International Association of Societies
of Design Research Congress 2023
LIFE-CHANGING DESIGN

Politecnico di Milano
9-13 October 2023

AIXE. Building a scale to evaluate the UX of AI-infused products

Sciannamè, Martina*[a]; Spallazzo Davide[a]

[a] Politecnico di Milano - Department of Design, Milano, Italy
* martina.scianname@polimi.it

doi.org/10.21606/iasdr.2023.355

Despite the diffusion of artifacts integrating AI systems, current UX evaluation methods are not yet prepared nor comprehensive enough to include the unique traits characterizing them. That is the main premise of the Meet-AI project, which developed a new method to assess AI-infused artifacts. The contribution traces all the research steps that have been necessary to build AIXE, a specific and comprehensive scale framed as a questionnaire with 33 items and aimed to support the understanding of the core UX qualities of this spreading technology. Specifically, it presents the three main phases of the research, which include: (i) the exploration of the state-of-the-art of current UX methods and reflections about AI-infused objects, (ii) the identification of dimensions and descriptors (second and first order variables) to construct an attitude scale using mixed methods sharing a human-centred approach, and (iii) the validation of the scale with an exploratory and a confirmatory factor analysis. AIXE represents one of the first attempts for the design field to approach the development and, primarily, the assessment of AI-infused products and services from a UX standpoint. In particular, it is aimed at guiding practitioners in understanding and properly evaluating artifacts integrating AI capabilities based on their characterizing features. This should encourage a more aware and effective design of such artifacts, emphasizing their core traits.

Keywords: AI-infused products; UX dimensions; UX evaluation method

1 Introduction

AI-infused products are game-changing as they are going to affect people’s lives and behaviours in unprecedented manners. New ways of understanding these kinds of interactive systems and their agency are required, as they question the ability of human-centred design to effectively give form to this technology (Giaccardi & Redström, 2020).

The paper discusses the creation of the AIXE (AI eXperience Evaluation) scale, result of the Meet-AI research project, funded by the Design Department of Politecnico di Milano. It is a specific and comprehensive evaluation method framed as a questionnaire with 33 items that aims to aid the comprehension of the core UX qualities of this rapidly spreading technology.

It describes the three main phases of the research: (i) the exploration of current UX methods and reflections on AI-infused objects, (ii) the identification of dimensions and descriptors (second and first-order variables) to construct an attitude scale using mixed methods with a human-centred approach, and (iii) the validation of the scale using exploratory and confirmatory factor analysis.

The presented study and its principal output – the AIXE questionnaire – address designers, UX researchers, and companies involved in designing and assessing AI-infused systems. Its main contribution is to support the UX testing of high-fidelity prototypes or products needing improvements for a minimum of one- or two-week period. However, the basic assumptions on which the scale is built (the selected dimensions and descriptors) may also inform the design process from the earliest stages.

2 AI-infused artifacts and their UX evaluation

Introducing a move from the paradigm of embodiment (Dourish, 2001a) to alterity, AI-infused artifacts may be understood as counterparts, as suggested in Hassenzahl and colleagues' definition of these products as otherware (2020).

Nevertheless, new interaction paradigms must be introduced to see interactive items as other entities rather than users' extensions. We chat with voice assistants rather than physically interacting with them. It is no longer sufficient to rely solely on the system's robustness. We must place our trust in them since most AI-infused products' operations are typically opaque to end users. Their results are mainly unpredictable, especially when systems like deep learning are involved. Novel kinds of interaction suggest a new dynamic between people and machines, which inevitably influences how people view these devices and the experiences they offer.

AI-enhanced systems can be proactive and visibly demonstrate their agency to end users by learning, reflecting, and conversing. These systems go beyond delegated agency (Kaptelinin & Nardi, 2009). They can disappoint users, act independently, or – even better – select the ideal solution to the issue at hand.

These systems are experiencing an ever-growing market success and are expected to grow exponentially. Nevertheless, the HCI community and the design discipline have only recently developed a comparable interest. AI has been flaunted as a new material for designers (Antonelli, 2018; Holmquist, 2017). Some studies how design may approach machine learning (Dove et al., 2017) and how virtual assistants are used in real-life (Sciuto et al., 2018). Others focused on conversational interfaces, reviewing the UX research (Zheng et al., 2022), or evaluating the user experience enabled by voice-based interactions (Kocaballi et al., 2018).

However, research on the user experience provided by AI-infused devices is still required. AI-enabled products (such as the Amazon Echo family) may appear as innocuous ornaments. Indeed, these technologies are typically viewed as gadgetry satisfying users' desire for novelty rather than playing a substantial part in their lives. Still, they reveal an inherent complexity amplified by features that make such products challenging to analyse from a UX standpoint. Similarly, they are not remarkable regarding interaction quality, occasionally causing discomfort and an inability to fully realize their potential (Sciuto et al., 2018).

Accordingly, tools and methods to assess the UX experience enabled by AI-infused systems may help UX researchers and companies understand the users' points of view and implement better solutions. Indeed, to the best of the authors' knowledge, there are still no methods that specifically address these issues.

Looking at the most recent systematic literature reviews on UX evaluation methods (Bargas-Avila & Hornbæk, 2011; Lachner et al., 2016; Pettersson et al., 2018; Rivero & Conte, 2017; Vermeeren et al., 2010), it is clear that none of them has been specifically designed to assess the UX of AI-infused systems. Furthermore, current UX evaluation methods are inadequate to holistically assess the user experience these devices enable (Spallazzo et al., 2020). Indeed, they may provide a general understanding of the UX but do not analyse the core of such systems.

UX research has not explicitly addressed AI-infused products and services. A novel, bespoke method to address such systems is explicitly needed to guide their development and improvement.

From these premises, the main aim of Meet-AI is to understand better the relevant UX qualities that can describe artefacts integrating AI systems and provide a new method to assess them.

3 Methodology

3.1 State-of-the-art exploration

Even though preliminary secondary research has contributed to developing the hypothesis at the basis of the Meet-AI project, the inquiry began from a thorough understanding of currently identified and agreed qualities for describing the UX of products and services, both in general and explicitly targeting AI-infused applications.

To this end, a twofold strategy has been employed. On one side, a wide-range critical analysis of qualitative and quantitative UX evaluation methods was meant to get a comprehensive picture of the state of the art in the design field and related social sciences experimentations. The research was limited to contributions from the entry of "UX evaluation" and "UX assessment" keywords and published in the ACM Digital Library and Springer Link between 2000 and 2020. The All About UX repository has also been a source for integrating methods that might have been missed. Each of the five researchers involved in the project simultaneously but independently explored and examined the identified scales and methods, collectively reaching a total of 129. These have been analysed according to various criteria (Spallazzo, Sciannamè, et al., 2021), primarily dimensions, general qualities describing people's experience of products, and descriptors, the specific features explaining the nuances of such overarching qualities; but also the collection method(s) (tools and modalities used to retrieve UX evaluations); the possible triangulation of multiple different methods; whether a lab or field context is required; the support materials used; the qualitative or quantitative or mixed nature of the investigation; the product's development phase (concept, early prototype, functional prototype, or market level), and associated period of experience (before use, after an episodic interaction, an accomplished task or long-term utilization) in relation to which the evaluation can be carried out; the object(s) of study; the evaluators required (single user, groups, expert users), and the level of consistency with AI-infused products as perceived by the researchers. Finally, spaces for references and personal remarks were added.

