, they can reveal how people perceive and understand the proposed dimensions based on their level of coherence, appropriateness, and personal contribution.

For instance, dimensions like the *hedonic*, *affective* and *meaningfulness* proved quite difficult for the respondents, whose answers revealed respectively high subjectivity, shortcomings, and both flaws combined. In these categories, the responses were mostly long-winded—despite the request to elaborate one-word attributes—and inconsistent. In fact, some of the valuable traits that emerged here could actually be better applicable to other dimensions. The ultimate demonstration of the complexity of the concept of meaning emerged also in the open statement of some advanced users, who admitted that they were not able to provide any answer.

However, other dimensions, such as the *pragmatic* one, resulted more straightforward and familiar to handle.

Some were characterised by rich responses, both in quantitative terms (it is the case of *conversational* and *intelligence* dimensions), and because of their clear articulation (*trustworthiness*). Additionally, the identification of attributes in these classes perfectly fit their overarching qualities and the subjects of the study, and traces of their appreciation and perceived relevance could be found throughout the questionnaire, as related concepts were redundant.

The only dimension with poor-quality data was *aesthetics*. The low perceived relevance in regard to the purposes of the investigation was patent. It was already explicitly marked in the related question and reinforced by answers that address specific features of the products on the market and hint at a certain superficiality.

4.4 Phase 2: Insights from an Intertwined Analysis of AI-Related Descriptors

Once the main dimensions to portray AI-infused systems have been identified, the second step in building a proper method consisted of going into more detail, to understand all the facets within the overarching qualities to generate a comprehensive view.

The collection of attributes prompted by the survey was intended for this purpose, but some analysis was needed to make the information actionable, involving Meet-AI researchers in the preparation of the raw data and in an assessment activity.

To begin, two of the researchers redacted a homogenous list, translating sentences and Italian responses into single English words in order to ensure that the survey results were in line with the original request. The resulting one-word descriptor lists were then compared to create a uniform one [1].

and they have been analysed by computing the mean and z score for each descriptor and according to the two parameters (consistency and relevance). They have been then segmented into quartiles to make it easier to detect the most significant. Finally, by comparing the relevance z scores of all the descriptors, a full overview (Table 4.2) was produced. Similarly, dividing these into quartiles, it emerges that the section >75% contains 134 descriptors, a too large number to be used as the basis for the scale. For this reason, it has been considered more reasonable to highlight and further operationalise the 36 among them that unanimously received the highest overall score (the “golden” ones). Some educated assumptions might be drawn after this processing—which is depicted in Table 4.3—and they are here explored and discussed according to their overarching dimensions.

Pragmatic Dimension Consistency characterises the responses in this category, which is probably indicative of the involved advanced users’ familiarity with it. In fact, one-word attributes have been suggested, in accordance with the request, and the descriptors marked one of the highest overall consistency scores in the intercoder assessment. A total of 134 items have been submitted for the pragmatic dimension, from which 46 descriptors emerged after the synthesising work and only two have been discarded before reaching the judges.

Coherence also with the features appearing in literature has meant that no major novelties have occurred, but some new elements directly related to AI-infused systems have been highlighted. It is the case of *smartness*, *customisation*, *responsiveness*, *adaptability*, *connectivity*, *unobtrusiveness*, and different concepts linked to *trustworthiness*.

In terms of relevance, it received the second-highest score in both the mean of evaluations and the overall “golden” descriptors, likely indicating that this basic dimension for evaluating UX is still significant or, at the very least, that respondents attributed the majority of the relevant characteristics of AI-infused systems to this dimension.

Aesthetic Dimension As anticipated, the responses in this category were profoundly influenced by the practices currently adopted in the industry of AI-infused products. Unequivocally, a lot of the items directly referred to the specific features that can be found on the market (e.g., *white colour*, *small size*, *rounded shapes*, etc.), instead of indicating broader parameters to describe the aesthetics of an object. This is why, a great effort has been necessary to generalise them to compile an adequate list for the judges. Nonetheless, despite the revision, the descriptors reported in this dimension attained the lowest scores in consistency and relevance (with an insufficient mean of 1.76 out of 4), only a few reached the >75% quartile for the overall relevance (the smallest number among the dimensions), and no one appears in the “golden” list.

To a qualitative look, though, one descriptor stands out with a relevance mean of 3.8: *personality*. What is interesting is the fact that this quality quite diverges from the most traditional conception of aesthetics to embrace studies in the perception of products. Analogously curious is the next descriptor of this class in terms of importance: *mimesis*, which underlines the relationship between UX and the field of

Table 4.2 List of the “golden” descriptors with the related dimensions

Source	Descriptor	R1 EV	R2 EV	R3 EV	R4 EV	R5 EV	R6 EV
CONV-L	Voice naturalness	4	4	4	4	4	4
CONV-L	Voice pleasantness	4	4	4	4	4	4
CONV-Q	Accuracy	4	4	4	4	4	4
CONV-Q	Context awareness	4	4	4	4	4	4
CONV-Q	NLP quality	4	4	4	4	4	4
CONV-Q	Reliability	4	4	4	4	4	4
CONV-Q	Understanding	4	4	4	4	4	4
HED-Q	Empathy	4	4	4	4	4	4
INT-Q	Accuracy	4	4	4	4	4	4
INT-Q	Empathy	4	4	4	4	4	4
INT-Q	Context awareness	4	4	4	4	4	4
INT-Q	Understanding	4	4	4	4	4	4
MEAN-Q	Usefulness	4	4	4	4	4	4
PRAG-L	Functionality	4	4	4	4	4	4
PRAG-L	Helpfulness	4	4	4	4	4	4
PRAG-L	Intelligibility	4	4	4	4	4	4
PRAG-L	Intuitivity	4	4	4	4	4	4
PRAG-L	Learnability	4	4	4	4	4	4
PRAG-L	Reliability	4	4	4	4	4	4
PRAG-L	Understandability	4	4	4	4	4	4
PRAG-Q	Customization	4	4	4	4	4	4
PRAG-Q	Ease of use	4	4	4	4	4	4
PRAG-Q	Transparency	4	4	4	4	4	4
PRAG-Q	Trustworthiness	4	4	4	4	4	4
TRUS-L	Access to data	4	4	4	4	4	4
TRUS-L	Human oversight	4	4	4	4	4	4
TRUS-L	Non-discrimination	4	4	4	4	4	4
TRUS-L	Privacy	4	4	4	4	4	4
TRUS-L	Quality of data	4	4	4	4	4	4
TRUS-L	Transparency	4	4	4	4	4	4
TRUS-L	Unfair bias avoidance	4	4	4	4	4	4
TRUS-Q	Accuracy	4	4	4	4	4	4
TRUS-Q	Data management	4	4	4	4	4	4
TRUS-Q	Data protection	4	4	4	4	4	4
TRUS-Q	Reliability	4	4	4	4	4	4
TRUS-Q	Transparency	4	4	4	4	4	4

Table 4.3 Descriptors performances, synthesised according to the related dimension, in the various steps of the analysis

Dimension	Submitted items	Resulting descriptors	Excluded items (S)	Mean descriptors consistency	Mean descriptors relevance (S)	Mean descriptors relevance (L)	Golden descriptors
Pragmatic dimension	136 (S) + 49 (L)	46 (S) + 49 (L)	2	2.53	2.67	2.68	11
Aesthetic dimension	132 (S) + 43 (L)	37 (S) + 30 (L)	1	2.03	1.76	1.92	0
Hedonic dimension	133 (S) + 32 (L)	55 (S) + 30 (L)	2	2.00	2.05	2.31	1
Affective dimension	158 (S) + 219 (L)	49 (S) + 96 (L)	52	2.69	2.47	1.64	0
Trustworthiness	140 (S) + 41 (L)	39 (S) + 41 (L)	2	2.50	2.62	3.04	12
Conversational dimension	161 (S) + 28 (L)	60 (S) + 22 (L)	1	2.64	2.70	2.93	7
Intelligence	141 (S)	53 (S)	0	2.19	2.30	/	4
Meaningfulness	115 (S)	49 (S)	2	2.24	2.39	/	1

(S) from survey, (L) from literature

ubiquitous computing. It received a mean of 3, but also occurred in the nuances of *invisibility* and *unobtrusiveness*.

Hedonic Dimension Seemingly promising in terms of descriptors collected from the survey—it has the second larger number (55)—this dimension is among the most delusional for its performance, recording slightly sufficient consistency and relevance scores, as reflected in the overall ranking of relevant descriptors, where the hedonic ones play a little role.

However, some qualities that can be directly related to AI emerged and two have been judged particularly noteworthy: *empathy*, also appearing among the “golden” descriptors, and *adaptability*, which is immediately behind with a 3.8 mean. Yet, both are present in 6 out of 8 dimensions, which reveals the ambiguity, possibly connected to the subjectivity, of the descriptors in this category. This common issue is also manifest in the judges’ evaluation of other features (e.g., *multifunctionality*, *responsiveness*, *voice interaction*), which have been deemed more appropriate for other dimensions or just not particularly significant.

Affective Dimension The responses in the affective sphere of the UX assessment brought to light great confusion in the advanced users, an insight utterly in antithesis with the pervasive role this dimension has in current UX evaluation methods (from which 96 descriptors are drawn and 219 occurrences are counted). Even though only single words were requested, the answers presented a great amount of articulated sentences. They pointed to the cause of emotional states instead of explaining the affective responses themselves, which are evident symptoms of the difficulty of correctly expressing one’s feelings. From these premises, it is not surprising that this dimension reported the highest number of items excluded for unequivocal inconsistency even before the judges’ evaluation, but it is impressive that their number is around 1/3 of the total submissions.

In the end, the judges evaluated the remaining descriptors quite positively and they were found to be more consistent with their related dimension. Moreover, some truly relevant qualities for AI-infused products deserve to be mentioned, such as the more general and empowering *feeling in control* and *feeling understood*, or those arising from the direct interaction with such devices: *attraction*, *challenge*, *disappointment*, *frustration*, and *satisfaction*. Nonetheless, in the overall ranking of relevant descriptors, the affective dimension marks an unsatisfying second place (following only the aesthetic one) for having the least representation in the highest quartile, with no “golden” items as well. Additionally, to further prove the fact that this category has not been judged valuable to describe the UX of AI-infused systems, also the commonly assessed qualities from the literature received a very low rating, with a negative-scented mean of 1.64 out of 4.

Trustworthiness Like the previous one, also the responses related to this dimension were well-articulated. Yet, they had a very different connotation, as they generally expressed a desire for a better explanation rather than a misunderstanding or difficulty in answering. Additionally, items reflecting an attention to ethical and other trust-related concerns pervaded all other dimensions indiscriminately. Both hints

indicate how prominent this matter was perceived by respondents, even before they encountered the related question in the survey. Overall, these qualitative considerations, along with more quantitative inferences, support that *trustworthiness* is highly relevant when dealing with AI-infused systems. In fact, despite being a not-so-easy topic for advanced users with no formal education in this, the number of descriptors gathered from the questionnaire was quite high, and their evaluations excelled on all criteria. Indeed, they make up one-third of the list of “golden” descriptors, making this dimension the one contributing the most to the highest quartile in the overall relevance ranking. To give some examples of the punctual specificity the descriptors here presented, the most accredited ones were *accuracy*, *data management*, *data protection*, *reliability*, and *transparency*, which somehow echo the European guidelines for trustworthy AI [4].

Conversational Dimension It has been the most prolific dimension in absolute terms, reporting 60 descriptors and 160 submitted items, and its significance in the respondents’ opinion is reinforced by the quality of the suggestions. In fact, even if they do not really match the terminology found in the literature, they give precise and granular information. Some are very specific and almost technical in describing the characteristics that require consideration due to the introduction of AI capabilities. It is the case of *NLP quality*, *accent and dialect recognition*, *voice quality*, *character*, etc. Others, instead, might also be applied to a broader set of behaviours of the systems, like *accuracy*, *context awareness*, *understanding*, *feedback quality*, but also *fluidity* and *naturalness*.

Either way, also the intercoder assessment reveals the peculiar role that conversational interactions might have in defining products and services integrating AI. Consistency and relevance rates confirm high performances of the descriptors, many of which reached the top section of the overall relevance ranking and are included in the “golden” shortlist, making this dimension the third force—after *trustworthiness* and *pragmatic*—for number of representatives.

Intelligence Defining intelligence is undoubtedly a challenging task that has been subject of interrogation and discussion in different disciplines. For this reason, it was not a foregone conclusion to receive consistent responses. Maybe unexpectedly, though, the advanced users involved in the study seemed to have clear ideas about the traits that can characterize a perceived intelligent behaviour in AI-infused systems, providing quite satisfactory suggestions. All of the 141 submitted items have been synthesised in the 53 descriptors that reached the judges for their evaluation. Indeed, it has been the only case in which no items needed to be discarded.

Their performance was also quite good, receiving evaluations in both consistency and relevance parameters that positioned them in an average place among the descriptors of all the dimensions, as testified also by the overall relevance ranking. Some of the most peculiar features of this dimension—namely *accuracy*, *adaptability*, *context awareness* and *understanding*—have been quite pervasive in all the survey, but here is where the judges considered that they were more appropriate. Moreover, it is curious how already the elite descriptors reveal the classic dualism underlying the history of AI. In fact, some of them remind human capabilities, such as *learning*,

understanding needs, companionship, while others are more strictly connected to the realm of machines, like *data elaboration* and *connectivity*.

Meaningfulness The last dimension proposed in the survey is certainly among the toughest to depict, especially for users who, however advanced, are not familiar with academic debates at different disciplinary levels. Already intuitively, in fact, the complexity and variety for interpreting which qualities might fit within this domain can be detected, making it understandable why people encountered difficulties in formulating their contributions.

Then, it does not surprise that *meaningfulness* has been the dimension with the smallest number of submitted items, only 115, and some respondents explicitly asserted they were not able to answer at all. However, the perceived significance of this category can be attested in the effort of providing straightforward items, with no long-winded digressions.

