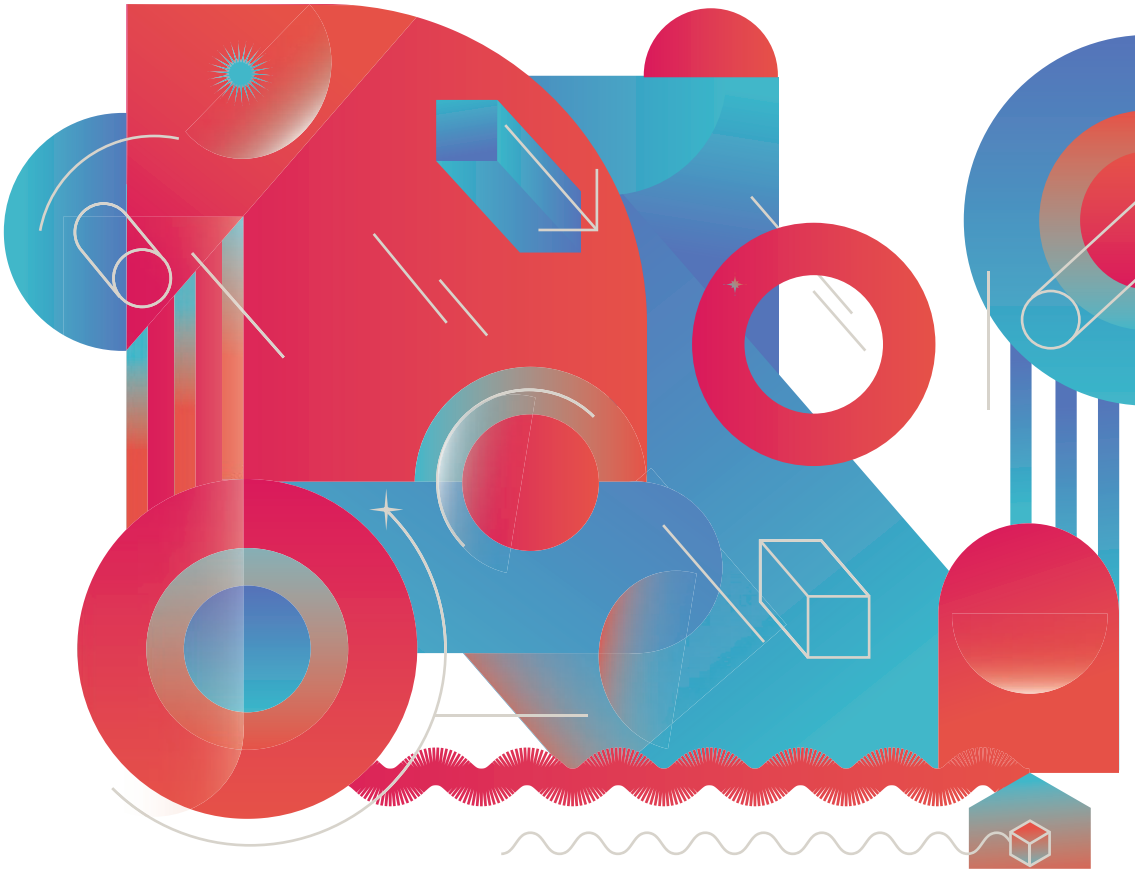
# EMBEDDING INTELLIGENCE

## Designerly reflections on AI-infused products

edited by Davide Spallazzo, Martina Sciannamè



D.I. FrancoAngeli
DESIGN INTERNATIONAL

The Design International series is born in 2017 as a cultural place for the sharing of ideas and experiences coming from the different fields of design research, becoming a place in which to discovering the wealth and variety of design, where different hypotheses and different answers have been presented, drawing up a fresh map of research in international design, with a specific focus on Italian design.
Different areas have been investigated through the books edited in these years, and other will be explored in the new proposals.
The Scientific Board, composed by experts in fashion, interior, graphic, communication, product and industrial, service and social innovation design, interaction and emotional design, guarantee the high level of the accepted books. After the first selection by the Scientific Board, the proposals are submitted to a double review by other international experts.