Conversely, an exploratory literature review on the intersection of AI, interaction design, and HCI was performed to explore the unique features of AI-infused products that could not yet be covered by current UX evaluation methods, as stated in the premises of the Meet-AI project. For this task, the researchers qualitatively investigated how the relationship between people and AI is addressed in the literature using a snowball sampling approach as, at the beginning of the study, a UX and interaction design discourse on AI was in its infancy and related scholarly contributions were scarce. They expanded the research into related fields, such as HCI, computer ethics, and AI, based on three broad thematic strands: non-human intelligence, emotion, and meaning. Later, the exploration brought to light further relevant issues in the current debate, which were then included for a complete overview. These are conversational interactions and ethical implications.

3.2 Scale construction

Completed the in-depth study of currently used qualities, some clarity emerged on the structure that most commonly characterizes UX evaluation models: they are usually based on UX dimensions and relative descriptors inside a measurement framework. From a statistical point of view, they correspond to latent constructs (latent variables). Meanwhile, if more detailed questions are used to assess the object of study, these describe the manifest or observable variables. Manifest and latent variables can be associated, and factor loadings can express their strength.

Hence, a conceptual model for building the scale could be depicted. The hypothesis formulated on the nature of the links between latent and manifest variables contributes to the selection of methodological landscapes. To develop the AIXE scale, a reflective hierarchical approach characterized by a third-order model was adopted (Figure 1). The reflective approach assumes that the latent constructs are well defined in the respondent's mindset. In other words, the choice of the evaluation level is an expression of the idea of the constructs of the respondent.

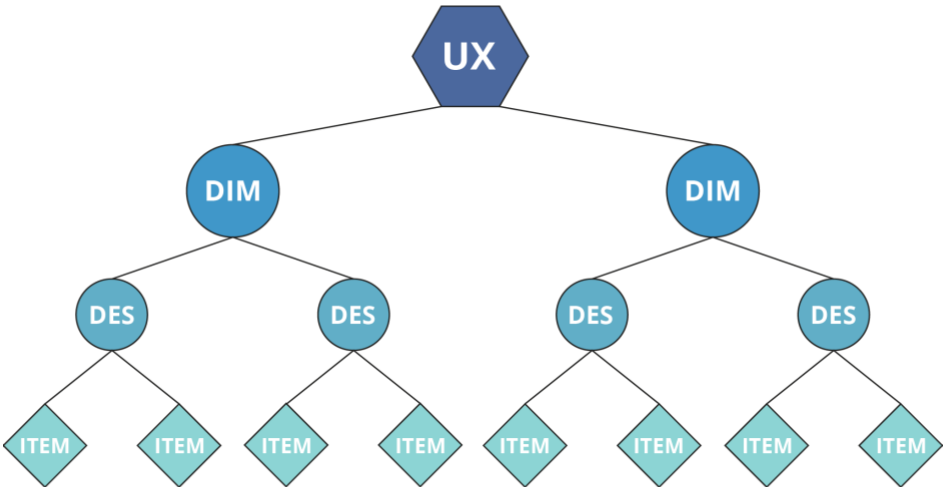


Figure 1. Path diagram portraying the conceptual model underlying the scale.

It 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 (Jöreskog, 1978), 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).

Therefore, we introduce the main research question: What dimensions and descriptors govern the UX of AI-infused artefacts? As several thematic possibilities could be recognized from the analysis of current UX evaluation methods and AI-related literature, further reasoning and research actions were needed to properly identify first- and second-order variables.

Specifically, the researchers sought an approach to defining the relevant dimensions and descriptors characterizing AI-infused products that could preserve as much objectivity as possible. This is why they opted for co-creation and engineering of the selection of the latent variables so that the final scale could also be statistically reliable.

3.2.1 Survey to validate and co-create dimensions and descriptors

A mixed-methods and multi-level protocol pointing in this direction was established. The first, central step was a survey intended to verify the assumptions about the dimensions describing products integrating AI systems (namely those identified as prominent in the previous stage: pragmatic, aesthetic, hedonic, affective, intelligence, trustworthiness, conversational, meaningfulness), and to elicit creative inputs to implement the non-comprehensive set of descriptors gleaned from the literature review. It was sent in a digital form to a total population of 110 students from MSc in Digital and Interaction Design and 47 young researchers from the Design Department of Politecnico di Milano, and reported a response rate of 26.75%, as 42 out of the 157 contacted responded. The respondents were considered advanced users, as they were familiar with using AI-infused products but also had a developed sensitivity and comprehension of the design of interactive objects.

The purposes of the survey were plainly stated at the beginning, along with precise explanations and examples of the typology of artefacts to be assessed (AI-based smart speakers, learning thermostats, and smart cams) and any proposed concepts (i.e., the dimensions). Subsequently, the inquiry focused on two main parts: one seeking to populate a consistent set of descriptors and the other requiring feedback on the UX dimensions depicting AI-infused artefacts. Specifically, after carefully portraying each dimension, an invitation to express attributes (adjectives, nouns, or verbs) that qualify the UX accordingly could suggest listing at least three or, alternatively, two positive and two negative features concerning the possible difficulty of providing heterogeneous answers. Then, straightforward questions shift the attention to the dimensions themselves as elements to be evaluated concerning their effectiveness in outlining AI-infused artefacts and their relevance for this purpose. As well, additional suggestions to integrate the proposed ones were welcomed. Finally, some profiling information was collected.

3.2.2 Data analysis and inter-coder agreement to select descriptors and dimensions

The analysis of the survey results for the dimensions was both quantitative (with the limitation of a small number of responses), computing the ratings for the different questions to understand AI-infused products' performances, and qualitative, as a content analysis was performed to extract more subtle indications. Further research steps were necessary to derive information about the suggested descriptors. First, the provided responses were processed to attempt their homologation to the original requests. Two researchers independently refined the outputs and then confronted them to compile a shared and uniform list (Spallazzo & Sciannamè, 2021). Sentences were reduced, and answers were translated to obtain English single-word descriptors. All entries were reported on a board, differentiated according to the study dimensions, and eight affinity maps were generated to better understand the polished results. Figure 2 shows an example. They allowed to visualize semantic concentrations, synthesize repetitions, filter out-of-context responses, and extrapolate univocal descriptors. Finally, an intercoder evaluation (Creswell, 2014) involved all the researchers in assessing the obtained unambiguous results, based on the consistency of the descriptors with the related dimension and their relevance for AI-infused products. Each descriptor was displayed with its frequency of occurrence inside the dimension for which the respondents suggested it. The descriptors from the literature review (L) were also added for a comprehensive overview. As judges, the researchers had to rate each one on a scale from 1 (not consistent/relevant at all) to 4 (very consistent/relevant). The results have been examined by computing the mean and z score for each descriptor and in accordance with each parameter (consistency and relevance), and they are available at (Spallazzo, Ajovalasit, et al., 2021).

Ultimately, the most significant were extracted after calculating quartiles and comparing the relevance z scores of the descriptors. Yet, 134 descriptors were included in the fourth quartile (>75%). To limit this number, only the 36 "golden" descriptors – those receiving the maximum score from all the researchers – were retained and used as a basis for an internal workshop within the Meet-AI team to conclusively elaborate a first draft structure for the scale to be built.