Concerning the contents of the entries, most of them are characterized by fuzzy boundaries and attributes like *trustworthiness, multipurposeness, personality, empathy* and *understanding* emerged. The most relevant ones, instead, appealed to the human-artefact (computer or product) relationship, reminding *pragmatic* features. They are *usefulness, being beneficial, and helpfulness*.

4.5 Phase 3: A Research Workshop to Systematise the Findings

Before being able to proceed with the construction of an evaluation scale for AI-infused products, a conclusive systematisation was necessary to operationalise the findings from the prior research activities—the systematic review and mapping of UX evaluation methods (described in Chap. 3), as well as the survey analysis. For this purpose, a workshop within the research group has been organised to make sense of the values resulting from the intercoder assessment in a collective discussion and finally select the most promising descriptors for the construction of a UX evaluation scale.

As anticipated in the previous paragraph, a list of 36 so-called “golden” descriptors (which received the maximum score from all judges) had ultimately been extracted to portray a synthetic picture of the most relevant attributes to describe systems integrating AI-enabled capabilities. Yet, it was still too extensive for the purposes of the investigation.

Therefore, the main purpose of the conclusive workshop within the research team was to analyse the obtained results qualitatively. With the support of a collaborative Miro board, the identified “golden” descriptors were systematically categorised based on their suitability for inclusion in the scale. This process further refined the descriptors, ensuring a consistent and reliable foundation for the development of the evaluation method.

Specifically, descriptors related to *data protection*, *data quality*, *unfair bias avoidance*, *trustworthiness*, and *non-discrimination* were deemed problematic in terms of measurability and were subsequently excluded as “not usable.” Descriptors with multiple affiliations were consolidated into a single dimension, while redundant descriptors were classified under the “better to keep out” category. Attributes considered weaker, overly general, or already assessed by well recognised UX evaluation methods, such as *ease of use*, *functionality*, *understandability*, *intelligibility*, *learnability*, and *access to data*, were labelled as “could be in” but were not included in the essential list.

The final selection encompassed human-related qualities (*empathy*, *understanding*, and *usefulness*), characteristics intrinsic to the system (*helpfulness*, *intuitiveness*, *reliability*, *accuracy*, *adaptability*, and *context awareness*), attributes that integrate both sociotechnical elements [5] (*customisation*, *human oversight*, *data management*, *privacy*, *transparency*, and *reliability* as an ethical concern), and NLP qualities (*naturalness*, and *pleasantness*). Each descriptor was then elaborated in a suitable format for the scale, as described in the next chapter of this book.

4.6 Conclusions, Limitations, and Further Actions

The chapter presents and reflects on the general and specific features that can inform the construction of a UX evaluation method for AI-infused products. In particular, it is supported by the analysis of the results from a survey involving advanced users of the targeted devices.

The study, central to the Meet-AI project, moves from the premise that current methods are unable to capture the complexity and peculiarities of these novel products and services, which represents an incredible opportunity for design. It also stems from the results of the mapping of current UX tools and methods and from a systematic review of AI-based systems’ characteristics. Indeed, these produced eight possible dimensions to describe their UX: *pragmatic*, *aesthetic*, *hedonic*, *affective*, *intelligence*, *trustworthiness*, *conversational*, and *meaningfulness*.

To avoid tying the outcomes too much to the subjectivity of the investigators, advanced users of the products in question—who are sensitive to notions about UX of interactive objects—have been involved in a survey, the results of which eventually composed a set of 16 relevant descriptors (see Fig. 4.4) to build the evaluation scale.

Indeed, the study has limits in terms of number and similar background of the people participating, as well as because the methods, decisions and evaluations conducted by the researchers still present some degrees of subjectivity.

However, further developments should balance the qualitative character of work here presented, specifically pointing at a statistically valuable elaboration and validation of a UX evaluation scale for AI-infused systems.

Specifically, the identified descriptors will be the starting point for the elaboration of a set of questions to be submitted to a large number of smart speakers users (as they are the most widespread and used products embedding the technology under study)

Fig. 4.4 Ultimately selected descriptors to build a UX evaluation method for AI-infused products

	DIMENSION	DESCRIPTOR	
HUMAN	HED	Empathy	
	INT	Understanding	
	MEAN	Usefulness	
	PRA	Helpfulness	
SYSTEM	PRA	Intuitiveness	
	INT	Accuracy	
	INT	Adaptability	
	INT	Context Awareness	
SOCIOTECHNICAL ENSAMBLES	PRA	Customization	NLP Quality
	TRU	Human Oversight	CONV Pleasantness
	TRU	Data Management	CONV Naturaleness
	TRU	Privacy	
	TRU	Transparency	
	TRU	Reliability	

to derive the final scale and related method, as it will be presented in the following chapter.

After acquiring quantitatively solid results, the method should be generalisable and widely disseminated to support the design and consequent assessment of products integrating AI systems.

References

1. Spallazzo D, Sciannamè M (2021) UX descriptors for AI-infused products. 22224 Bytes
2. Creswell JW (2014) Research design: qualitative, quantitative, and mixed methods approaches. SAGE
3. Spallazzo D, Ajovalasit M, Ceconello M, et al (2021) Assessment of descriptors for UX evaluation of AI-infused products. 124653 Bytes
4. High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI
5. Johnson DG, Verdicchio M (2017) Reframing AI discourse. Minds Mach. <https://doi.org/10.1007/s11023-017-9417-6>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5

AIXE. A Method to Evaluate the UX of Systems Integrating AI



Abstract The chapter frames the AIXE (AI user eXperience Evaluation) scale, a statistically validated questionnaire to assess the UX of AI-infused products, describing its development process as well as its validation. AIXE is composed by 33 questions with 4 ordinal Likert-scale answers, organized around 12 descriptors related to the UX of the target systems. The questionnaire is meant to be proposed to the intended users of AI-infused products to quantitatively analyse the user experience they convey. The chapter further illustrates how the scale can be applied, its limitations and future opportunities.

5.1 Introduction

This chapter presents the final development and validation of the AIXE scale, a comprehensive tool specifically designed to assess the UX of artefacts that incorporate AI systems. With the increasing integration of AI-based functionalities in various products, it has become critical to evaluate their performance and interaction quality from the user's perspective, addressing the unique characteristics of these systems. For detailed background and the motivation that led to the development of this scale, the reader is referred to the preceding chapters.

The AIXE scale is intended to provide a holistic evaluation, focusing on multiple layers of user experience to capture the complexity and nuances involved in interacting with AI-infused products. Prior research has identified a structure commonly employed in UX evaluation models, typically based on UX dimensions and corresponding descriptors within a measurement framework. From a statistical point of view, these dimensions correspond to latent constructs (or latent variables), while more detailed questions, designed to assess the object of study, represent manifest or observable variables. The association between manifest and latent variables is typically quantified using factor loadings, which indicate the strength of these relationships.

To construct the AIXE scale, a conceptual model was developed, and hypotheses regarding the relationships between latent and manifest variables were formulated,

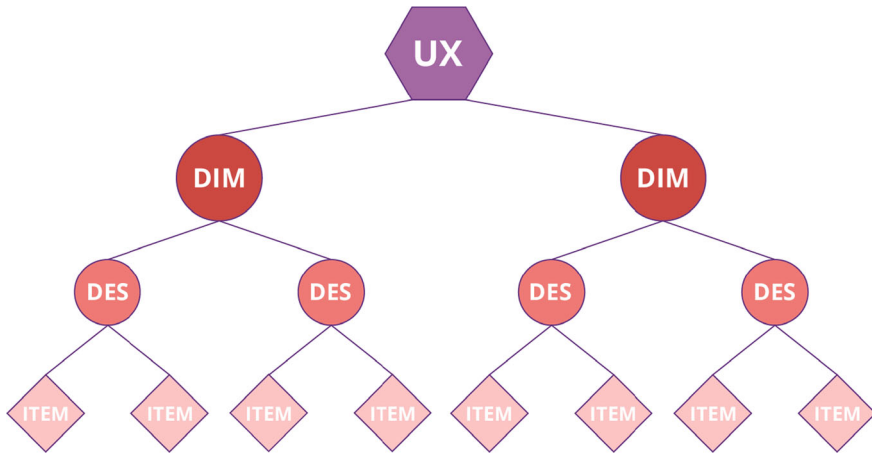


Fig. 5.1 Path diagram portraying the conceptual model underlying the scale

guiding the methodological approach. The scale was developed using a reflective hierarchical approach, characterised by a third-order model (Fig. 5.1). This reflective approach assumes that the latent constructs are well-defined in the respondent's mindset, meaning that the patent evaluation level reflects their conceptual understanding of the constructs. The decision to adopt a reflective rather than a formative model was driven by the nature of the constructs under investigation. In a reflective model, the indicators (items) are seen as manifestations of the underlying latent variables, which is particularly suitable when the objective is to assess how well the items reflect the broader dimensions of UX. This approach aligns with the research objectives of capturing the respondents' perception of the UX of AI-infused products through clearly defined constructs, ensuring that each item contributes meaningfully to the overall assessment.

Indeed, the scale includes a *measurement model*, referring to all items associated with the latent variables (denoted by squares in the diagram), and a *structural model*, including the latent variables (indicated with circles) and their nested relationships. In formal terms, the model is oriented to estimate the structure of the covariance [1], and it allows for the estimation of the relationships that exist between first-level (descriptors) and second-level (dimensions) latent variables, as well as those between latent dimensions and the overall UX (third-order factorial variable).

The scale's structure is divided into four distinct levels, presenting the broad concept of UX at its core. Based on prior research, six dimensions have been identified as particularly relevant for characterising the UX of AI-infused products: *intelligence*, *trustworthiness*, *meaningfulness*, *pragmatic*, *hedonic*, and *conversational* aspects. These dimensions provide a comprehensive framework for understanding how users perceive and interact with AI systems. Each of these six dimensions is further broken down into descriptors, which specify various attributes contributing to the overall user experience. 16 descriptors were selected as the most suitable for constructing

the scale, and these are to be further delineated into items, or manifest variables, in the form of close-ended questions.

The final phase of the AIXE scale's construction is the focus of this chapter. It details the identification of first- and second-order variables (items and descriptors) through a statistically validated process, leading to the definition of the scale and the related guidelines and tools for its application.

In the following sections, the chapter describes the methodology employed to engineer the selection of latent and manifest variables to ensure the final scale's statistical reliability. The validation process employed robust statistical techniques to ensure the reliability and validity of the scale. Exploratory Factor Analysis was used to identify the underlying structure of the data, followed by Confirmatory Factor Analysis to verify the factor structure and assess the goodness-of-fit of the model. Additionally, reliability testing was conducted using Cronbach's alpha, which is widely recognised as a standard measure of internal consistency.

Following this, the results from the initial draft of the AIXE scale, along with findings from Exploratory and Confirmatory Factor Analyses, are illustrated and discussed. The chapter concludes by outlining how to apply the AIXE scale and identifying future opportunities that the construction of this evaluation method might present.

5.2 Methodology

5.2.1 *Items and Questionnaire Elaboration*

The development of the AIXE scale was the concluding phase of the Meet-AI project, building on a comprehensive analysis of UX-related attributes that are crucial for characterising AI-infused products. This analysis is detailed in the preceding chapters of this book. The collaborative effort involved five researchers who systematically categorised 36 "golden" descriptors identified from previous investigations based on their perceived relevance to the UX assessment of AI-integrated products.

A digital Miro board facilitated the display of these descriptors on sticky notes, each associated with its corresponding dimension. Through collective discussion, 16 descriptors were selected to form the foundation of the scale. Recognising that each descriptor could encompass various semantic nuances, the researchers independently elaborated multiple questions for each descriptor to ensure a comprehensive and diverse set of items. These questions were then refined to maintain homogeneity and eliminate redundancy.

The initial questionnaire, prepared for testing, consisted of 65 items deemed sufficiently clear and distinct. The questionnaire adopted an ordinal scoring system, aligning with recent statistical developments favouring ordinal scales over summated scales with equal intervals [2]. This approach was also considered more intuitive for

respondents when addressing potentially complex questions and for validating the evaluation method.

To mitigate neutrality and reduce ambiguity, especially among non-expert respondents, the questionnaire offered four possible responses: *Not at all*, *A little*, *Rather*, *Very much*. This forced-choice format encouraged respondents to provide a definitive positive or negative response, yielding clearer data for analysis.

To validate the scale, the questionnaire included demographic profiling questions (age, gender, region of provenance) and contextual questions related to the use of smart speakers, which are prevalent examples of AI-infused products. Respondents were asked about which device they owned or were most familiar with and the usage frequency of such smart speakers.

5.2.2 *Statistical Validation*

Considering that the scale is intended for companies or designers to test their AI-infused products with users, the target audience to validate AIXE only had to respect two main requirements: being (i) anglophone to ensure language consistency and avoid biases, and (ii) familiar with AI-infused devices, smart speakers in particular.

Then, the validation of the AIXE scale was developed in two iterative stages. First, an Exploratory Factor Analysis (EFA) was chosen as the initial step to explore the underlying factor structure of the AIXE scale without preconceived hypotheses. It was followed by a Confirmatory Factor Analysis (CFA) to corroborate the identified structure in a new, independent sample, thereby ensuring the robustness and generalizability of the scale.

Stage 1: Exploratory Factor Analysis (EFA)

The first version of the questionnaire was distributed to a random sample of 671 individuals from the UK and USA through a specialised agency, yielding 601 valid responses after excluding 48 incomplete and 22 inapplicable responses (from individuals who never used smart speakers). The data were analysed using EFA [3, 4] to identify dimensions and reduce the number of items for a manageable tool. Given the ordinal nature of the responses, a polychoric correlation matrix was employed, with the WLS estimation method applied [5].