 ISBN 9788835141914




Copyright © 2022 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788835141914

# EMBEDDING INTELLIGENCE

Designerly reflections on AI-infused products

edited by Davide Spallazzo, Martina Sciannamè

D.I. FRANCOANGELI DESIGN INTERNATIONAL OPEN ACCESS

 ISBN 9788835141914

Cover image by Sara Sciannamè

ISBN e-book Open Access: 9788835141914

Date of first publication: September 2022



Copyright © 2022 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788835141914

# Contents

# 3. The qualities of AI-infused products. Reflections on emerging UX dimensions

*Martina Sciannamè*
*Politecnico di Milano, Department of Design*

*Emma Zavarrone*
*Iulm University, Humanities Department*

The flourishing moment of AI that yields the first waves of materialization of this technology offers a fertile ground for UX and interaction designers, who might bring added values to the currently technology-driven AI-infused products.

However, these artefacts present inherent complexity and distinctive features that significantly affect the user experience (UX) and should be considered for their development and evaluation. The dimensions commonly considered in current UX assessments, though, result insufficient and inconsistent for this task. For this reason, Meet-AI, a research project funded by the Design Department of Politecnico di Milano, focuses on identifying the most fitting qualities to describe AI-infused products and ultimately aims to create a specific UX evaluation method.

Building on the premises described in chapter 2, this contribution portrays the process that led to the delineation of seventeen qualities at the basis of such method. Carried out within the Meet-AI project, the preliminary research is divided into two phases. The first investigates more and less traditional UX dimensions coming respectively from a wide-range critical analysis of existing UX evaluation methods and a literature review on AI and humans. The second aims at assessing the feasibility of the assumptions from the first phase and explores new qualities through a survey submitted to advanced users, which responses have been analyzed in subsequent steps culminating in a workshop within the research group. This eventually produced the synthesis from which starting to build the UX evaluation method, final objective of the Meet-AI project.

## 3.1 The UX of AI-infused products: a challenge to take on

The evolution and democratization of personal computers and the Internet demonstrates that seemingly niche technologies can successfully spread among the lay public if less technical and more *designerly* factors are considered in their development. Hence, the current materialization of AI in everyday products and services is a great opportunity for UX and interaction design.

Between the hype for novelty and the disillusion caused by unfulfilled promises and incomprehension, AI-infused products offer several possibilities for experimentations that appeal to basic design principles like discoverability, proper exploitation of functions (Hekkert, 2006; Kinsella, 2018; White, 2018), and personal significance, to finally let this technology bring richness and enjoyment to people (Norman, 2004). The relevance of the limitations brought by a technology-driven concretization of AI is recognized and countered both in academia (Dove *et al.*, 2017; Yang, 2020) and by the biggest companies providing such products and services, who are themselves defining and publishing guidelines to support the design of AI-infused systems (Amershi *et al.*, 2019; Google PAIR, 2019).

However, as portrayed in chapter 2, the peculiar nature of these products requires additional, ad hoc preparation of interaction and UX designers, who primarily need to comprehend the key features they have to work with.

Framing the UX of this new generation of products, then, is a starting point for a deeper understanding of the limitations and the potentialities they entail. Moreover, in order to conceive, develop, and improve products and services integrating AI, some guidance is needed, and these are the main premises to the Meet-AI project: a one-year-long research project, funded by the Department of Design, Politecnico di Milano. Its objective is to build a new evaluation method that specifically addresses the UX of AI-infused products, comprehending and highlighting the peculiarities they have in relation to other interactive products of common use. Based on the principal hypothesis that current UX assessment methods cannot frame and analyze the UX enabled by such systems, appropriate UX dimensions must be detected.

The chapter discusses the project's preliminary research and findings, obtained with a multi-method approach framed in two main phases that share the common goal of identifying the most relevant qualities to

describe the UX of AI-infused products, to build the premises for a UX evaluation method. The first phase includes extensive research on existing UX evaluation methods and a literature review attempting to define AI as a UX matter. The second aims to verify the assumptions from the previous inquiries and further expand the investigation, including advanced users through a survey analyzed according to a multi-step protocol.

## 3.2 Phase one: setting the ground to understand and assess the UX of AI-infused products

### *3.2.1 Research methods*

As a first step, the research required acknowledging the state of the art in UX evaluation to get a comprehensive picture of the most considered qualities for describing interactions between people and products of various kinds and to understand which ones might characterize the unique relationship with AI-infused products. Hence, this preliminary phase of the investigation was twofold: it comprehended a wide-range critical analysis of qualitative and quantitative UX evaluation methods, and a literature review on the intersection of AI, interaction design, and HCI to explore possibly uncovered angles.