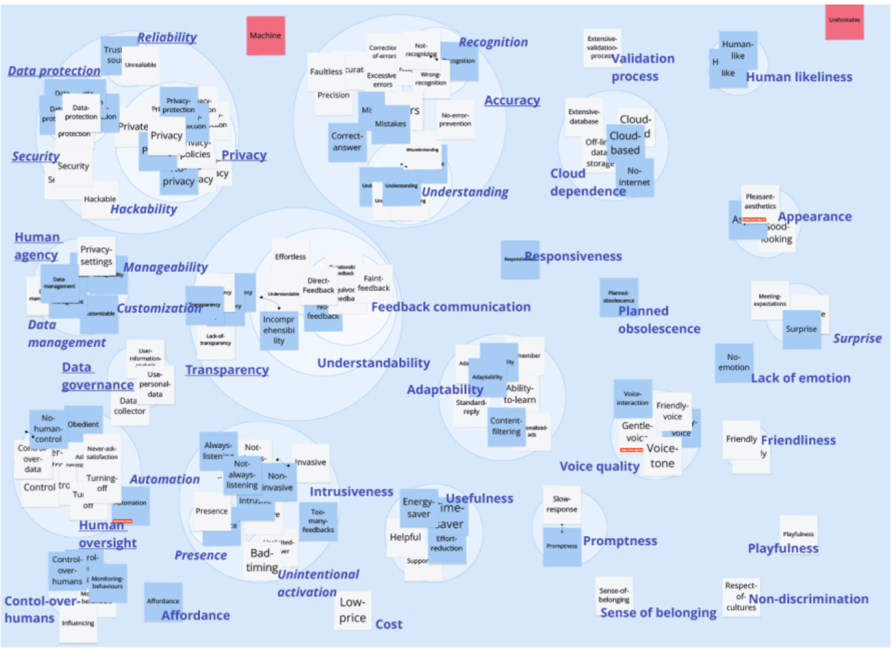


Figure 2. Affinity map of the descriptors from the Trustworthiness dimension. The picture provides an overview of the process, with the collected descriptors on post-its and the extracted ones written in blue.

3.2.3 Items and questionnaire elaboration

A board was used to display the "golden" descriptors with their related dimension on post-its to promote a collective discussion among the researchers to categorize them according to their perceived likelihood of being part of the UX assessment scale. Then, for each of the selected descriptors, a set of questions illustrating several potential nuances of the terms have been developed to serve as a foundation for the scale.

The first questionnaire to be tested was ultimately prepared using the 65 items that the researchers determined to be sufficiently clear and distinct from one another.

In light of recent developments, assigning a relationship of order to the scoring system is preferable to using summated scales (Spector, 1992) with equal distances. Hence, an ordinal approach was deemed more intuitive to answer possibly complex questions and validate the evaluation method. It presents four possible grades (Not at all, A little, Rather, Very much) to eliminate neutrality and reduce ambiguity, especially for non-expert respondents, who are forced to provide a positive or negative answer.

To form the validation questionnaire, the items have been complemented with profiling (age, gender, region of provenance) and contextual questions intended to acquire information about the smart speaker owned or used by the respondents and their daily frequency of use to guarantee their understanding of the devices. Through an appointed agency, it has been submitted to a random sample of 671 anglophone people, familiar with AI-infused devices, from the UK and USA to avoid language biases with English questions. While for the previous study it was important to involve people who also had a sensitivity on design matters to inform the development of the scale, for its validation it was important to have random respondents to limit biases and ensure that the scale reflected the perspective of AI-infused artifacts users.

3.2.4 Statistical validation

The results have contributed to the first step towards validating the scale: an Exploratory Factor Analysis (EFA) (Fabrigar et al., 1999; Spearman, 1904), intended to identify the dimensions and the number of items for a manageable tool.

Subsequently, the items selected from the EFA results have been retained in a second version of the questionnaire. Through the same agency as before, it was sent to a new sample of 736 new respondents from the UK and USA, and the collected data were employed to perform a Confirmatory Factor Analysis (CFA). It aimed to test the structure model emerging from the EFA and, eventually, get the final number of relevant latent variables and items.

4 Presenting the results

4.1 State-of-the-art exploration

4.1.1 Systematic analysis of current UX methods – depicting common formats and approaches

The 129 UX evaluation methods were analysed and mapped (Spallazzo, Sciannamè, et al., 2021) to infer some generalizations about current approaches and inform the definition of a new one as output of the project. As synthesized in Figure 3, the most common format employed to collect evaluations is the questionnaire (69 methods), interpreted in various ways and followed by others mainly reflecting a scientific tradition of evaluation (e.g., interviews, recordings, physiological measurements, etc.). Solutions rooted in design and social sciences (i.e., diaries or cultural probes) are less frequent. Despite these intriguing alternatives may offer engaging and qualitatively rich ways for people to report on their experiences, they tend to be directed to limited samples. Questionnaires, instead, are usually more straightforward and have a more extensive range of possibilities to acquire significant amounts of responses for quantitative analysis and to be effectively introduced in the design process, especially by technology-driven companies producing AI-infused artefacts and UX researchers (primary users of AIXE). Therefore, this format has been adopted for building the scale.

Always with the target audience in mind, further information affecting the construction of AIXE were retrieved, like a clear preference for quantitative collection methods (57% of the analysed methods) or of digital devices as support materials. Additionally, most cases (125) are intended for individual non-expert users to evaluate the artifacts under investigation after performing some tasks or activities (99 cases) when they are already at an advanced design level (i.e., functional prototypes or products on the market).