Stage 2: Confirmatory Factor Analysis (CFA)

Based on the EFA results, a refined version of the questionnaire was created and administered to a new sample of 733 respondents from the UK and USA, resulting in 702 valid responses. The sample sizes for both the EFA ($n = 601$) and CFA ($n = 702$) were deemed sufficient based on common recommendations for factor analysis, which suggest a minimum of 300 responses for reliable factor extraction [6, 7]. For security, a larger sample was considered because of the elevated number of items to be checked.

This dataset was used for CFA to test the structure model derived from the EFA and finalise the relevant latent variables and items. The ordinal nature of the scores necessitated the use of a diagonally weighted least squares (DWLS) estimator [8] to ensure robust analysis.

Through these rigorous validation processes, the AIXE scale was statistically validated, confirming its reliability and effectiveness in assessing the UX of AI-infused systems.

5.3 Results

5.3.1 Items and Questionnaire Generation

The foundation for the initial draft of the AIXE questionnaire was laid during the conclusive workshop among the Meet-AI research team, as comprehensively discussed in Chap. 4. This collaborative effort led to the selection of 16 descriptors that were considered pivotal for assessing the UX of AI-infused products. These descriptors include *accuracy*, *adaptability*, *context awareness*, *customisation*, *data management*, *empathy*, *helpfulness*, *human oversight*, *intuitiveness*, *NLP quality (pleasantness and naturalness)*, *privacy*, *reliability*, *transparency*, *understanding*, and *usefulness*.

Before proceeding with the independent elaboration of these descriptors into specific items (questions or statements for respondents to evaluate), the researchers reached a consensus on a synthetic definition for each descriptor, as summarized in Table 5.1. These definitions were intentionally kept generic to allow for broad interpretation, thus encompassing the diverse semantic nuances associated with each term. Based on their sensitivity and the dimensions associated to each descriptor, each researcher delineated the various meanings by translating them into questions or statements for users to assess. The finalized list of items is detailed in Table 5.2.

In certain occasions, the descriptor was simply contextualized into different scenarios where the quality might manifest, such as in the case of *accuracy*. Other times, the questions addressed specific interpretations of the descriptor's meaning. For instance, *usefulness* was unpacked into four distinct aspects: being valuable, meaningful, adding something to people's lives, or augmenting their capabilities. Moreover, the context in which *usefulness* is evaluated was also diversified, considering aspects like its relevance to one's daily routine, overall life experience, or personal development.

The objective of creating a broad set of questions was to uncover the most relevant and clear items during the validation process, thus ensuring that the final scale would be both comprehensive and applicable across a variety of use cases.

To ensure methodological rigor, once all researchers had contributed to the item list, the authors of this book further refined the content to produce a homogeneous and functionally coherent draft scale. The question format was specifically chosen over a

Table 5.1 Declination of the selected descriptors to compose the first draft of the AIXE scale, reporting the related dimension and a synthetic definition as shared among the researchers

Descriptor	Related dimension	Definition
Accuracy	Intelligence	The quality or state of being precise
Adaptability	Intelligence	The quality of being able to adjust to new conditions
Context awareness	Intelligence	Being aware of where one is and what is happening
Customisation	Pragmatic	The action of modifying something to suit a particular individual or situation
Data management	Trustworthiness	The way in which data are handled
Empathy	Hedonic	The ability to understand and share the feelings of another
Helpfulness	Pragmatic	The quality of giving or being ready to give help
Human oversight	Trustworthiness	The capability for human intervention during the design cycle of the system and monitoring the system's operation
Intuitivity	Pragmatic	The quality of being natural to learn, use, or understand
NLP Quality—naturalness, Pleasantness	Conversational	The capability to handle written or spoken text in a way that seems natural and pleasant to people, as they were talking to another human being
Privacy	Trustworthiness	Freedom from unauthorized intrusion
Reliability	Trustworthiness/ Pragmatic	The quality of performing consistently well
Transparency	Trustworthiness	Operating in such a way that it is easy for others to see what actions are performed
Understanding	Intelligence	The ability to understand something
Usefulness	Meaningfulness	The quality of being useful

statement format, along with a four-point response scale (*Not at all, A little, Rather, Very much*), as it was deemed more direct and engaging for respondents as well as consistent with contemporary statistical practices. This construction, coupled with a careful selection of items that were both clear and sufficiently diverse, formed the basis of the first version of the scale. Redundant questions were combined, ensuring that each focused on a single declination of the descriptor. Additionally, to minimise potential response biases, such as social desirability or acquiescence bias, the scale was designed with neutral wording and randomized question order, and the use of a four-point response scale without a neutral middle option was intended to encourage more definitive responses, thereby reducing the likelihood of non-committal answers.

As illustrated in Table 5.2, the first draft of the AIXE scale comprised 65 items.

Table 5.2 First draft of the AIXE scale, including all the 65 items elaborated by the research team and the indications of which items were excluded during the validation process

Dim	Descriptor	Question 1st part	Code	Question 2nd part	
<i>INT</i>	Accuracy	How accurate is the system in	D01	_1	Responding to your requests? ^a
				_2	Performing the task? ^a
				_3	Anticipating your needs?
				_4	Matching your needs?
<i>INT</i>	Adaptability	Is the system's behaviour adapting	D02	_1	To your habits?
				_2	To your needs?
				_3	Over time?
<i>INT</i>	Context awareness	Do you think the context in which the system is placed	D03	_1	Gives it important information to work accordingly? ^a
				_2	Affects its behaviour?
				_3	Affects its performance?
<i>PRA</i>	Customisation	Do you think you can customize	D04	_1	The system to your needs? ^b
				_2	The system's behaviour?
				_3	The system to your habits?
<i>TRU</i>	Data management	Do you feel you can manage the data	D05	_1	Affecting the information the system uses? ^b
				_2	Collected by the system?
				_3	That the system uses?
<i>HED</i>	Empathy	Do you feel the system is empathetic	D06	_1	With you?
				_2	And behaves according to the relationship it has built with you?
				_3	In anticipating your needs?
				_4	Towards your needs?
				_5	And this makes it perform better? ^a

(continued)

Table 5.2 (continued)

Dim	Descriptor	Question 1st part	Code	Question 2nd part	
<i>PRA</i>	Helpfulness	Do you think the system is helpful	D07	_1	In your daily life? ^a
				_2	In responding to your needs? ^a
				_3	In achieving your tasks? ^a
<i>TRU</i>	Human oversight	Do you feel you can control	D08	_1	The operations of the system?
				_2	How the system behaves?
				_3	How the system performs its tasks?
<i>PRA</i>	Intuitiveness	Is the system intuitive	D09	_1	And easy to use? ^a
				_2	Making you know what to expect? ^a
				_3	In manifesting its potentials? ^a
				_4	Making you comfortable in using it? ^a
<i>CONV</i>	NLP	Do you think the system	D10	_1	Lets you understand what it says? ^a
				_2	Understands what you say? ^a
				_3	Establishes a good dialogue with you? ^a
				_4	Has a good quality in terms of voice interaction? ^a
<i>CONV</i>	NLP (voice quality)	Do you perceive the system's voice as	D11	_1	Pleasant?
				_2	Natural?
				_3	Likable?
<i>TRU</i>	Privacy (passive)	Do you feel the system protects	D12	_1	Your privacy?
				_2	Your data?
				_3	Your private information?
<i>TRU</i>	Privacy (active)	How the system handles privacy makes you	D13	_1	Trust it? ^a
				_2	Share your data? ^a
				_3	Safely share your personal information? ^a

(continued)

Table 5.2 (continued)

Dim	Descriptor	Question 1st part	Code	Question 2nd part	
<i>TRU</i>	Reliability	Do you rely	D14	_1	The system's behaviour?
				_2	The system's responses?
				_3	The system increasingly over time?
				_4	What the system proposes? ^b
				_5	The system is doing what you expect? ^a
<i>TRU</i>	Transparency	Is the system transparent	D15	_1	In the way it adapts to your needs? ^a
				_2	About its processing?
				_3	In showing what its decisions depend on?
				_4	In the way it adapts to your interests? ^a
				_5	In communicating the processes it performs? ^a
				_6	In explaining where information is retrieved from? ^a
				_7	In the way it adapts to your habits?
				_8	In explaining how it works? ^a
<i>INT</i>	Understanding	Do you think the system understands	D16	_1	You? ^a
				_2	How to anticipate your needs? ^a
				_3	Your needs? ^a
<i>MEAN</i>	Usefulness	Do you think the system	D17	_1	Is valuable in your daily routine? ^a
				_2	Adds meaning to your life?
				_3	Adds something to your life?

(continued)

Table 5.2 (continued)

Dim	Descriptor	Question 1st part	Code	Question 2nd part
			_4	Has value for you?
			_5	Augments your capabilities? ^a

^aitems excluded after the EFA, ^bitems excluded after the CFA

5.3.2 Statistical Validation

Exploratory Factor Analysis

The statistical validation of the AIXE scale started with an EFA performed on the responses to the first version of the questionnaire. Out of the 671 initial responses, those from participants who reported never using smart speakers or had missing data were considered non-usable. Consequently, 70 responses were excluded, leaving a sample of 601 valid responses for the analysis.

To ensure the robustness of the EFA, the weighted least squares (WLS) estimation method was employed. This method was chosen specifically because the data were ordinal in nature, necessitating the use of an asymptotic covariance matrix to generate accurate estimates. The EFA aimed to identify the most statistically relevant items and descriptors that could reliably measure the intended constructs of the AIXE scale.

From the original set of 65 items and 17 descriptors, the EFA identified 36 items linked to 12 descriptors as the most relevant for the scale. In Table 5.2, they are the ones with no symbol and those with a^b next to them. These selected items were the ones presenting a factor loading value greater than 0.5, indicating a strong correlation with the underlying factors. To maintain a balanced and concise scale, the analysis was designed to retain a maximum of three items per descriptor. However, an exception was made for the descriptor *empathy*, which was the only one representing the *hedonic* dimension. Indeed, four items were retained because of their high factor loadings and their qualitative significance and diversity.

Additionally, all the items associated with the *privacy* descriptors (D12 and D13) exhibited high factor loadings. However, due to the similarity in the meanings of the questions, it was determined that such redundancy could potentially confuse respondents. Therefore, only the items associated with D12 were retained, as they had higher factor loadings, indicative of a more straightforward and distinct question structure.

The overall outcomes of the EFA were promising, as the goodness-of-fit indices denoted a well-fitting model. Specifically, the analysis reported a Cumulative Explained Variance of 65%, a Tucker-Lewis Index (TLI) of 0.904, and an RMSEA index of 0.047. Considering widely accepted guidelines in structural equation modelling, a TLI close to 0.95 or higher, and an RMSEA below 0.05, are considered indicative of a well-fitting model [9].

Moreover, the selection of goodness-of-fit indices and the use of WLS and DWLS methods were guided by established statistical best practices in factor analysis for

ordinal data [10, 11] and the results suggest that the model adequately represents the underlying structure of the data.

Hence, a new version, with only the best performing items, had to be tested.

Confirmatory Factor Analysis

To further validate the AIXE scale, a CFA was conducted using a new sample of 736 participants. They were submitted the reduced questionnaire, consisting of the 36 items identified through the EFA. As before, responses from participants who reported not using smart speakers were excluded, resulting in the removal of 31 responses. There were no cases of missing information, leaving a final sample of 705 valid responses for the CFA.

The CFA aimed to confirm the factor structure identified during the EFA and assess the overall fit of the model. The analysis led to the removal of three additional items (D04_1, D05_1, D14_4), which were identified as suboptimal based on their performance in the model. These have a^b symbol next to them in Table 5.2. The final set is therefore composed of 33 items, displayed in Table 5.3.

The validity of the model was confirmed by calculating various goodness-of-fit indices using the diagonally weighted least squares (DWLS) method. The results were highly favourable, with a Comparative Fit Index (CFI) of 0.986, a Tucker-Lewis Index (TLI) of 0.987, and an RMSEA of 0.038. The Composite Reliability was calculated to be 0.99, and the Average Variance Extracted (AVE) was 0.89. These indices suggest a robust model with strong internal consistency and an excellent fit to the data.

The finalised version of the AIXE scale comprises 6 dimensions and 12 descriptors that serve as latent variables to measure the user experience (UX) of AI-infused products. The reliability of each dimension was assessed using Cronbach's alpha coefficients, which are summarised in Table 5.4. In line with the standards [12], values above 0.70 indicate an acceptable internal consistency, values between 0.8 and 0.9 are good, while over 0.9 the results are excellent.

Ultimately, the final and registered scale illustrates that the descriptors of *accuracy*, *adaptability*, and *context awareness* are indicators of *intelligence*; *customisation* and *reliability* represent the *pragmatic* dimension; the latter also embodies *trustworthiness* alongside *data management*, *human oversight*, *privacy*, and *transparency*. *Empathy* remains the sole descriptor for the *hedonic* dimension, while *natural language processing (NLP) qualities* and *usefulness* represent the *conversational* and *meaningfulness* dimensions, respectively.

As the results confirm the soundness of the AIXE scale, verifying its capacity to measure the intended constructs with high reliability and validity, the research questions driving the development of this scale can thus be considered statistically validated.