Firstly, the five researchers involved in the study independently identified and examined relevant scales and methods to assess the UX, both within the field of design and in related social sciences experimentations. The research was limited to articles published in the ACM Digital Library and Springer Link between 2000 and 2020, resulting from the entry of "UX evaluation" and "UX assessment" keywords. To spot and integrate potentially missing methods, the All About UX website – the largest repository of UX evaluation methods available at the time of the study – has also been used as a reference. In the end, a list of 129 UX evaluation methods emerged, and they have been analyzed according to various criteria (Spallazzo, Sciannamè, *et al.*, 2021). Central for the inquiry was to highlight the *UX dimensions* and *descriptors* addressed in each case, here respectively intended as the general qualities that significantly describe people's experience of products and the specific features explaining the nuances of such overarching qualities. Additionally, to understand how the existing evalua-

tion methods are operationalized and can inspire the construction of a new one (the primary expected outcome of the Meet-AI project) other relevant pieces of information have been pinpointed, namely: the *collection method(s)* (tools and modalities used to retrieve UX evaluations); whether more than one method has been put in place (*triangulation*); the context (*lab/field*), and the *support materials* used; the nature of the investigation (*qualitative/quantitative/both*); 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 kind of *object(s) of study*; the *evaluators* required (single user, groups, expert users), and the researchers' perceived *level of consistency* with AI-infused products. To complement this, also sources and personal notes were added.

Because one of the premises of the Meet-AI project is the alleged absence of UX evaluation methods able to capture the essence of artifacts integrating AI systems, a deepening on the theme was also necessary. To this end, the researchers collected insights and reflections from a literature review revolving around the relationship between people and AI. The facets of human-AI interaction have been initially investigated according to three main thematic strands: non-human intelligence, emotion, and meaning. These, later on, evolved to include other relevant aspects in the current debate on the topic, namely, conversational interaction and ethical implications.

At the time of the study, UX and interaction design interest in AI was in its infancy, therefore, not many references could be found in the disciplinary literature. This is why the research expanded into related fields, such as HCI, computer ethics and AI itself.

### 3.2.2 What can be gleaned from current UX evaluation methods?

The critical analysis of existing UX evaluation methods eventually resulted in an extensive table (Spallazzo, Sciannamè, *et al*., 2021), from which some inferences can be derived.

The most easily quantifiable considerations concern the framing of the assessment methods (Fig. 3.1). As expected, the most frequently