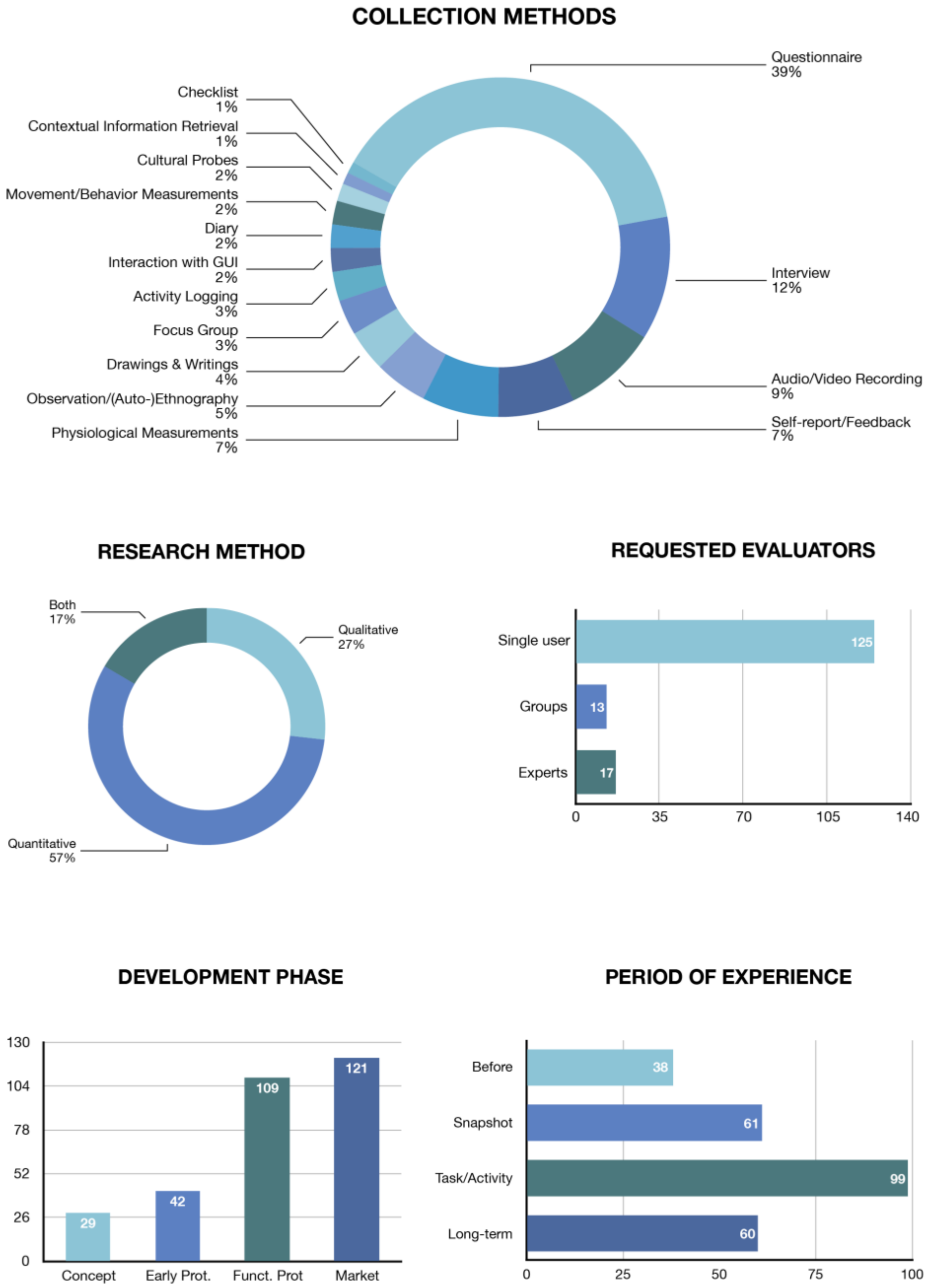


Figure 3: Synthetic overview of the UX evaluation methods analysis

4.1.2 Systematic analysis of current UX methods – dimensions and descriptors

The core ingredients of a UX evaluation method are the qualities of products and services that they aim to assess. A patent issue is the lack of agreement on terminology, and the same terms can be found indistinctly as dimensions and descriptors.

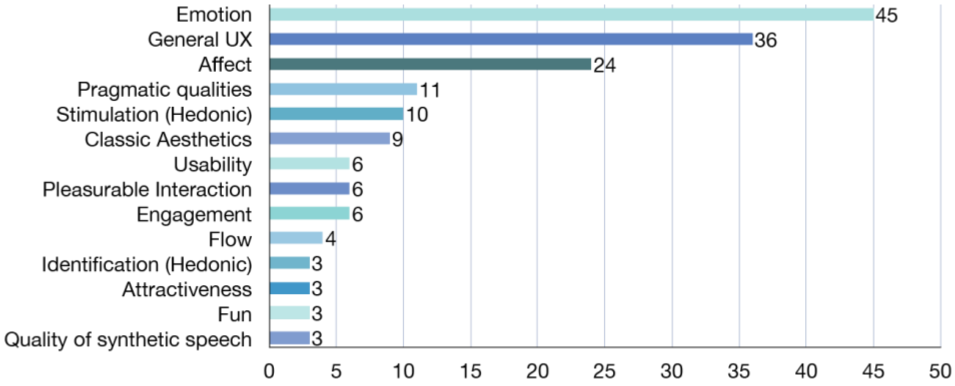


Figure 4: Prevailing dimensions in the mapping of UX evaluation methods

57 different UX evaluation dimensions emerged from the analysis of the 129 methods. The most recurrent ones are emotion and affect, respectively appearing in 45 and 24 cases (Figure 4). In consonance with different studies that, at the dawn of the new millennium, steered the design field from the dominance of functionality and usability toward softer considerations such as pleasure (Jordan, 2000), positive emotions (Norman, 2004), and aesthetics (Tractinsky et al., 2000) – just to cite a few.

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), complement the practical facets of an experience.

Considering the outcomes of the comprehensive work done, the researchers have retained these four dimensions (affective, pragmatic, hedonic, and aesthetic) as the most significant for building a holistic UX evaluation method that synthesizes 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 are starting to be developed and that they need to introduce more precise attributes.

To deduce information from the many descriptors collected, they have been first 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 (J. A. Russell, 1980) 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 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.

4.1.3 Systematic analysis of current UX methods – overall considerations

Further insights, useful for constructing AIXE, concern the number of dimensions each method considers and their perceived applicability to evaluate artefacts integrating AI systems.

For the former issue, most cases evaluate between one and two dimensions (1.7 average), denoting the generality and inevitable limitation of current assessment methods in dealing with complex and multi-faceted products with unique UX features. To meet this need, with a more comprehensive perspective on UX, only a few examples were found, namely SASSI - Subjective Assessment of Speech System Interfaces (Pettersson et al., 2018), 12 dimensions; SUISQ – Speech User Interface Service Quality (Polkosky & Lewis, 2003), 8; UEQ – User Experience Questionnaire (Polkosky, 2005), 6; and AttrakDiff (Laugwitz et al., 2008), 4.

Overall, the baseline hypothesis of the research was confirmed by the researchers' assessments. They had to evaluate the consistency of the analysed methods to address AI-infused products on a 1 to 5 scale, but none was rated with a 5, and most examples (45) obtained an average score of 3.

4.1.4 Literature review to understand AI-infused products qualities

An in-depth study to identify the peculiar and latent qualities of AI-infused artifacts was necessary as current UX evaluation methods proved their ineffectiveness in doing that. The main findings are synthesized here.

The concept of non-human intelligence was investigated and selected among the unique qualities of these systems. It brought out a foundational dichotomy between its machine and human-like nature. (S. Russell & Norvig, 2020) soundly describe its current prevailing interpretation as the rationally acting of AI systems, defined as rational agents that can operate autonomously (with no step-by-step programming), pursue goals to achieve the best-expected outcome, improve over time by learning from past experiences, perceive their environment and respond within it, and adapt to change. (Rijsdijk & Hultink, 2009) advance the definition of intelligence, including reactivity, multifunctionality, ability to cooperate, human-like interaction, and personality, while (Aarts & Ruyter, 2009) also includes customizability, context-awareness, and proactivity. These disruptive qualities inevitably entail a different UX based on unpredictable behaviours of AI systems, as confirmed in the fields of HCI and ambient intelligence by (Amershi et al., 2019), who proposed 18 design guidelines for human-AI interaction, and (Dove et al., 2017), who recognized them as no regular products that can be framed just within familiar features like usability, utility, and aesthetics.