Table 5.3 Final structural organization and list of items of the AIXE scale

Dim	Descriptor	Question 1st part	Question 2nd part
<i>INT</i>	Accuracy	How accurate is the system in	Anticipating your needs?
			Matching your needs?
<i>INT</i>	Adaptability	Is the system’s behaviour adapting	To your habits?
			To your needs?
			Over time?
<i>INT</i>	Context awareness	Do you think the context in which the system is placed	Affects its behaviour?
			Affects its performance?
<i>PRA</i>	Customisation	Do you think you can customize	The system’s behaviour?
			The system to your habits?
<i>TRU</i>	Data management	Do you feel you can manage the data	Collected by the system?
			That the system uses?
<i>HED</i>	Empathy	Do you feel the system is empathetic	With you?
			And behaves according to the relationship it has built with you?
			In anticipating your needs?
			Towards your needs?
<i>TRU</i>	Human oversight	Do you feel you can control	The operations of the system?
			How the system behaves?
			How the system performs its tasks?
<i>CONV</i>	NLP (voice quality)	Do you perceive the system’s voice as	Pleasant?
			Natural?
			Likable?
<i>TRU</i>	Privacy (passive)	Do you feel the system protects	Your privacy?
			Your data?
			Your private information?
<i>TRU</i>	Reliability	Do you rely	The system’s behaviour?
			The system’s responses?
			The system increasingly over time?
<i>TRU</i>	Transparency	Is the system transparent	About its processing?
			In showing what its decisions depend on?
			In the way it adapts to your habits?
<i>MEAN</i>	Usefulness	Do you think the system	Adds meaning to your life?
			Adds something to your life?
			Has value for you?

Table 5.4 Cronbach coefficients for each dimension

Latent variable	N. of Items	Cronbach coefficients
General UX	33	0.967
Intelligence	7	0.873
Pragmatic	2	0.8
Hedonic	4	0.914
Trustworthiness	14	0.941
Conversational	3	0.866
Meaningfulness	3	0.873

5.4 Discussion

5.4.1 Reflecting on the Results

For usability purposes, the scale needed a significant reduction of questions with respect to the ones initially depicted. Still, for a granular assessment of AI-infused products, too few or general items, descriptors, and dimensions were not acceptable either. A balanced result has been obtained through the EFA and CFA validation steps.

All the dimensions on which the scale was drafted were confirmed in the validation process, ensuring an interesting breadth of qualities. In line with the results of both the systematic literature review related to AI-infused systems and the investigation with advanced users previously conducted, *trustworthiness* has a clear predominance of representation in the final scale, encompassing five descriptors and 14 items. With respect to the most recurrent clusters emerged in the analysis (*transparency, acceptability, privacy and safety, legality, data concerns*), only legal concerns are not covered in the scale. Indeed, due to the sensitivity of the topic, the professional expertise required to handle these evolving matters, and the consistency with UX assessment, it makes sense that legal considerations are out of scope. Comparing the final descriptors with the key requirements identified by the European Commission, three out of seven are explicitly covered by AIXE. Namely, (1) *human agency and oversight*, (3) *privacy and data governance*, (4) *transparency*. These are, in fact, the most coherent with assessing UX. The *reliability* descriptor, instead, can be considered as halfway through (2) *technical robustness and safety* and (5) *diversity, non-discrimination and fairness*, although it does not uniquely refer to one or the other. In general, *trustworthiness* descriptors and items performed very well in the EFA, and the only reason why one descriptor was excluded is because of its redundancy. Indeed, *privacy* was initially included from two angles: a “passive” one, only referring to the capability of the product to protect people’s privacy, personal data and information; and an “active” one, focusing on the consequences on people’s behaviour. Possibly because of the more convoluted construction of the questions,

the latter performed slightly worse and was excluded, with no significant loss of collected information.

Even though the previous stages of the research seemed more in favour of the *conversational* dimension rather than *intelligence*, their fates are reversed in the final scale. Probably, this is not an indication of their importance, but more of the ease with which they can be faceted. Indeed, *intelligence* is expressed by three descriptors and seven items and somehow echoes the findings from the systematic review. *Adaptability* and *context awareness* are still pillars in the articulation of this dimension, that additionally includes *accuracy*. While this quality might be attributed to several dimensions, it is significant that it was selected and validated as a measure of the system's *intelligence*. What proved relevant for the assessment of the UX is the effective ability to anticipate and match people's needs. Instead, considering *accuracy* in a more technical sense—i.e., in relation to the performance of the task or the quality of the response, has also been proven of secondary importance by the EFA, and the related questions have been discarded. Also, the questions related to *understanding* did not obtain satisfying results, maybe because they are quite blurry and overlapping with other more easily quantifiable items.

The outcomes of the items in the *conversational* dimension were less expected. The richness observed in both the systematic review and the survey submitted to advanced users—including, for instance, *conversational attributes*, *language property*, and *understanding*—solely reduced to voice qualities. This result might suggest an overarching relevance of this aspect in the dialogic interaction with AI-infused products. Yet, if the *conversational* dimension is particularly important for a specific artefact, one might consider to complement this wide-ranging assessment, with one of the many evaluation methods addressing NLP.

To conclude the overview of the dimensions emerged from the systematic review of the studies related to AI-infused products, *meaningfulness* is represented by one descriptor, *usefulness*, and three items. As remarked in the previous stages of the research, *meaningfulness* proved to be a difficult dimension to address because of its many possible definitions. It reached the first draft of the scale with a quite practical descriptor, associating meaning to the actual *usefulness* that people can find in the product under evaluation, which passed the barrier of the EFA and CFA. Interestingly, however, its declination into items left a margin of interpretation, being articulated as the capability of the product to add meaning or something to users' life, or to have value for them. The more specific questions about the system being valuable in the daily routine or being augmenting users' capabilities did not reach satisfying results, hinting to a possible scarce generalisability of these items. The openness of the included questions, though, favours a subsequent deepening with qualitative studies to get to know people's individual perspectives.

Moving on to more traditional factors in the UX assessment, the *pragmatic* and *hedonic* dimensions remained with just one descriptor each. The first suffered a significant downsizing, especially if considering its relevance in the vast majority of the evaluation methods analysed (both AI-related and non-related). Of the descriptors and items proposed in the first draft, *helpfulness* was excluded by the EFA results, possibly because of its redundancy with *usefulness*, and the same happened

to *intuitiveness*, which instead embedded fundamental concepts for assessing UX, like ease-of-use, intelligibility, familiarity, etc. This result came unexpected, yet it reinforced the need for the scale to be shaped by attributes closer to AI-based qualities. Indeed, *customisation* is the only *pragmatic* descriptor in the final method. This becomes particularly relevant for products integrating ML systems, as they can evolve and assume behaviours tailored to their users.

On a different note, *empathy*, a descriptor which importance was underlined by the fact that it was associated with almost all the dimensions but performed better in relation to the *hedonic* one, almost retained all the proposed items as they got excellent results. The only question discarded was indeed the most stretched in meaning and possibly hard to fully comprehend. This might suggest that the human-like connection with the product should require careful attention when developing AI-infused products. Perhaps a legacy of the early and still current discourse on AI as a mimic of human beings, users might have built the expectation that these systems can be empathetic with them and their needs and, at least in our study, it is the only measure for determining the pleasure of use of AI-infused products.

5.4.2 Strengths and Limitations

Although the methods, decisions, and evaluations employed throughout the research project to identify the latent and manifest variables for the AIXE scale inherently involve a degree of subjectivity, the validation process adhered strictly to established statistical methodologies. Additionally, a large sample size and an iterative process characterised this stage. These elements aim to ensure the reliability of the results. However, to further confirm its robustness, future studies should aim to replicate these findings in different contexts and with more diverse populations.

Moreover, applying the AIXE scale in real-world scenarios, such as in the assessment of AI-infused products by companies or start-ups, will provide valuable opportunities for further refinement. The direct interaction with designers and developers during these applications could highlight areas for improvement and either reaffirm or challenge the validity of the evaluation method, thereby enhancing its robustness and practical relevance.

Ultimately, the scale is comprehensive and oriented towards AI-related features. It presents more dimensions and descriptors than most of the current evaluation methods analysed, which hopefully can help in addressing the complexity and nuances of AI-infused products across different aspects of the UX. Because of the variety and non-situatedness of the included qualities, AIXE should be versatile and applicable to a wide range of AI products and in different fields, from home and entertainment devices to industrial or healthcare implementations.

Nonetheless, as the final elaboration of the validated scale points more at peculiar AI-related attributes than at classic UX concerns, a triangulation of the data from AIXE with other, well-established, methods might be a preferable option to gather a more comprehensive picture.

5.5 Application of the AIXE Scale

5.5.1 *Setting*

The AIXE scale is a quantitative evaluation method intended for companies and professionals involved in the design, development, and distribution of AI-infused products. Its application is not limited to fully operational products but also extends to prototypes, including those at Technology Readiness Level 7 (TRL7) or higher. This flexibility makes the AIXE scale a valuable tool for evaluating AI products at different stages of their development, providing insights that can inform both product improvement and user satisfaction, even before they are released on the market.

To be successfully implemented, the selection of participants is a key factor in ensuring the accuracy and reliability of the results. While no specific experience level with AI-infused artefacts is needed, respondents should have used the product to evaluate for a sufficient period, typically no less than 15 days, to ensure they have had enough time to form a meaningful experience. This period allows users to fully engage with the AI-infused product and provides a more accurate reflection of its performance and user experience.

The sample can be randomly selected, or specific criteria may depend on the interests of the study. Following the common “5–10 respondents per item” rule of thumb for determining sample size in surveys or questionnaires in the fields of psychometrics, social sciences, and UX research, the ideal sample size can range from 165 to 330 respondents for reliable quantitative results when using all the 33 items of the scale. However, the AIXE scale is supposed to be modular and adaptable to contextual needs. Therefore, a smaller sample size might suffice if not all the items are used for testing.

Administering the AIXE scale involves providing a structured questionnaire to the selected users in a paper-based or digital format, only after they have had substantial interaction with the AI-infused product being evaluated. The validated questionnaire is in English and consists of 33 questions, each requiring a response on an ordinal 4-point Likert scale, ranging from “Not at all” to “Very much.” This format allows for capturing clearly positively or negatively connotated user feedback.

Before submitting the scale, it is important to provide clear instructions to the participants, including specifying the purpose of the questionnaire—which is evaluating their UX with the AI-infused product—and underlining the relevance of honest responses. It is at the client’s discretion whether to remunerate respondents.

5.6 Conclusion and Future Work

The development of the AIXE scale is grounded in an extensive review of existing UX evaluation methods. Despite the proliferation of tools for assessing user experience, a significant gap was identified in the availability of instruments tailored specifically

for AI-integrated products. The AIXE scale, therefore, represents a pioneering effort to fill this gap, offering a unique and structured approach to evaluating the UX of AI-driven products. By capturing the distinct qualities of AI interactions, the AIXE scale provides valuable insights that can guide the design and development of more user-friendly AI-infused products.

Through the adoption of a reflective hierarchical approach, the AIXE scale has been meticulously constructed and validated. The scale's four-level structure—comprising UX as the overarching concept, six dimensions, 12 descriptors, and 33 items—ensures that it can accurately measure the diverse aspects of user interactions with AI-infused products. The rigorous process of developing and validating the scale involved both exploratory and confirmatory factor analyses, confirming the reliability and validity of the measurement model.

Looking forward, the AIXE scale might open new avenues for research and application in the field of AI, human–computer interaction, and UX design. While it has been tested and validated considering physical products integrating AI systems, its application to digital AI-infused products, presenting similar challenges to the UX, might be further investigated. Additionally, applying it to very specific niches of products can be an opportunity for exploring the modularity and versatility of the scale to different necessities. Indeed, the scale has been designed to be comprehensive enough to adapt to diverse types of products. Still, more in-depth research might strengthen this hypothesis.

Further future developments include creating a digital version of the AIXE scale, with a user-friendly dashboard including automatic calculations and useful visualizations, to make it publicly accessible and usable. Additionally, translating the core qualities identified by the AIXE scale into meta-design principles and practical tools might be a valuable opportunity to influence the foundational stages of AI-infused product design. By embedding these principles early in the design process, both educational and professional contexts can harness their potential to inspire meaningful innovation, contributing to the creation of AI-infused products that better serve and enrich human experiences across various domains.

References

1. Jöreskog KG (1978) Structural analysis of covariance and correlation matrices. *Psychometrika* 43:443–477. <https://doi.org/10.1007/BF02293808>
2. Spector PE (1992) *Summated rating scale construction: an introduction*. Sage Publications, Newbury Park, Calif
3. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ (1999) Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods* 4:272–299. <https://doi.org/10.1037/1082-989X.4.3.272>
4. Spearman C (1904) “General intelligence”, objectively determined and measured. *Am J Psychol* 15:201–292. <https://doi.org/10.2307/1412107>
5. Flora DB, Curran PJ (2004) An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol Methods* 9:466–491. <https://doi.org/10.1037/1082-989X.9.4.466>

6. Comrey AL, Lee HB (2013) *A first course in factor analysis*, 2nd edn. Psychology Press, New York
7. Hinkin TR (1995) A review of scale development practices in the study of organizations. *J Manag* 21:967–988. [https://doi.org/10.1016/0149-2063\(95\)90050-0](https://doi.org/10.1016/0149-2063(95)90050-0)
8. Li C-H (2016) Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behav Res Methods* 48:936–949. <https://doi.org/10.3758/s13428-015-0619-7>
9. Hu L, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model Multidiscip J* 6:1–55. <https://doi.org/10.1080/10705519909540118>
10. Brown T, Malmgren D, Stringer M (2017) Design for augmented intelligence. In: Medium. <https://medium.com/ideo-stories/design-for-augmented-intelligence-9685c4db6fbb>. Accessed 15 Dec 2020
11. Kline RB (2016) *Principles and practice of structural equation modeling*, 4th edn. Guilford Press, New York, NY, US
12. Tavakol M, Dennick R (2011) Making sense of Cronbach’s alpha. *Int J Med Educ* 2:53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6

Applying AIXE to Compare Domestic Smart Speakers



Abstract The chapter describes and analyses the results of a comparative study conducted on domestic smart speakers and aimed at assessing the user experience they entail. About 1400 respondents from US and UK answered the AIXE (AI user eXperience Evaluation) questionnaire, evaluating the smart speaker they commonly use in their daily life (typically, Amazon Echo, Google Nest, and similar). The results provide an overview of different UX dimensions deemed relevant by the users and highlight different performances of the analysed devices.