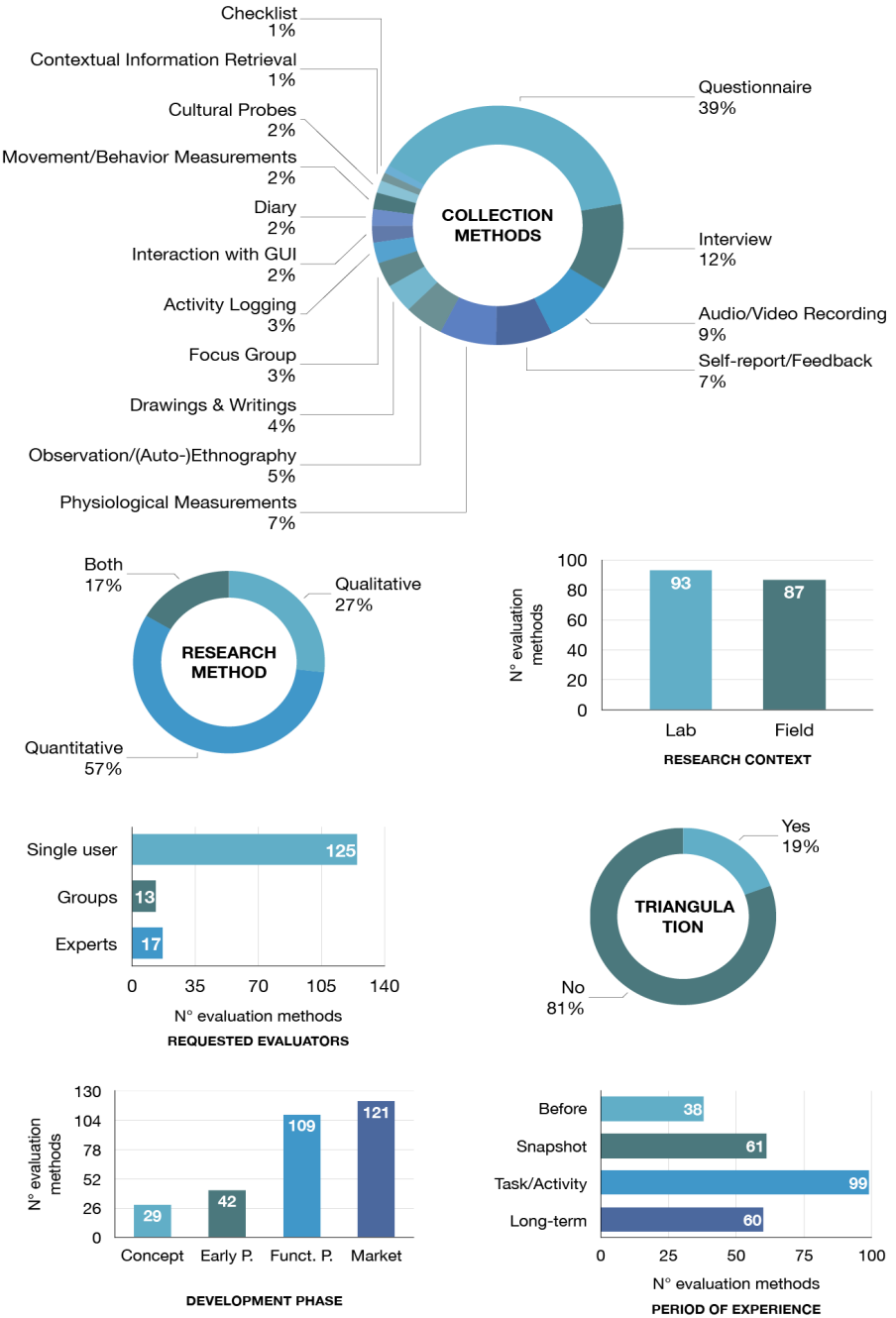



*Fig. 3.1 – Synthetic overview of the UX evaluation methods analysis.*

employed format for gathering evaluations is the questionnaire (69 methods), which has been interpreted both in traditional (questions and scales) and in more creative ways (with graphic, pictorial, auditory versions, and even randomly appearing when the phone screen is unlocked). Other prominent collection methods are interviews (21), video/audio recording (16), physiological measurements, and self-reports/feedbacks (13). These mostly reflect scientific approaches to evaluation, while modalities rooted in the design and social sciences fields, like diaries (4) and cultural probes (3), are less frequent.

As one can notice, the sum of collection methods exceeds the total of UX evaluation methods analyzed. 19% of them triangulate information retrieved in different ways to add soundness to more experimental or qualitative studies. The latter still represents a minority (27.6%) compared to quantitative practices (57%), although 16.4% opted for mixed methods.

Of course, digital devices (computers and mobile phones) stand out among support materials as they can easily process data coming from questionnaires, sensors, activity logs, and video/audio recordings, sometimes using custom software or apps.

The analyzed UX evaluation methods are equally submitted in a lab (93) and/or in actual contexts of use (87 occurrences) to test a wide variety of products and services. The majority is versatile and can encompass as many industrial products, as systems, environments, and events. Few specifically address reduced niches of interactive content such as visual interfaces or video games.

Additionally, this kind of investigation mostly require single non-expert users to evaluate the study objects when they are at an advanced design level – 109 methods can be applied to functional prototypes, and 121 when the product is already on the market (the same evaluation method can be submitted at different stages of development) – and after exploiting a precise task or activity (99). Debatably, the initial phases of the design process do not seem to be given much consideration for an early evaluation that may lead to a quick and rapid iteration of the product.

However, the wide-range analysis focuses on the qualities defining the UX dimensions and descriptors to evaluate products and services. Before getting into the issue, it is essential to state that literature reveals no agreement on terminology, and sometimes the exact words are used

without careful attention to the semantic nuances or interchangeably as dimensions and descriptors.

If chapter 2 shows the recent dominance of generic UX in UX evaluation methods, our content analysis (Fig. 3.2) traces a slightly different story.