The second essential UX dimension identified is trustworthiness, tightly related to the new agency AI-infused artefacts have and are at the centre of debates and studies, such as the several guidelines collected in (Algorithmic Watch, 2020). The main and most comprehensive one was published by the European Commission (High-Level Expert Group on Artificial Intelligence, 2019) and is declined as (i) human agency and oversight; (ii) technical robustness and safety; (iii) privacy and data governance; (iv) transparency; (v) diversity, non-discrimination and fairness; (vi) societal and environmental well-being; and (vii) accountability. Other key principles include aligning machine values with people's ones (S. Russell & Norvig, 2020), clarity on the intertwined relationship between humans and AI systems (Johnson & Verdicchio, 2017; van de Poel, 2020), and their functioning, capabilities, and limitations. Indeed explainability is crucial for ethicists (Kulesz, 2018), designers (Yang, 2020), and computer scientists who developed specific research strands like explainable AI – XAI (Confalonieri et al., 2021) and interpretable ML (Molnar, 2019).

The deepening of AI, design, and emotion-related studies highlighted the importance of understanding and making sense of AI-infused products and services as a measure for the UX they foster, as affect and cognition are related to an evaluation process (Norman, 2004). A lack of them may cause uncertainty, frustration, doubt, mistrust and inevitably influence people's experiences (Fruchter & Liccardi, 2018). Meaningfulness needs to be considered when dealing with these artefacts, both in terms of a cognitive issue (High-Level Expert Group on Artificial Intelligence, 2019) and as the satisfaction of psychological needs, such as autonomy, competence, relatedness, popularity, stimulation, security, through human-product interaction (Dourish, 2001b; Hassenzahl et al., 2013; Mekler & Hornbæk, 2019).

Ultimately, conversational interactions must be included. Even if they do not characterize all kinds of AI-infused artefacts, voice assistants are among the most widespread manifestation of this technology. As demonstrated, traditional UX evaluation methods are unsuitable to address this dimension, and new methods, reflections, and experimentations are emerging in the HCI field. In their overview, (Clark et al., 2019) point out interesting qualities to keep in consideration, such as user attitudes (towards the interface), task performance (total of dialogue turns, task completion, etc.), lexis and syntax choice, perceived usability, user recall (of specific aspects and outputs), and physiological qualities (like speech loudness and pitch). While heuristics (Maguire, 2019) and other experimentations (Bartneck et al., 2009; Garcia et al., 2018) highlighted more human-related features like anthropomorphism, animacy, likeability, and, above all, the agent's personality, but also accommodating conversational speech and ensuring high accuracy.

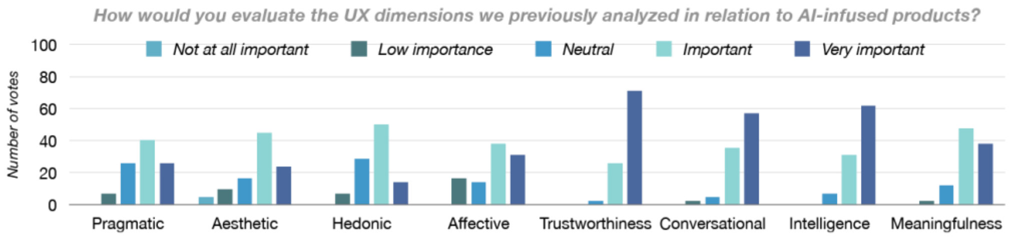


Figure 5. Survey results on the evaluation of the proposed UX dimensions for AI-infused products

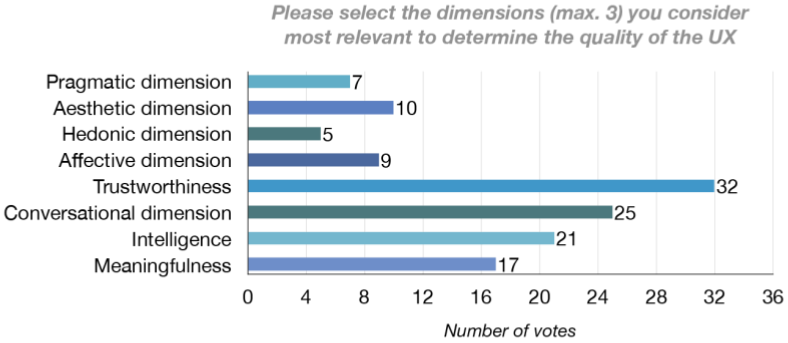


Figure 6. Survey results highlighting the most relevant UX dimensions for AI-infused products.

4.2 Scale construction

Overall, the direct assessment from the survey (Figures 5 - 6) reveals a positive perception of the proposed dimensions by the advanced users. In accordance with the research expectations, the dimensions more oriented towards AI-infused products (trustworthiness, intelligence, conversational, and meaningfulness) achieved better scores, while those more frequently included in current methods (pragmatic, aesthetic, hedonic, and affective) were considered less relevant in influencing their UX. Conversely, the request for suggestions on dimensions to complement or substitute the proposed ones provided no useful addition: a few replies confirmed the researchers' selections, while others provided qualities better suited to descriptors, or were off topic. Further interesting insights were gained with the collection of descriptors. These were evaluated by the researchers through an inter-coder agreement, and more qualitative information was derived by the contents and modalities in which people answered. Although single-word attributes were expected for each dimension, the format of the responses varied, and they were analysed based on their length, intelligibility, precision, appropriateness, and coherence. Key considerations are summarized in the following.

- Trustworthiness was the most successful dimension from every perspective. It produced a lot of descriptors, all highly rated in consistency and relevance, very much aligned with the European guidelines. Indeed, accuracy, data management, data protection, reliability, and transparency were the most significant features proposed. Also from a qualitative perspective, its perceived importance is denoted by articulated answers aiming at comprehensive explanations, and by the pervasiveness of related qualities throughout all the dimensions.
- Pragmatic dimension. The abundant and straightforward propositions of descriptors indicate a still prominent and consolidated role of this dimension in assessing UX. Both consistent and relevant from the researchers' evaluations, the majority of entries reports aspects already covered in current methods. However, some new qualities, directly influenced by the integration of AI, were proposed, like smartness, customization, responsiveness, adaptability, connectivity, and unobtrusiveness, along with other concepts related to trustworthiness.
- Conversational dimension. Peculiarly depicting new interactive possibilities unlocked by AI-infused systems, the significance of this dimension was relatively unquestioned. It was the most prolific in terms of received suggestions and presented precise and granular responses. Even if not entirely accurate from a technical standpoint, the related descriptors were mostly consistent and quite relevant, portraying a twofold picture. On the one hand, they

were strictly related to the qualities of speech (NLP quality, accent & dialect recognition, voice quality, character, etc.). On the other, more general features were highlighted (accuracy, context awareness, understanding, feedback quality, fluidity, and naturalness).