6.1 Introduction

This chapter presents a comparative analysis of user experience results obtained from administering the AIXE questionnaire to evaluate common domestic smart speakers in 2021 and 2023.

The study had two primary objectives: first, to further validate the AIXE questionnaire by evaluating its effectiveness in assessing market products, and second, to offer a comprehensive understanding of the UX performance of widely used smart speakers over time.

While the previous chapter extensively covered the validation process, this chapter shifts focus to the results obtained from using the AIXE questionnaire, specifically evaluating and interpreting the UX performances of the most common devices.

Since the introduction of the Amazon Echo in 2014, smart speakers have rapidly gained traction in the market, attracting growing interest and capturing an increasing share while becoming more affordable. The smart speaker market was valued at approximately USD 8.02 billion in 2021, marking a significant milestone in its development. This growth trajectory accelerated from 2021 onward, with the market expected to experience a robust compound annual growth rate (CAGR) of around 16.65% from 2022 to 2030 [1].

By 2022, the global smart speaker market had expanded to approximately USD 10.06 billion, reflecting the steady rise in consumer adoption and the increasing presence of these devices in households and businesses worldwide [2].

This upward trend continued into 2023, with the market value reaching an estimated USD 12.52 billion. Projections for 2024 suggest the market will further grow to USD 15.00 billion, and by 2032, it is anticipated to reach USD 61.40 billion, underscoring the sustained demand and integration of smart speakers into the domestic environment [3].

While the rapidly growing market demonstrates strong consumer interest in these products, with millions of households now relying on them daily, we may recognise that they still have perceivable limitations in terms of user experience (UX) [4].

The present study aims to delve deeper into this widespread perception of smart speakers, seeking to thoroughly understand how the most commonly used devices perform not only in terms of overall user experience but also across the specific UX dimensions and descriptors evaluated by the AIXE scale.

6.1.1 Sample of Respondents and Methodology

The study involved a sample of 1608 respondents coherent with the following selection criteria: (i) aged 18–65, (ii) anglophone, and (iii) already familiar with at least one AI-enabled smart speaker. They have been recruited through an agency that remunerated them upon completing the questionnaire, which is shared through a proprietary platform.

In detail, the study has been conducted twice, in 2021 and 2023, recruiting 722 respondents in 2021 (366 from the UK and 356 from the US) and 866 respondents in 2023 (462 from the UK and 424 from the US).

All respondents completed the 33 questions of the AIXE questionnaire, rating each item on a 4-point ordinal Likert scale that ranged from “Not at all” to “Very much.” In addition to their responses, participants provided demographic information such as age, the specific smart speaker device they own or use, and the frequency of their interactions with it.

The collected data were systematically processed to generate results at three distinct levels: (i) 12 specific descriptors, (ii) 6 broader UX dimensions, and (iii) an overall general UX score. These three levels are summarized in Table 6.1.

To enhance the clarity and comparability of the findings, all results were normalized to a percentage scale, enabling a more immediate and intuitive interpretation of the data.

Table 6.1 Three levels of analysis: (i) Descriptors, (ii) UX dimensions and (iii) General UX

Descriptors	UX dimension	
Accuracy	Intelligence	General UX
Adaptability		
Context awareness		
Customisation	Pragmatic	
Reliability	Trustworthiness	
Data management		
Human oversight		
Privacy		
Transparency		
Empathy	Hedonic	
Voice quality	Conversational	
Usefulness	Meaningfulness	

6.2 How Smart Speakers Performed

6.2.1 The Big Picture

Examining the entire dataset, which includes both UK and US respondents from 2021 and 2023, smart speakers’ general user experience (UX) is perceived as slightly above average. The overall UX score is 54%, indicating that while respondents find the user experience acceptable, it falls short of being impressive.

A closer analysis of the individual UX dimensions (Fig. 6.1) reveals some variation. The *pragmatic* dimension scores 57%, and the *intelligence* dimension achieves 58%, both slightly higher than the other dimensions. Meanwhile, *meaningfulness* scores 53%, *trustworthiness* and *hedonic* both report 52%, and the *conversational* dimension trails slightly at 51%. These results suggest that while certain aspects of UX are stronger, the overall experience remains modest across all dimensions.

A closer examination of the descriptors (Fig. 6.2) within the UX dimensions reveals a more varied picture. For instance, certain descriptors perform better than others within the *intelligence* dimension. *Accuracy* reaches 63%, and *adaptability* scores 59%, indicating that respondents perceive their devices as precise and capable of adjusting to different situations. However, *context awareness* lags behind at 45%, suggesting that users feel their devices are less effective at understanding the surrounding environment and act accordingly.

A similar pattern emerges within the *trustworthiness* dimension. Users feel they understand what is happening with their devices, as reflected by a *transparency* score of 61%. However, concerns arise around *privacy*, which amounts to just 47%, and *reliability*, which stands at 46%. This indicates that while users appreciate the clarity of their device’s operations, they remain uneasy about privacy intrusions and question the overall reliability of their smart speakers.

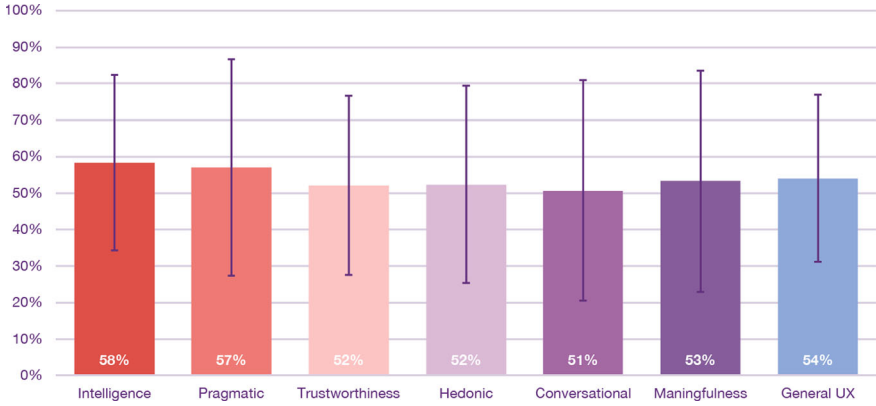


Fig. 6.1 Scores of the UX dimensions and General UX for the entire dataset

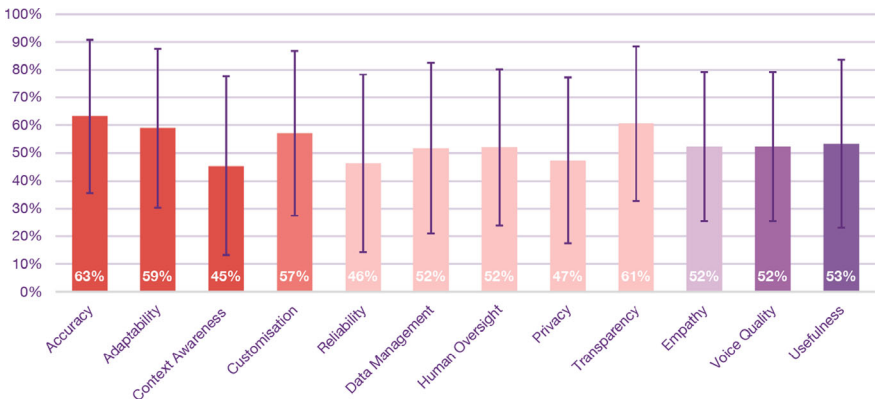


Fig. 6.2 Scores of the descriptors for the entire dataset

In general, we may say that only three descriptors—out of twelve—are under the threshold of 50%: *context awareness* (45%), *reliability* (46%) and *privacy* (47%).

6.2.2 Performances Over Time: 2021 Versus 2023

The analysis of the entire dataset reveals a picture of average user satisfaction with their smart devices. It highlights stronger performances in areas like the *accuracy* of the intelligent systems and their perceived *transparency*. However, the lowest scores are seen in the systems’ ability to understand their surroundings (*context awareness*), their sense of *reliability*, and users’ perception of *privacy* protection.

A valuable aspect of the study lies in comparing the overall performance of smart devices between 2021 and 2023, which allows us to observe how user opinions have shifted over time. At first glance, the *general UX* score has remained steady at 55% across both years, suggesting that user satisfaction has neither significantly improved nor declined—remaining just above the average 50% threshold.

However, a closer look at the data reveals differences across the six UX dimensions, pointing to subtle shifts in user perceptions (Fig. 6.3). Most notably, there have been improvements in the *intelligence* and *conversational* dimensions. The *intelligence* dimension, for instance, saw a modest increase from 57 to 59%, reflecting a 2% rise that suggests users are increasingly recognising advancements in the devices’ ability to understand and process information. The *conversational* dimension also improved, growing from 50 to 51%, indicating that users are perceiving slightly better interaction and dialogue capabilities in smart devices. These trends indicate that while the overall satisfaction has decreased, users are gradually recognising improvements in how these devices function and interact. Specifically, they have noticed that the devices have become more intelligent and that their voice interaction quality is starting to be better.

Additionally, the *pragmatic* dimension has remained constant. It traditionally relates to ergonomic factors and, in this context, reflects the devices’ ability to adapt to user preferences and customisation. While the performance on this trait is slightly above the average, and the problems associated with this dimensions are increasingly recognized, no improvements have been remarked.

On the other hand, not all dimensions have followed a positive or static trend. Three key areas have shown a decline in performance. The *meaningfulness* dimension experienced the most significant drop, falling from 56% in 2021 to 51% in 2023. This suggests that users are significantly recognizing that these devices are less providing a personal or practical value over time, possibly not meeting the initial expectations. Similarly, the *trustworthiness* and *hedonic* dimensions saw a 4% and a 3% decline

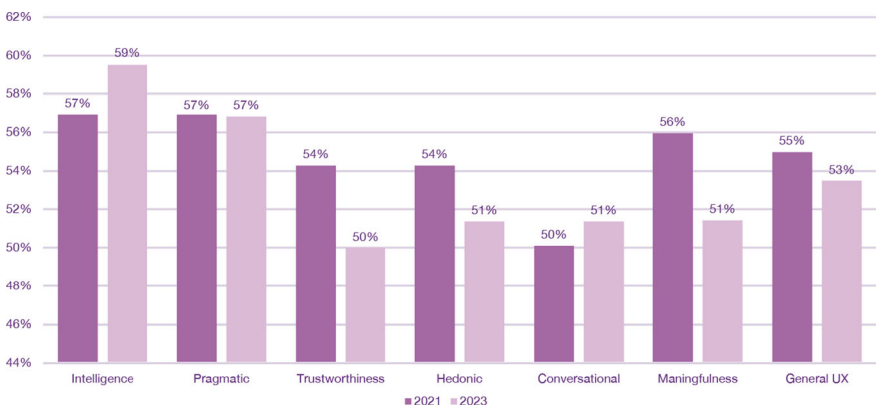


Fig. 6.3 Scores by survey year: 2021 versus 2023

respectively, starting both from a 54%. This indicates that users perceive the devices as slightly less trustworthy and less capable of evoking a positive emotional response.

Taking a broader view, it becomes clear that users are experiencing a loss of trust in their devices (as indicated by the decline in *trustworthiness*), find them less enjoyable and emotionally engaging (reflected in the drop in the *hedonic* dimension), and more significantly, see them as less valuable or integral to their daily lives (as shown by the decrease in *meaningfulness*).

The findings suggest that although smart devices are becoming more advanced in terms of their technical capabilities, they are falling short in addressing key areas that contribute to users' sense of security and personal relevance.

Overall, the study highlights improvements in dimensions directly related to technological advancements, especially from an AI perspective. Users view the devices as smarter and better at human interaction. However, despite these improvements, the overall user experience has not significantly advanced. In fact, there is growing scepticism about the usefulness of smart speakers in everyday life, and users are becoming less trusting of these devices.

6.2.3 Performances in the UK

Another layer of analysis in the study focuses on the performance of smart devices in the UK and USA over time.

Examining the 2021 AIXE questionnaire results for the UK (Fig. 6.4) a pattern similar to the overall dataset is observed. The *general UX* score stands at 54%, matching the overall analysis. There are only minor variations across the other UX dimensions, with the most notable difference found in *trustworthiness*, which scores 54%, 2% higher than the overall dataset.

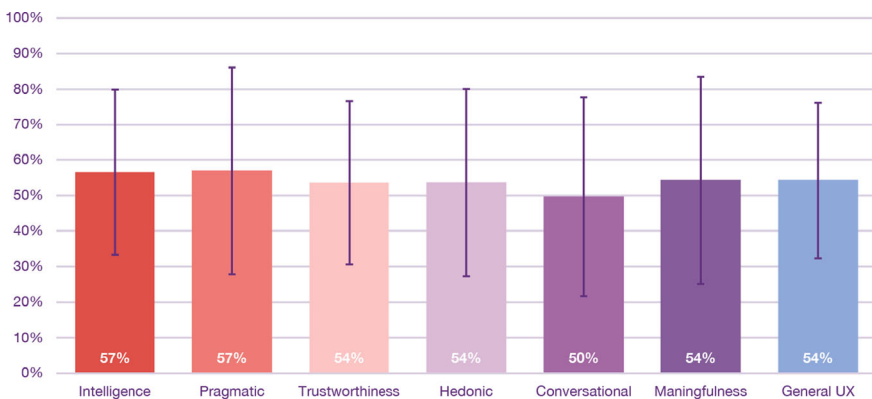


Fig. 6.4 Dimensions scores UK 2021

A closer look at the more nuanced level of the descriptors yields similar observations (Fig. 6.5). The most significant difference is found in the *data management* descriptor, where UK respondents in 2021 expressed greater confidence, scoring 55% compared to the overall average of 52%. For the other descriptors, the differences are minimal, with none exceeding 2%.