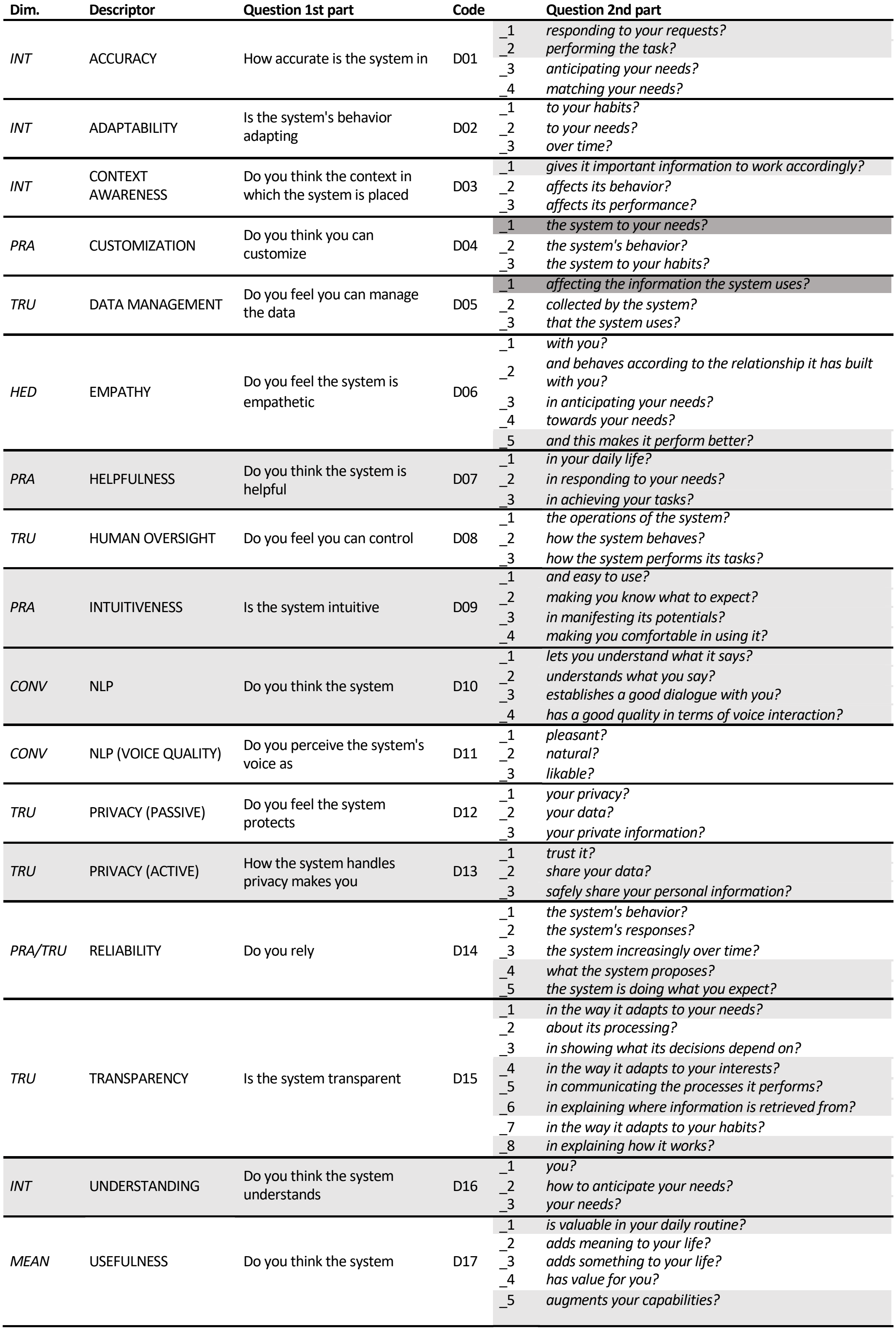
- Intelligence. The other dimension inevitably linked to AI-infused products also performed quite well in terms of consistency and relevance. Although some difficulties were expected in the definition of intelligence, a good variety of qualities were submitted. Accuracy, adaptability, context awareness, and understanding stood out among the given traits. The duality between characteristics related to the human sphere (learning, understanding needs, companionship) and the machine one (data elaboration, connectivity) emerged.
- Hedonic dimension. Probably difficult to understand, this dimension performed poorly both in consistency and relevance evaluations. In many cases, even if some AI-related quality emerged, the responses were incoherent or more appropriate for other dimensions. The most noteworthy descriptors were empathy and adaptability. Though, their connection with making a product attractive, engaging, and arising pleasant sensations during use is weak, as they both occur in 6 over 8 dimensions.
- Meaningfulness. This was undoubtedly the toughest for advanced users to contribute to. Even if the collected outputs were compliant with the requests, some stated they could not answer, and overall, the smallest number of entries was registered in this section. Difficulties in defining the boundaries of this dimension can be inferred from the identified qualities. Some examples are trustworthiness, multipurposeness, personality, empathy, and understanding. However, the best-rated ones reveal a connection with the human-computer/product relationship (e.g., usefulness, being beneficial, and helpfulness).
- Affective dimension. Long-winded, convoluted, and manifestly inconsistent responses (one-third of them were discarded even before the inter-coder agreement) characterized the definition of this dimension. The confusion and difficulty in expressing one’s emotions resulted in many answers pointing at the cause of emotions and not the affective responses themselves. However, some portray relevant qualities for AI-infused products, such as feeling in control and understood. Others are linked to the interaction with such devices (attraction, challenge, disappointment, frustration, and satisfaction). Yet, overall, the results show the inadequacy of affection to characterize and communicate the UX AI-infused products.
- Aesthetic dimension. The data collected demonstrated poor quality, which, in the end, was echoed by poor performance. The outputs revealed a superficial approach from the advanced users, many of whom indicated specific characteristics of the products on the market (e.g., white colour, small size, rounded shapes, etc.). and aesthetics got the lowest consistency and relevance averages from the researchers’ evaluations. Diverging from common conceptions of aesthetics, though, personality and mimesis (reinforced by invisibility and unobtrusiveness as dear concepts to the field of ubiquitous computing) stand out as significant descriptors.

Ultimately, the descriptors obtaining the highest possible score from all the judges (“golden” descriptors) were retained – with the related dimensions – and are portrayed in Table 1.

Table 1. List of the 36 "golden" descriptors with the related dimensions

Source	Golden Descriptors	Source	Golden Descriptors
Conversational dimension (literature)	Voice naturaleness Voice pleasantness	Pragmatic dimension (literature)	Functionality Helpfulness Intelligibility Intuitivity Learnability Reliability Understandability
Conversational dimension (questionnaire)	Accuracy Context awareness NLP quality Reliability Understandability	Pragmatic dimension (questionnaire)	Customization Ease of use Transparency Trustworthiness
Hedonic dimension (questionnaire)	Empathy	Trustworthiness (literature)	Access to data Human oversight Non-discrimination Privacy Quality of data Transparency Unfair bias avoidance
Meaningfulness (questionnaire)	Usefulness		
Intelligence (questionnaire)	Accuracy Empathy Context awareness Understanding	Trustworthiness (questionnaire)	Accuracy Data management Data protection Reliability Transparency

Starting from this list the researchers excluded those that were difficult to measure or too general, identifying the most promising ones for the scale to build. The final selection included human-related qualities (empathy, understanding, and usefulness), characteristics of the system itself (helpfulness, intuitiveness, reliability, accuracy, adaptability, and context awareness), and features merging both in a sociotechnical ensemble (customization, human oversight, data management, privacy, transparency, and reliability – as an ethical concern). Then, each descriptor was declined into different questions (items) to compose the first version of the AIXE questionnaire to be validated. As shown in Figure 7, it included 65 items with no logical order to the sequence of questions, not to influence the results.



Dim.	Descriptor	Question 1st part	Code		Question 2nd part
INT	ACCURACY	How accurate is the system in	D01	_1	*responding to your requests?*
				_2	*performing the task?*
				_3	*anticipating your needs?*
				_4	*matching your needs?*
INT	ADAPTABILITY	Is the system's behavior adapting	D02	_1	*to your habits?*
				_2	*to your needs?*
				_3	*over time?*
INT	CONTEXT AWARENESS	Do you think the context in which the system is placed	D03	_1	*gives it important information to work accordingly?*
				_2	*affects its behavior?*
				_3	*affects its performance?*
PRA	CUSTOMIZATION	Do you think you can customize	D04	_1	*the system to your needs?*
				_2	*the system's behavior?*
				_3	*the system to your habits?*
TRU	DATA MANAGEMENT	Do you feel you can manage the data	D05	_1	*affecting the information the system uses?*
				_2	*collected by the system?*
				_3	*that the system uses?*
HED	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?*
PRA	HELPFULNESS	Do you think the system is helpful	D07	_1	*in your daily life?*
				_2	*in responding to your needs?*
				_3	*in achieving your tasks?*
TRU	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?*
PRA	INTUITIVENESS	Is the system intuitive	D09	_1	*and easy to use?*
				_2	*making you know what to expect?*
				_3	*in manifesting its potentials?*
				_4	*making you comfortable in using it?*
CONV	NLP	Do you think the system	D10	_1	*lets you understand what it says?*
				_2	*understands what you say?*
				_3	*establishes a good dialogue with you?*
				_4	*has a good quality in terms of voice interaction?*
CONV	NLP (VOICE QUALITY)	Do you perceive the system's voice as	D11	_1	*pleasant?*
				_2	*natural?*
				_3	*likable?*
TRU	PRIVACY (PASSIVE)	Do you feel the system protects	D12	_1	*your privacy?*
				_2	*your data?*
				_3	*your private information?*
TRU	PRIVACY (ACTIVE)	How the system handles privacy makes you	D13	_1	*trust it?*
				_2	*share your data?*
				_3	*safely share your personal information?*
PRA/TRU	RELIABILITY	Do you rely	D14	_1	*the system's behavior?*
				_2	*the system's responses?*
				_3	*the system increasingly over time?*
				_4	*what the system proposes?*
				_5	*the system is doing what you expect?*
TRU	TRANSPARENCY	Is the system transparent	D15	_1	*in the way it adapts to your needs?*
				_2	*about its processing?*
				_3	*in showing what its decisions depend on?*
				_4	*in the way it adapts to your interests?*
				_5	*in communicating the processes it performs?*
				_6	*in explaining where information is retrieved from?*
				_7	*in the way it adapts to your habits?*
				_8	*in explaining how it works?*
INT	UNDERSTANDING	Do you think the system understands	D16	_1	*you?*
				_2	*how to anticipate your needs?*
				_3	*your needs?*
MEAN	USEFULNESS	Do you think the system	D17	_1	*is valuable in your daily routine?*
				_2	*adds meaning to your life?*
				_3	*adds something to your life?*
				_4	*has value for you?*
				_5	*augments your capabilities?*

Figure 7. AIXE questionnaire at different stages. In white the questions of the final scale, in light gray those excluded after the EFA, in dark gray those excluded after the CFA.

4.3 Statistical validation

4.3.1 Exploratory Factor Analysis

Overall, the EFA had positive results as the goodness of fit indexes denote a good model. 601 answers to the first version of the questionnaire were admissible for the EFA (70 were excluded because missing information or declaring that smart speakers are never used). A set of 36 items out of the 65 submitted and the related 13 descriptors (out of the initial 17) were selected based on higher performances (factor loadings > 0.5), and limited to the best three for the descriptors presenting more possible values. An exception was made for the *empathy* descriptor (the only one representing the *hedonic* dimension), for which four items were kept because of the high factor loadings of all the related items and the qualitative significance and diversity the researchers attributed to them. Additionally, although all the items related to the two *privacy* descriptors (D12 and D13) reported high factor loadings, the questions had similar meanings, and differentiation might confuse the respondents. For this, only those belonging to D12 were kept, as all the items reported higher factor loadings than the others, indicative of a more straightforward question structure.

4.3.2 Confirmatory Factor Analysis

A reduced questionnaire, including only the 36 acceptable items resulting from the EFA, was then submitted to a new sample of 736 people for a total of 705 admissible answers for the CFA. This brought to discard three more items: D04_1, D05_1, D16_4 (dark grey in Figure 7), and the final calculation of goodness of fit indexes confirmed the model's validity, with no further necessary reduction.

Therefore, the ultimate version of the scale (white background in Figure 7) counts six dimensions and twelve descriptors as latent variables expressing the UX of AI-infused products. The Cronbach coefficients are summarized in Table 2.

Table 2. Cronbach coefficients for each dimension

Latent variable	N. of Items	Cronbach coefficients
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 Discussion and future work

The paper summarizes all the research processes and findings developed within the Meet-AI project that led to the construction of AIXE, a scale to evaluate AI-enabled experiences. The initial hypothesis about the inappropriateness of current methods to frame the complexity and uniqueness of wide-spreading AI-infused artifacts was confirmed in the first phase of the research through a thorough analysis of 129 current UX evaluation methods and an exploratory literature review aiming to identify the peculiarities of the objects under investigation. This opened an interesting gap for research and represented an opportunity for UX design to join the conversation and bring a different perspective. Hence, the basic nature of the inquiry was suitable with an iterative approach to identify the core latent qualities that characterize AI-infused products and services. From the state-of-the-art exploration, the first phase of the research, eight possible dimensions emerged: pragmatic, aesthetic, hedonic, affective, intelligence, trustworthiness, conversational, and meaningfulness. To start building the scale, the second phase consisted of different subsequent research activities, including a survey to bring external perspectives into the reasoning and further steps of analysis, engaging the research team to a better refinement of the previous findings and point out consistent foundations for an AI-related evaluation method. Finally, six dimensions were retained, declined into seventeen descriptors, and a total of sixty-five items. They composed the first draft of an ordinal questionnaire submitted to two random samples of anglophone people for validation. The data from the first were used for an Exploratory Factor Analysis to identify a viable structure for the evaluation method, while the Confirmatory Factor Analysis performed on those from the submission of the second reduced version statistically established the solidity of the model.

Despite the limitations of the most exploratory stages of the inquiry, such as the number and background of the advanced users engaged with the survey and the subjectivity of methods, decisions, and evaluations conducted by the researchers to identify the latent and manifest variables to build the AIXE scale, the validation process observed a statistical rigor based on well-established methods, which testify the reliability of the results.

Meant to be employed by designers, UX researchers, companies and start-ups for long-term users' evaluation (over a period of at least one or two weeks) of the AI-infused artefacts they develop, the actual application of the AIXE scale will further enrich the results, enabling the researchers to spot possible spaces for improvement or confirming the validity of the evaluation method.

Further future developments include a digital version of the scale to make it publicly available and usable and the translation of the identified core qualities for evaluating the UX into meta-design principles and tools to inform the early stages of designing AI-infused artifacts. Indeed, this could impact educational and professional contexts allowing creative processes that could beneficially result in more reliable, pleasant, trustworthy, intelligent, and meaningful objects and interfaces integrating AI systems and capabilities.