Looking at the 2023 results (Fig. 6.6), we observe a slight decline in the *general UX* score, dropping from 54 to 53%. This decrease reflects a general reduction across nearly all six UX dimensions. Notably, *trustworthiness* and *hedonic* dimensions show the most significant drops, both falling from 54 to 50%. The *pragmatic* and *meaningfulness* dimensions also declined by 3%, with *pragmatic* decreasing from 57 to 54%, and *meaningfulness* from 54 to 51%.

In contrast, there were small improvements in the *intelligence* and *conversational* dimensions, both registering a modest 1% increase in 2023.

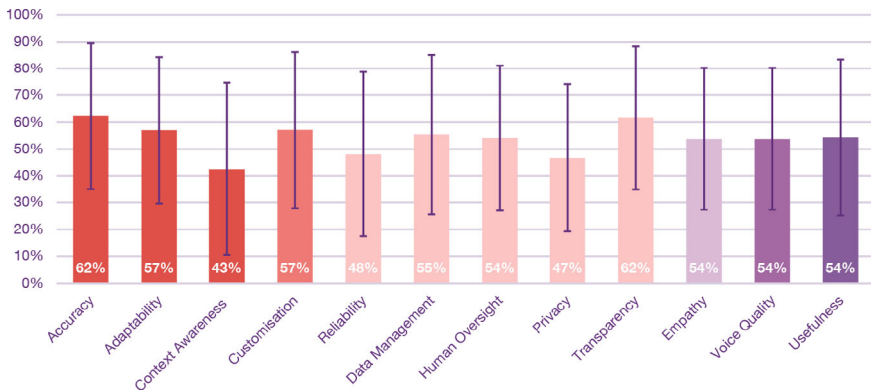


Fig. 6.5 Descriptors scores UK 2021

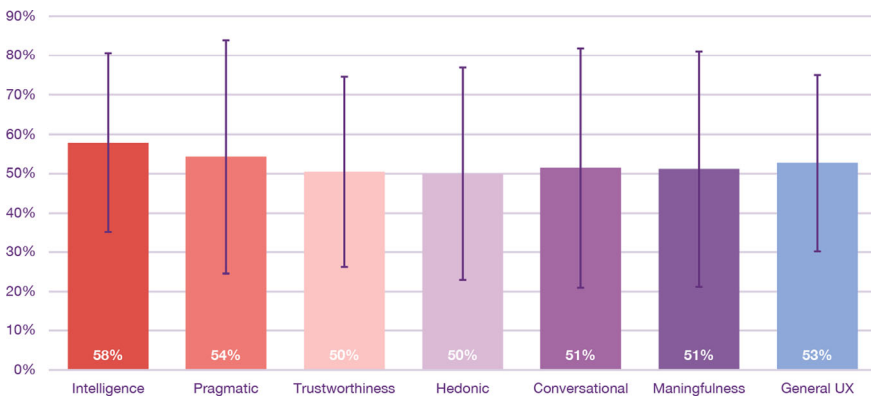


Fig. 6.6 Dimensions scores UK 2023

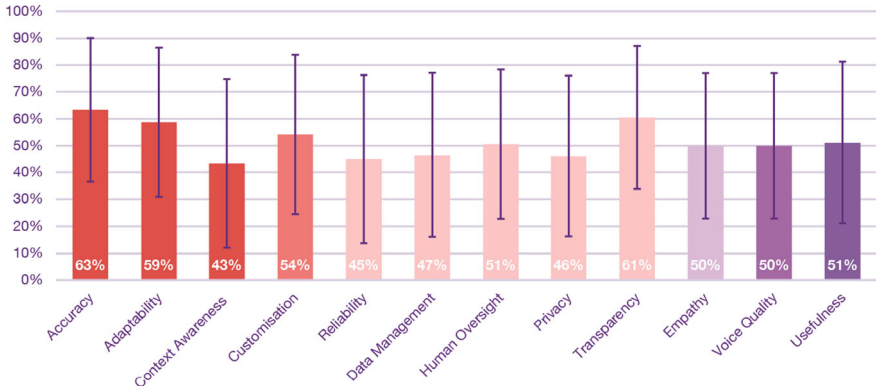


Fig. 6.7 Descriptors scores UK 2023

These results largely align with the overall findings, showing a moderate increase in dimensions related to technological advancements for the UK, such as *intelligence* and *conversational* capabilities. However, there is a significant decline in the other dimensions, particularly those related to *trust*, *empathy*, and perceived *usefulness*, reflecting a broader trend of diminishing user confidence in non-technological aspects of the devices.

The analysis at the descriptor level offers a more detailed view of the situation (Fig. 6.7). Notably, a significant drop in the *data management* score—47% compared to 55% in 2021—contributes to the overall decline in the *trustworthiness* dimension. This suggests that UK respondents in 2023 have become considerably less confident in how smart devices manage their data, highlighting a notable erosion of trust in this area.

Additionally, a 3% reduction in the *customisation* score contributed to the decline in the *pragmatic* dimension. The remaining descriptors showed minimal change compared to 2021, with variations not exceeding 2%.

6.2.4 Performances in the USA

Analysing the results of the first survey for the US (Fig. 6.8) reveals a slightly better performance in *general UX* compared to the overall findings, with a score of 56% versus 55%. This improvement is driven by higher scores in three specific UX dimensions: *trustworthiness*, *hedonic*, and *meaningfulness*. Each of these dimensions shows a 1% increase over the overall average, indicating a slightly higher level of satisfaction among US respondents.

In the detailed analysis of the descriptors (Fig. 6.9), the results closely align with the overall scores, with average differences of around 2%. Notably, the *usefulness* descriptor stands out, scoring 4% higher than the overall average. This indicates that

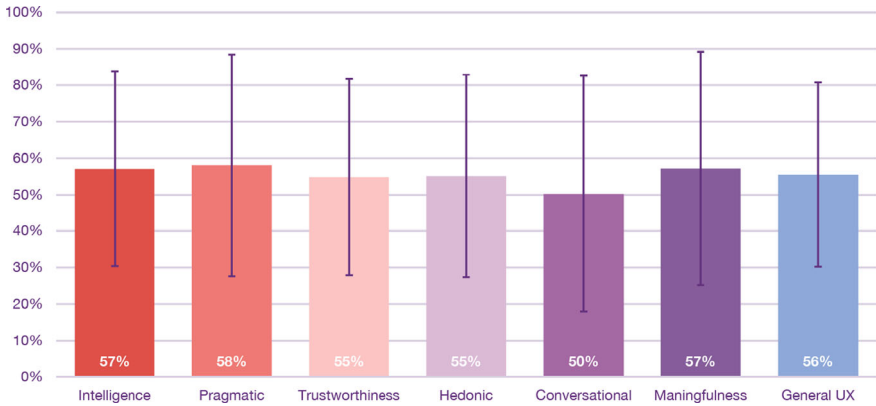


Fig. 6.8 Dimensions scores USA 2021

US respondents have a more positive perception of the actual value of smart speakers in their daily lives compared to the broader dataset.

Two years later, the results of the AIXE questionnaire denote notable changes (Fig. 6.10). The *general UX* score dropped by 2%, from 56 to 54%, indicating an overall decline in user perception. A closer look at the six UX dimensions reveals a pattern similar to that of previous analyses. While *intelligence*, *pragmatic*, and *conversational* dimensions improved compared to 2021, *trustworthiness*, *hedonic*, and *meaningfulness* experienced significant declines. Notably, *trustworthiness* and *meaningfulness* each decreased by 5%.

These findings paint a clearer picture of a trend: despite advances in the technological aspects of smart speakers, the user experience related to trust and perceived value has notably deteriorated. This underscores the non-linear relationship between technical improvements and overall user satisfaction.

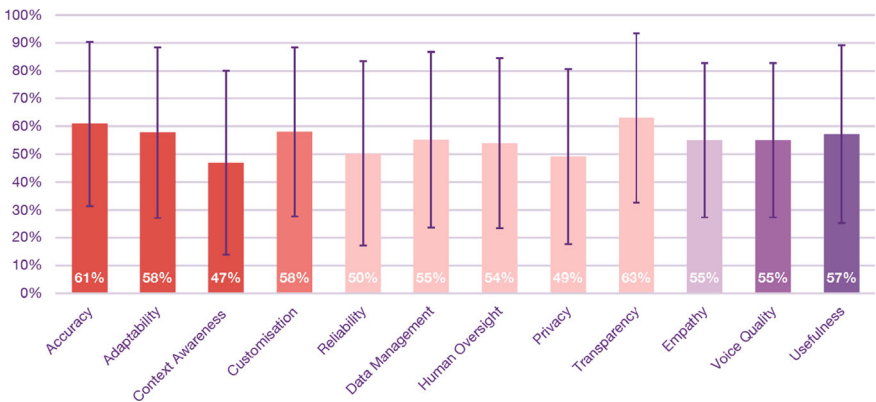


Fig. 6.9 Descriptors scores USA 2021

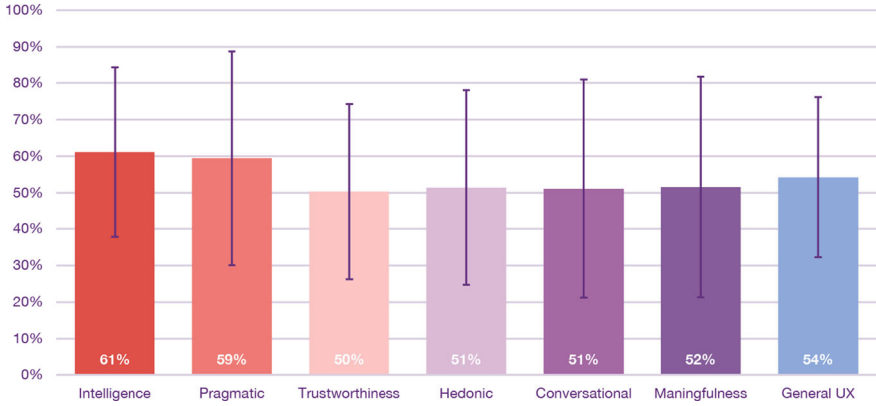


Fig. 6.10 Dimensions scores USA 2023

A deeper analysis of the individual descriptors provides further insights (Fig. 6.11). For instance, *accuracy*—an important marker of *intelligence*—saw the largest improvement, rising from 61 to 65%. However, this improvement contrasts sharply with the 7% drop in *reliability*, which fell from 50 to 43%. This contrast suggests that while users now see smart speakers as more intelligent and capable of delivering accurate responses, their trust in the devices’ overall dependability has significantly eroded. This imbalance between growing intelligence and diminishing trust is a key challenge for future developments.

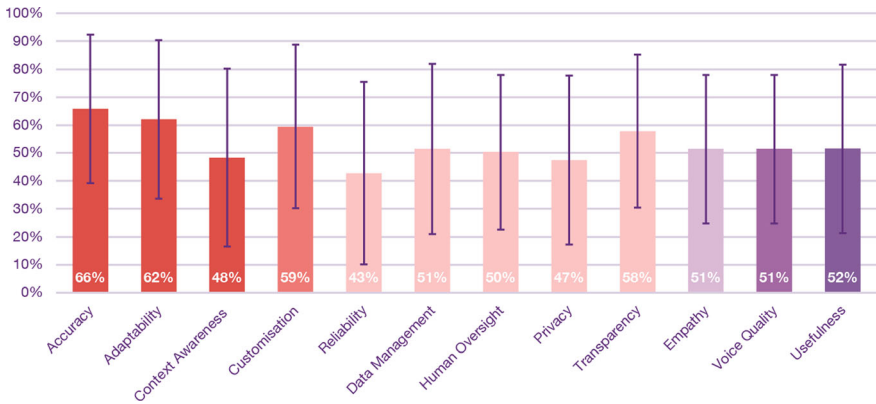


Fig. 6.11 Descriptors scores USA 2023

6.2.5 UX Results by Age Group

The demographic information gathered before administering the AIXE questionnaire enables the differentiation of results by age groups, revealing potential patterns that connect perceived user experience across the six UX dimensions to age. Figure 6.12 illustrates the *general UX* and individual dimension results based on data from 2021 and 2023 in both the UK and USA.

At first glance, a clear trend emerges: as age increases, both the *general UX* and each individual dimension show a noticeable decline in scores. This suggests that older users report lower satisfaction with their smart speaker experience than younger users.

This trend is clearly portrayed by the *general UX*. Users aged 18–29 score 58%, while users between 55 and 65 report 45%, marking 13% of difference. The decrease is almost linear, from 58 to 55% for the 30–44 group to 52% for the 45–54 group.

Gen Z (18–29 years old) registers the highest scores across all UX dimensions, indicating stronger satisfaction. Notably, this generation has a particularly higher opinion about the *pragmatic* dimension (62%), and a good one also for *intelligence*, shared with the 30–44 group.

Millennials (30–44), in general, show similar trends, with all UX dimensions scoring equal or above 50%, though generally lower than Gen Z. *Intelligence* remains the only dimension to reach 60%, while the others fall short. The most significant differences are found in the *conversational* (56% vs. 50%) and *pragmatic* dimension (62% vs. 58%). In other words, Millennials are less satisfied with the way they can adjust the interaction to their needs and preferences and perceive a notably lower quality in conversational interactions compared to Gen Z.