References

Aarts, E., & Ruyter, B. (2009). New research perspectives on Ambient Intelligence. *JAISE*, *1*, 5–14. https://doi.org/10.3233/AIS-2009-0001

Algorithmic Watch. (2020, April). *AI Ethics Guidelines Global Inventory*. https://inventory.algorithmwatch.org/

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Antonelli, P. (2018, February 8). *AI Is Design's Latest Material* (PAIR, Interviewer) [Interview]. https://design.google/library/ai-designs-latest-material/

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–2698. https://doi.org/10.1145/1978942.1979336

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., & Cowan, B. (2019). The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, *31*(4), 349–371. https://doi.org/10.1093/iwc/iwz016

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining and Knowledge Discovery*, *11*(1), e1391. https://doi.org/10.1002/widm.1391

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE.

Dourish, P. (2001a). *Where the Action Is: The Foundations of Embodied Interaction* (New Ed edition). The MIT Press.

Dourish, P. (2001b). *Where the Action Is. The Foundations of Embodied Interaction*. MIT Press.

Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 278–288. https://doi.org/10.1145/3025453.3025739

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fruchter, N., & Liccardi, I. (2018). Consumer Attitudes Towards Privacy and Security in Home Assistants. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3170427.3188448

Garcia, M. P., Lopez, S. S., & Donis, H. (2018). Voice activated virtual assistants personality perceptions and desires: Comparing personality evaluation frameworks. *Proceedings of the 32nd International BCS Human Computer Interaction Conference*, 1–10. https://doi.org/10.14236/ewic/HCI2018.40

Giaccardi, E., & Redström, J. (2020). Technology and More-Than-Human Design. *Design Issues*, *36*(4), 33–44. https://doi.org/10.1162/desi_a_00612

Hassenzahl, M., Borchers, J., Boll, S., Pütten, A. R. der, & Wulf, V. (2020). Otherware: How to best interact with autonomous systems. *Interactions*, *28*(1), 54–57. https://doi.org/10.1145/3436942

Hassenzahl, M., Eckoldt, K., Diefenbach, S., Laschke, M., Len, E., & Kim, J. (2013). Designing Moments of Meaning and Pleasure. Experience Design and Happiness. *International Journal of Dsign*, *7*(3), 21–31.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

Holmquist, L. E. (2017). Intelligence on tap: Artificial intelligence as a new design material. *Interactions*, *24*(4), 28–33. https://doi.org/10.1145/3085571

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI Discourse. *Minds and Machines*. https://doi.org/10.1007/s11023-017-9417-6

Jordan, P. W. (2000). *Designing Pleasurable Products: An Introduction to the New Human Factors*. CRC Press. https://doi.org/10.1201/9780203305683

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*(4), 443–477. https://doi.org/10.1007/BF02293808

Kaptelinin, V., & Nardi, B. A. (2009). *Acting with Technology. Activity Theory and Interaction Design.* MIT Press.

Kocaballi, A. B., Laranjo, L., & Coiera, E. (2018, July 4). *Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires*. https://doi.org/10.14236/ewic/HCI2018.21

Kulesz, O. (2018). *Culture, platforms and machines: The impact of Artificial Intelligence on the diversity of cultural expressions* (DCE/18/12.IGC/INF.4) [Information Document]. UNESCO. https://en.unesco.org/creativity/sites/creativity/files/12igc_inf4_en.pdf

Lachner, F., Naegelein, P., Kowalski, R., Spann, M., & Butz, A. (2016). Quantified UX: Towards a Common Organizational Understanding of User Experience. *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, 1–10. https://doi.org/10.1145/2971485.2971501

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger (Ed.), *HCI and Usability for Education and Work* (pp. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Maguire, M. (2019). Development of a Heuristic Evaluation Tool for Voice User Interfaces. In A. Marcus & W. Wang (Eds.), *Design, User Experience, and Usability. Practice and Case Studies* (pp. 212–225). Springer International Publishing. https://doi.org/10.1007/978-3-030-23535-2_16

Mekler, E. D., & Hornbæk, K. (2019). A Framework for the Experience of Meaning in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300455

Molnar, C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/

Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things*. Basic Books.

Pettersson, I., Lachner, F., Frison, A.-K., Riener, A., & Butz, A. (2018). A Bermuda Triangle? A Review of Method Application and Triangulation in User Experience Evaluation. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3173574.3174035

Polkosky, M. D. (2005). *Toward a Social-Cognitive Psychology of Speech Technology: Affective Responses to Speech-Based e-Service* [Ph.D.].

Polkosky, M. D., & Lewis, J. R. (2003). Expanding the MOS: Development and Psychometric Evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology*, *6*(2), 161–182. https://doi.org/10.1023/A:1022390615396

Rijsdijk, S. A., & Hultink, E. J. (2009). How Today's Consumers Perceive Tomorrow's Smart Products *. *Journal of Product Innovation Management*, *26*(1), 24–42. https://doi.org/10.1111/j.1540-5885.2009.00332.x

Rivero, L., & Conte, T. (2017). A Systematic Mapping Study on Research Contributions on UX Evaluation Technologies. *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3160504.3160512

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. *Proceedings of the 2018 Designing Interactive Systems Conference*, 857–868. https://doi.org/10.1145/3196709.3196772

Spallazzo, D., Ajovalasit, M., Ceconello, M., Sciannamè, M., & Vitali, I. (2021). *Assessment of Descriptors for UX Evaluation of AI-infused Products* (p. 124653 Bytes) [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.14387468.V1

Spallazzo, D., & Sciannamè, M. (2021). *UX Descriptors for AI-infused Products* (p. 22224 Bytes) [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.14345498.V1

Spallazzo, D., Sciannamè, M., Ajovalasit, M., Ceconello, M., Vitali, I., & Arquilla, V. (2021). *UX Evaluation Methods Mapping* (p. 100905 Bytes) [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.14350553

Spallazzo, D., Sciannamé, M., & Ceconello, M. (2020). Towards a UX Assessment Method for AI-Enabled Domestic Devices. In N. Streitz & S. Konomi (Eds.), *Distributed, Ambient and Pervasive Interactions* (pp. 336–347). Springer International Publishing.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Sage Publications.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145. https://doi.org/10.1016/S0953-5438(00)00031-X

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Vermeeren, A. P. O. S., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: Current state and development needs. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 521–530. https://doi.org/10.1145/1868914.1868973

Yang, Q. (2020). *Profiling Artificial Intelligence as a Material for User Experience Design* [Carnegie Mellon University]. http://reports-archive.adm.cs.cmu.edu/anon/hcii/abstracts/20-100.html

Zheng, Q., Tang, Y., Liu, Y., Liu, W., & Huang, Y. (2022). UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–24. https://doi.org/10.1145/3491102.3501855

About the Authors:

Martina Sciannamè: Ph.D. in Design and research fellow at the Design Department of PoliMi. She focuses on human-centered and responsible design for meaningful technological solutions. She is developing theoretical knowledge and practical tools to include ML and related ethics in design education.

Davide Spallazzo: Ph.D. in Design and Associate Professor at the Design Department of PoliMi, he’s the coordinator of Meet-AI. Active in Interaction Design and HCI, he investigates human-centred approaches to digital innovation and meaning-making applied to fields like cultural heritage, serious gaming, and AI.