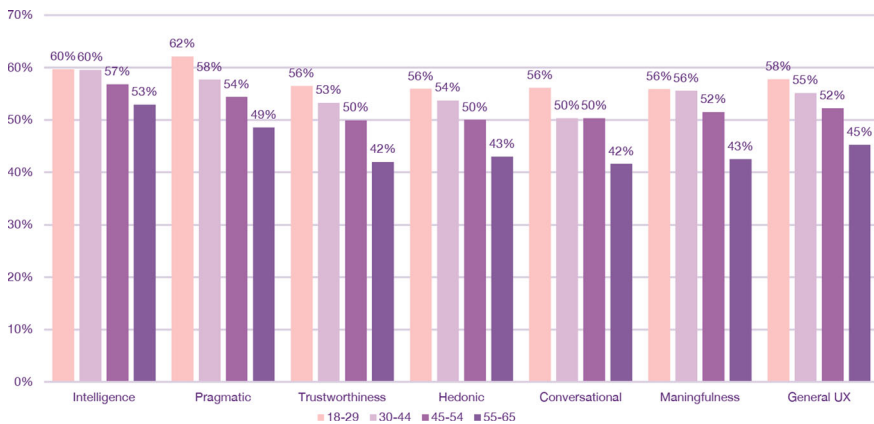


Fig. 6.12 UX dimensions results by age range

For Gen X (45–54), the tendencies are quite similar to those of Millennials, with differences in each dimension around, but not exceeding, 4%. Specifically, this generation shows the most noticeably lower results in *pragmatic*, *hedonic*, and *meaningfulness* dimensions. These findings suggest that Gen X users, compared to younger age groups, exhibit less enthusiasm in the capability of smart speakers to add value to their lives, from a practical and a sense-making perspective, encompassing the aspects related to finding pleasure in the experience.

As previously noted, the 55–65 age group (Boomers) registers the lowest scores across all UX dimensions. Only the *intelligence* dimension stands above the 50% threshold, while the *pragmatic* reaches 49%. The remaining dimensions show significantly lower results, with *trustworthiness* and *conversational* scoring 42%, and *hedonic* and *meaningfulness* reporting 43%. This indicates a marked decline in satisfaction among Boomers, particularly in areas related to trust, conversational quality, and the perceived enjoyment and value of smart speakers in their lives.

6.2.6 UX Results by Device

The final analysis of the dataset focuses on the performance of the three most commonly used smart speakers among respondents: Amazon Echo, Google Nest, and Apple HomePod. The data clearly show that Apple HomePod outperforms the other two devices, with Google Nest slightly ahead of Amazon Echo (Fig. 6.13).

Apple HomePod scores 9% higher in *general UX* compared to its competitors. It shows particularly strong performance in *trustworthiness*, *hedonic*, *conversational*, and *meaningfulness* dimensions, with nearly a 10% advantage over both Google Nest and Amazon Echo. The gap is smaller in the *intelligence* dimension, where Apple leads by 4% over Google and 6% over Amazon.

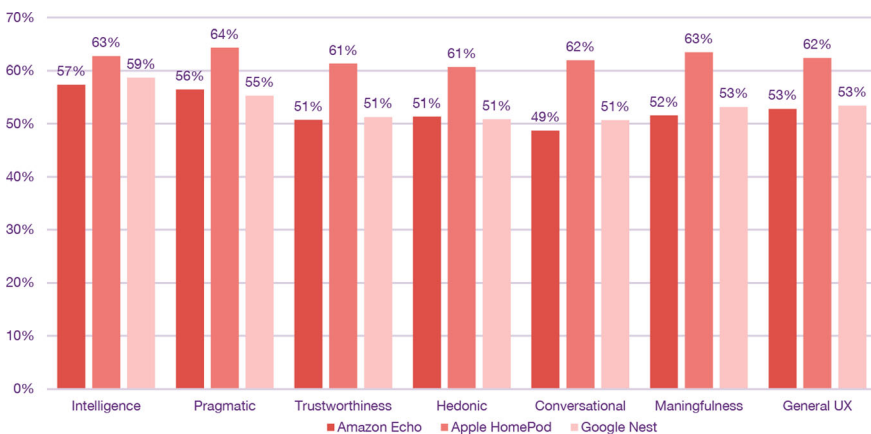


Fig. 6.13 Performances by device 2021 + 2023 data

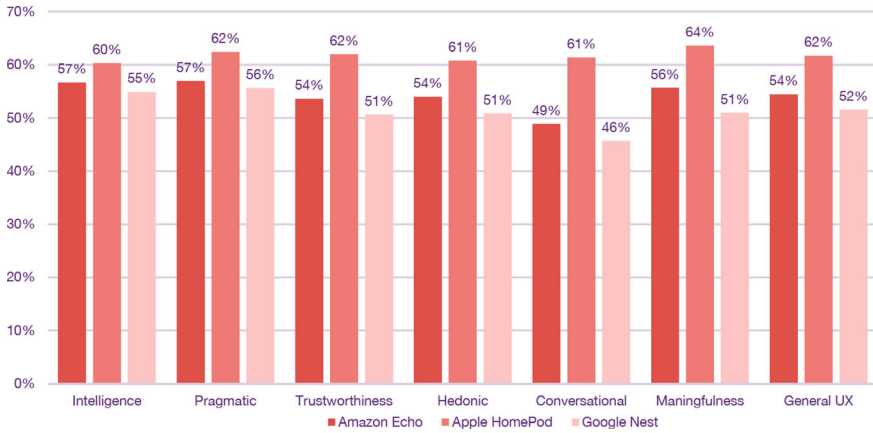


Fig. 6.14 Performances by device in 2021

If Apple, with its devices, leads the comparison, Google Nest and Amazon Echo have very similar results, with the Amazon devices leading over Google only in the *pragmatic* dimension (56% over 55%). Users perceive Google Nest as more intelligent and appreciate the conversational quality more. The other dimensions are absolutely comparable, showing equivalent performances between the two devices leading the market in the UK and USA.

Analysing the performance of the three devices in both 2021 (Fig. 6.14) and 2023 (Fig. 6.15), we can observe how the scores have evolved over time. At a first glance, the average scores remain consistent, with Apple HomePod continuing to outperform in every dimension in both years. Over time, both Apple HomePod and Google Nest have improved their scores across most dimensions, whereas Amazon Echo has experienced a decline.

Focusing on the dimensions closely linked to technological advancements, we see a widespread improvement in the *intelligence* dimension, with Google Nest showing a 6% increase and Apple HomePod a 5% increase. The *pragmatic* dimension saw a significant rise for Apple (+4%), while Amazon and Google experienced a slight decline. Amazon showed losses in the *conversational* dimension, while Google Nest marked a notable 5% increase and Apple 2%.

Trustworthiness and *meaningfulness* show a decline for both Amazon Echo and Apple HomePod. Amazon Echo experienced a significant drop, with *trustworthiness* decreasing by 4% and *meaningfulness* by 8%. In comparison, Apple HomePod’s losses in these areas were more moderate, showing a smaller decline overall. On both dimensions, Google Nest experiences an increase, even if not significant.

Finally, the *hedonic* dimension marks a decrease for Amazon Echo (−5%) and stable scores for Google Nest and Apple HomePod.

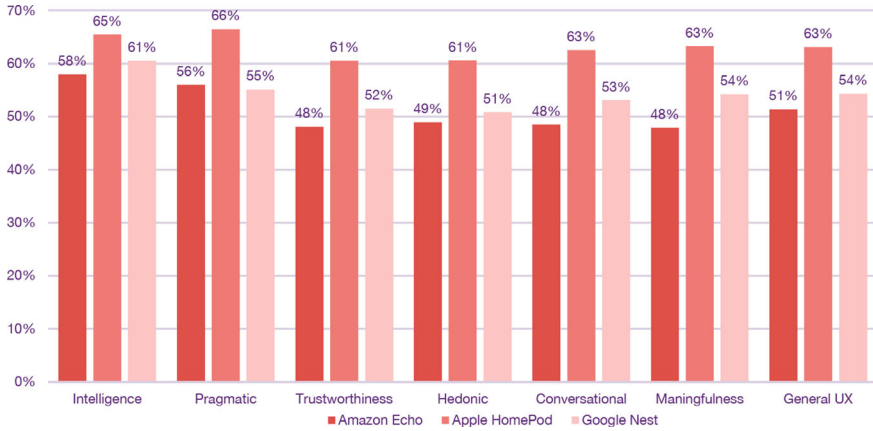


Fig. 6.15 Performances by device in 2023

6.3 Discussion

The study employed the AIXE questionnaire to assess the user experience (UX) of about 1600 respondents in the US and UK who used common smart speakers such as Amazon Echo, Google Nest, and Apple HomePod. This analysis, conducted in 2021 and 2023, provides valuable insights into how users perceive these devices and the factors influencing their satisfaction.

Overall UX Score

The results show that the UX of smart speakers is perceived as slightly above average, with an overall score of 54%. This suggests that while users find their experience generally acceptable, there is room for improvement. The fact that the overall UX score falls short of being impressive indicates that despite their growing presence in households, smart speakers have not yet fully captured the level of satisfaction users might expect from such devices.

UX Dimensions

The study reveals that smart speakers perform best in the *pragmatic* and *intelligence* dimensions. This indicates that users value the peculiar functionalities offered by the AI systems embedded in these devices, appreciating their ability to adapt to various tasks and situations and perform actions accurately. However, the lower scores in *meaningfulness*, *trustworthiness*, and *hedonic* dimensions suggest that users may not find these devices as valuable or enjoyable in their daily lives. These aspects are critical to fostering long-term engagement with technology, and the lower scores indicate that users are not fully convinced of the benefits beyond basic functionality. Indeed, these results confirm the general trends that lead AI-based products to fail because of lack of human factors [5].

Performance Over Time

Interestingly, the overall UX scores remained relatively stable between 2021 and 2023, indicating that users' satisfaction levels have not drastically changed. However, when examining individual UX dimensions, there are notable shifts. Dimensions closely tied to technological advancements—*intelligence*, *pragmatic*, and *conversational*—showed improvements, suggesting that users have noticed developments in how these devices understand, respond to, and interact with them. This positive acknowledgement highlights the impact of ongoing technical updates.

On the other hand, *trustworthiness*, *hedonic*, and *meaningfulness* showed a decline over the same period. This decline points to growing scepticism, particularly around data privacy and the perceived value of these devices in users' lives. As technology becomes more pervasive and the topic is more broadly addressed in contemporary discourses, users may have rising concerns about how their data is managed, which undermines their trust in these devices. However, the results observed in this study portray an inversed tendency with regards of how much people are inclined to entrust the companies implementing AI systems, which is increasing in the latest AI Index Report [6]. Of course, the difference might not surprise, as the AIXE scale has specifically addressed smart speakers and, in the same report 52% of the respondents have also stated that products and services using AI make them nervous, versus the 39% of 2022—with no clearer specification of what *being nervous* might imply. Therefore, while still achieving average results, the general trust toward the companies employing AI is improving, as opposed to what is perceived about the products that materialise it the most, smart speakers.

Performance by Country

The study also reveals regional differences in UX perception. Users in the UK reported slightly lower overall UX scores and showed a declining trend in trust, particularly in relation to data management. This could reflect heightened concerns about privacy or perhaps differences in European regulatory environments.

In contrast, US respondents displayed slightly higher UX scores and a higher perception of *usefulness*, indicating a more positive view of smart speakers. The divergence between these two regions highlights the relevance for device manufacturers to address regional expectations and concerns, especially those related to data security and privacy.

Performance by Age Group

The study underscores significant differences in smart speaker satisfaction across age groups. Younger users—particularly those under 30—reported significantly higher levels of satisfaction. This age group is likely more comfortable with evolving technologies and adapts more easily to innovations in smart speaker functionalities.

Conversely, older users reported a noticeable decline in satisfaction as age increased. This could be due to a variety of factors, including less familiarity with technology, higher expectations for tangible benefits, and more scepticism about data privacy and trust. These generational differences suggest that manufacturers need to consider how they market and develop devices to cater to diverse user needs.

Performance by Device

When looking at performance by device, Apple HomePod consistently outperformed its competitors, Amazon Echo and Google Nest. Apple's devices achieved higher overall UX scores, particularly in dimensions like *trustworthiness*, *hedonic*, *conversational*, and *meaningfulness*, positioning Apple HomePod as the leader in delivering a well-rounded user experience.

Google Nest, however, demonstrated marked improvements over time, especially in *intelligence* and *conversational* dimensions, signalling progress in how users perceive its technological sophistication and interaction capabilities. Meanwhile, Amazon Echo, while still a popular choice, showed a decline in overall UX, with users experiencing drops in *trustworthiness* and *meaningfulness* perception, suggesting that Amazon's device may have struggled to keep up with evolving user expectations. The sharp decrease in trust-related dimensions reflects growing concerns about *privacy* and *data management*, areas where Amazon could focus future improvements to regain user confidence.

6.4 Conclusion

The declining scores in *trustworthiness* and *meaningfulness* reflect a gap between technical advancements and user experience. While devices are becoming more intelligent, users may not necessarily perceive them as trustworthy or valuable. This growing disconnect is particularly evident in the declining satisfaction among older users and those in the UK, underscoring the importance of addressing concerns related to data privacy, security, and the meaningful integration of these devices into users' lives.

The study also highlights the critical importance of focusing on broader UX dimensions beyond technical features. As smart speakers evolve, *trustworthiness*, *meaningfulness*, and the *hedonic* dimensions are essential to a positive user experience. These elements will become even more critical as users grow more sophisticated in their expectations of technology.

Additionally, the generational differences in satisfaction suggest that younger users may be more accepting of new technologies, while older users require more demonstrable value and trust. This generational gap in user experience highlights the need for device manufacturers to tailor their approaches to different demographics.

Finally, the evolving nature of the smart speaker market calls for continuous improvement. The study suggests that manufacturers must prioritise addressing user concerns around trust, data privacy, and perceived value to enhance satisfaction and maintain market leadership.

6.4.1 Limitations and Future Research

While the study provides valuable insights, it is not without limitations. The sample is restricted to respondents from the UK and USA, which may limit the generalizability of the findings to other regions with different cultural and technological landscapes. Additionally, the data are collected after a period of, at least, 14 days and not right after a task has been completed. This might introduce biases due to the overall users' perception of the device, meaning that this is not the precise assessment of single tasks but reflects the respondent's broader opinion. As mentioned, the data also have a quantitative nature which, while providing terms for comparability, it only allows for the researchers' speculations about the motivations behind the ratings.

Future research could expand the demographic scope to include a more diverse range of countries and cultural contexts. Investigating the specific factors that contribute to trust and meaningfulness could provide deeper understanding and inform more targeted interventions to enhance user satisfaction across all age groups. Moreover, additional qualitative investigations can enrich the overall picture.

References

1. Smart speaker market size, share, trends, opportunities & forecast. In: Verified market research. <https://www.verifiedmarketresearch.com/product/global-smart-speaker-market-size-and-forecast/>. Accessed 3 Sept 2024
2. Smart speaker market size, share and growth report, 2030. <https://www.grandviewresearch.com/industry-analysis/smart-speakers-market>. Accessed 3 Sept 2024
3. Smart speaker market size, share & industry trends [2032]. <https://www.fortunebusinessinsights.com/smart-speaker-market-106297>. Accessed 3 Sept 2024
4. Spallazzo D, Sciannamè M, Ceconello M (2019) The domestic shape of AI: a reflection on virtual assistants, pp 52–59
5. Yildirim N, Oh C, Sayar D, et al (2023) Creating design resources to scaffold the ideation of AI concepts. In: Proceedings of the 2023 ACM designing interactive systems conference. Association for Computing Machinery, New York, NY, USA, pp 2326–2346
6. Maslej N, Fattorini L, Perrault R et al (2024) The AI index 2024 annual report AI. Index Steering Committee, Institute for Human-Centered AI, University of Stanford, Stanford, CA

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 7

Conclusions



Abstract The conclusive chapter summarises the main takeaways of the book, highlighting the primary contribution of the Meet-AI research project to the design field. It further highlights the limitations of the study and suggests future research paths.

7.1 Summarizing the Contribution

The Meet-AI project represents one of the first systematic attempts to understand and assess the UX of AI-infused products. It entailed understanding the peculiarities of such products to frame their inner complexity. Then, through an in-depth exploration of existing tools and methods for UX assessment, the research identified a critical gap: they are inadequate to holistically assess the UX of such dynamic systems.

As a result, the project moved towards understanding which UX dimensions are essential to properly frame the evaluation. This entailed different stages of literature review, followed by a validation process that involved expert users, according to a user-centred approach.

This step produced a list of UX dimensions and descriptors, which, through an iterative and collaborative process among the researchers, culminated in the foundation of the construction of a scale for the evaluation of AI-infused products (AIXE).

The AIXE questionnaire was then tested and refined through multiple iterations, ultimately taking its final form. It has since been employed to evaluate the UX of domestic smart speakers, providing valuable insights into how these products perform in real-world contexts over time.

7.1.1 *Contribution to Design and UX Assessment*

The Meet-AI project has made significant contributions to the field of UX assessment and design, particularly in the context of AI-infused products. A major achievement of the project was the identification and incorporation of four new UX dimensions

that are critical for evaluating AI systems but are often overlooked in traditional UX methods: *intelligence, trustworthiness, meaningfulness, and conversational*. These dimensions are central to understanding how users engage with AI products on a deeper level, where factors like trust and the perceived intelligence of the system significantly influence the overall experience.

Overall, the AIXE scale was designed to capture six core UX dimensions, as the traditional pragmatic and hedonic ones proved worthy to be retained and were articulated consistently with the objects of the investigation. This multidimensional framework allows for a more holistic evaluation of AI products, going beyond the conventional aspects that characterise other products to capture the essence of the issues and concerns that AI systems bring to the experience.

The AIXE scale contributes to the field of AI and UX design by offering a structured and reliable tool for assessing the complex interactions between humans and AI technologies. Its development provides a quantitative method for evaluating AI systems, making it possible to gather statistically significant data on user experiences. This tool is valuable not only for academic researchers but also for designers, engineers, and developers, who can use it to assess both market products and prototypes, helping them make informed decisions about product design and iteration.

By introducing this scale, the Meet-AI project has addressed a critical gap in the field of UX assessment, offering a framework that captures the nuances of AI interactions while maintaining the rigour required for benchmarking and product evaluation. This ensures that the scale can be used for a wide range of purposes, from redesign and improvement of existing products to making informed comparisons of new AI-based artefacts.

7.1.2 Implications for AI-Infused Products Design

The findings from the Meet-AI project have far-reaching implications for the design of AI-infused products. One of the most relevant insights is that UX evaluation of AI systems must go beyond assessing their technical capabilities. While technical performance is essential, AI products must also deliver meaningful and trustworthy experiences to users. The project highlights that these dimensions—*trustworthiness* and *meaningfulness*—are not just desirable qualities but are crucial to the overall success of AI-infused products, as they directly impact user acceptance and long-term engagement.

The AIXE scale provides a user-centred framework for guiding the design and development of future AI products. This framework ensures that AI innovations are better aligned with user needs, expectations, and ethical concerns, addressing the core human-centred issues that are often neglected in the pursuit of technical sophistication. By focusing on these dimensions, the Meet-AI project reinforces the necessity for a change in how products integrating AI are designed—one that prioritises meaningful interactions and user trust as much as technical efficiency.

The project also underscores the effectiveness of the AIXE questionnaire in providing a layered approach to UX assessment. The ability to move from general UX scores to more detailed dimensions and descriptors offers flexibility in interpreting the data. Designers and developers can use this tool to conduct broad benchmarking of AI products or focus on specific aspects of the UX to identify pain points or areas for improvement. The scale's granular level of analysis allows teams to pinpoint the factors that most significantly enhance or hinder the user experience, providing insights to conduct well-oriented qualitative investigations.

In sum, the Meet-AI project provides a roadmap for creating AI-infused products that are both functional and meaningful. This contribution sets the stage for future research and experimentation, fostering a discourse that places **people** and **human factors** at the heart of AI innovation.

7.2 Strengths, Limitations and Future Opportunities

As the research presented in this book has underlined, the challenges and complexities of AI-infused products are manifold and extend beyond what could be addressed within the scope and timeframe of the Meet-AI project. Therefore, this section aims to draw some conclusions, highlighting which aspects were successfully covered and which expansions and further improvements can be envisioned. Starting from the most punctual, the interesting issues and opportunities about UX evaluation methods and AI systems identified throughout the research are retraced and discussed in the light of project outcomes.

7.2.1 *A Broader Access*

The first element requiring attention concerns the dissemination of the AIXE scale. Indeed, this open access publication is a first step toward making this resource and all the research findings publicly and freely available.

However, some barriers to a broader adoption can be recognised. Currently, submitting the questionnaire to a sample of users requires manual implementation into digital platforms, and a level of statistical expertise is needed to process the results. This presents a challenge for smaller organizations or individual practitioners who may not possess the technical resources or know-how required to fully utilize the scale.

To facilitate the employment of the scale, it would be beneficial to develop a preset digital support that could present the questionnaire and automate the processing and visualization of data.

A user-friendly dashboard could be created to provide a better understanding of how the products performs at different levels of granularity. From the overall UX score to detailed analyses of the six dimensions, 12 descriptors, and 33 items, the

key indicators derived from the structural model of the evaluation scale would be clearly measured. Such a system would make it easier for various stakeholders to interpret the data and gain actionable insights without needing extensive knowledge of statistics. For instance, the dashboard could include the automatic calculation of key metrics, such as Cronbach's Alpha, to evaluate the reliability of the results and, in case of low values, prompt the AIXE users to consider collecting more data.

By integrating these enhancements into a familiar format, such as a spreadsheet tool or a web-based application, the AIXE scale would become more accessible and scalable, thus encouraging wider adoption among researchers, designers, and companies.

7.2.2 Transcending Conventions to Embrace Evolution and the Whole Design Process

To avoid introducing a further layer of novelty and to facilitate a broader adoption, the questionnaire format was selected to materialise the evaluation scale. Being a conventional tool for assessment, it is familiar both to users and companies and allows for easy implementation, elaboration of data and statistical validation. Recurring to a well-known format, freed the space for focusing on the core aspects of the novel method to evaluate AI-infused products: the dimensions and descriptors that could best express their UX, and how they could be granularly captured through specific items.

Now that the Meet-AI project has addressed these foundational aspects and provided actionable insights, further experimentations might take a closer look at alternative ways of interrogating users and gathering useful information.

In particular, the initial stages of the research have highlighted how longitudinal studies may benefit from further exploration. Indeed, they could provide interesting opportunities to address AI-infused products as they would capture how the quality of the experience might evolve over time.

For this purpose, user data could be collected in different ways, different forms of diaries have been generated to let users report their experience through thorough documentation. As well, activity logs could be integrated in the devices to automatically get data that are not filtered by the users' subjectivity. These methods would help track real-time user interactions and gather insights that traditional questionnaires might overlook, such as subtle shifts in trust, engagement, or adaptability.

To address the evolving nature of AI-infused products (one of its objectives), the AIXE scale is built to capture the essential aspects of human-AI interaction, encompassing the characterising traits that can change and improve over time. Currently, the questionnaire has to be re-submitted at intervals of time (as shown in Chap. 6) to track the evolution of the UX, which is a highly recommended practice for the particular kind of products it targets. Of course, this requires additional work for

UX researchers and designers, but it also guarantees the possibility of iteratively improving the studied artefacts, which might be necessary regardless.

An additional interesting domain for exploration is the application of the research findings to the very early stages of the design process. While this has been recognised as an essential issue that would dramatically reduce the risk of failure of AI-based systems, producing tools to support this stage was beyond the possibilities of the Meet-AI project. Nonetheless, we deem it important to underline the crucial impact that such a research activity would have.

Starting from the essential qualities, dimensions, and descriptors outlined in this book, some meta-design principles could be inferred and translated into practical tools to support the envisioning of AI-infused products, as well as the early prototyping stages.

Yet, further elements should be taken into consideration, like the necessity to involve different kinds of expertise and the difficulties in prototyping AI systems. If successfully tackled, the emerging guidelines and tools would surely benefit the professional field, facilitating their operations and the collaboration among different professional figures considerably. Moreover, educational institutions could also be positively impacted by these results as they could adequately steer the preparation of design and engineering students who will deal with AI systems as objects of their work. In both cases, harnessing the potential of AI-infused products could inspire meaningful innovation, that eventually could produce enhanced human experiences in different domains.

7.2.3 A Multidimensional and Multi-method Approach

As anticipated in the book, the choice of the questionnaire as a scale format implied a commitment to a quantitative evaluation method. One of the reasons, in addition to those mentioned in the previous section, was to avoid it being a niche, non-reusable tool. Indeed, it allows for rigorous, generalisable results that prove very helpful in measuring relevant UX qualities and use these values for comparisons. Whether they are aimed at observing product performances over time or benchmarking them with others, these high-level insights do not have the depth needed to uncover actionable information for product improvement.

As emerged in the thorough review to assess the state of the art of UX evaluation methods, there is a lack of methods collecting rich qualitative data, and therefore offering a comprehensive multidimensionality of UX qualities.

Although it was not possible to further deepen and test a multi-method approach of which AIXE could be part, we can foresee how beneficial it would be, and highly recommend it. At least one qualitative tool, such as semi-structured interviews, should be introduced to follow-up and complement the quantitative nature of AIXE, to get the granular motivations behind the rates. Although the validation of the scale attests to its comprehensibility and reliability, there are still facets that can be interpreted

in multiple ways, and grasping them can provide invaluable advantages for product improvement.

In general, combining different UX evaluation methods can provide a more comprehensive understanding of the actual user experience, but identifying which methods work well together and how to effectively integrate and analyse data from multiple sources remains to be explored.

Considering the main features and challenges that AI-infused products bring to the attention of the UX field, the core qualities identified are actually covered by the AIXE scale. If we think of uncertainty, active agency, or lack of transparency, understanding, human factors in general, or new interaction paradigms accounting for decision-support and recommendations, it is recognizable how these are investigated by different items from different angles, but always putting the respondent's perceptions in the spotlight.

However, the issues reported to have a direct reflection in the interface, are not adequately addressed by the final questionnaire. The nature of this problems, though, lies in the betrayal of basic ergonomic and heuristic principles that UI and UX designers are well accustomed to, and for which traditional UX methods are a solid reference. As it turned out several times throughout the research, the pragmatic dimension of the user experience still has a primary importance and, even if it is not portraying the unique features of AI systems and is not well represented in AIXE items, a complementary investigation of basic usability principles would be valuable.

Another point for integration might relate to the conversational interface, if this is particularly relevant for the product being evaluated. As discovered in the systematic review, the conversational dimension can articulate in several ways that account also for more technical aspects that are not covered in AIXE, like the quality of the responses, the capability to entertain a dialogue, linguistic properties, and inclusiveness toward dialects and disabilities.

Further developments might also account more specifically for multi-device and multi-user experiences.

Finally, as the complexity of AI-infused products actually extends to several areas of expertise, it would also be interesting to involve different tools and professionals for a multidisciplinary evaluation.

7.2.4 A Broader Range of AI-Infused Products

The AIXE scale has been designed to capture the essence of AI-infused products and be comprehensive and flexible. Through the several stages of research, a good number of qualities has been thoroughly examined to meet these requirements, resulting in the articulation of six dimensions, 12 descriptors, and 33 items. While pointing at the unique traits that affect the UX with AI-infused artefacts, they are depicted in a sufficiently general manner to encompass multiple contexts, situations, and product typologies.

Nonetheless, the original target was limited to physical products—as they could be more problematic and nuanced. The scale was validated specifically with smart speakers, and a subsequent study tested the questionnaire on these devices as well.

Expanding the range of products on which to apply the AIXE scale would undoubtedly be interesting. Other AI-infused products could be explored, ranging from autonomous-vehicles to home appliances, and further industry contexts could be examined. Furthermore, it would be worth extending the application of AIXE to digital AI-infused products, as they may present very similar UX challenges.

Indeed, using AIXE in very specific niches of products can be an opportunity for exploring the modularity and versatility of the scale to different necessities and would strengthen the premises on which it was built.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