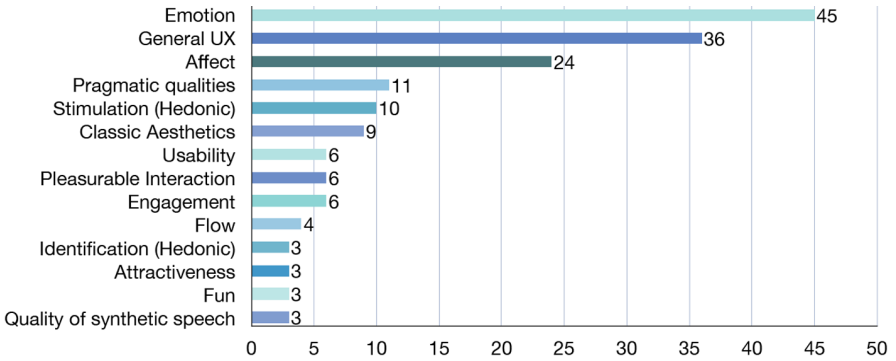



*Fig. 3.2 – Prevailing dimensions in the mapping of UX evaluation methods.*

The first relevant discrepancy concerning the previous studies is the prominent role that *emotion* and *affect* have come to play as UX evaluation dimensions: respectively, they appeared in 45 and 24 cases. *General UX*, instead, has been demoted to second place (with 36 occurrences). This is probably the result of the consistent number of experiments in the UX evaluation from the field of psychology.

Subsequently, *pragmatic* (17 occurrences if summed to usability), *hedonic* (10 stimulation + 3 identification) and *aesthetic qualities* (9) emerge, confirming the balance between practical and subjective sides of the overarching experience. More could be added to the list, that counts a total of 57 declared dimensions, like: *pleasurable interaction* (6), *engagement* (6), *flow* (4), *attractiveness* (3), *fun* (3), and so on. However, they mostly refer to soft characteristics that could be easily included in the previous, overarching, ones. Noteworthy is instead the *quality of synthetic speech* (3) – a new entry in the UX panorama – which is evidence that some methods are emerging to particularly address interfaces implying novel kinds of interaction modalities and they needed to introduce more specific attributes to assess the AI-infused products integrating them.

In the end, from the comprehensive work, the researchers highlighted the four most frequently assessed dimensions as the most significant for a holistic method, synthesizing the legacy of UX evaluation. They are the *pragmatic, hedonic, aesthetic,* and *affective* dimensions. As it refers to an overall perception of the user experience and lacks a clear measurability, *generic UX* has not been considered a valuable option for this study.

Afterwards, the descriptors collected from the UX evaluation methods have been organized according to these overarching dimensions to have the most recurrent ones depicting a nuanced picture of their dominant interpretations. It follows that the pragmatic dimension concerns the use of an artifact not just in terms of *helpfulness, efficiency,* and *functionality*, but also from more user-friendly characteristics such as *easiness, simplicity, clearness, navigation, learnability, reliability, and convenience*. Aesthetics presents multiple facets, probably due to its variability according to the object of the evaluation and subjective nature, but it can mainly be interpreted as *appearance* (where *clarity* and *sophistication* are two recurring themes), or *attractiveness* (*good* and *pleasant* qualities being frequently assessed). The hedonic dimension often overlaps with the affective one, as their borders are very subtle. *Enjoyability* and *excitement* are undoubtedly two factors that determine the pleasure of use, but also *creativity*, *inventiveness,* and *innovativity* appear to have their weight. Finally, the affective component is largely influenced by psychology. In fact, *valence* and *arousal* are by far the most common descriptors in this category, followed by the recognized basic emotions: *pleasure, fear, sadness, happiness, disgust, anger*, and *surprise*.

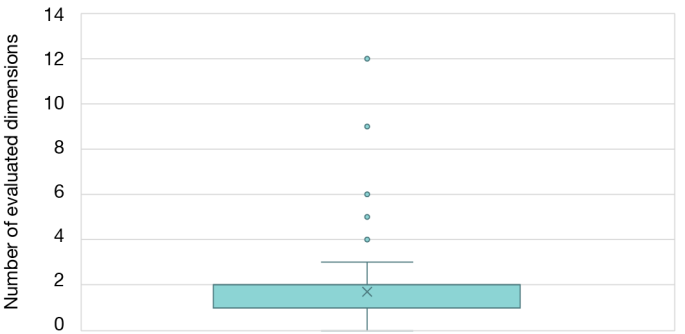



*Fig. 3.3 – Variability of the UX dimensions assessed by each evaluation method.*

The number of UX dimensions covered by each method represents another relevant information. The majority of the analyzed UX evaluation methods, in fact, moves within the boxplot (with a mean of 1.7 dimensions per method), underlining their limitation in dealing with complex, multi-faceted products, the interaction they imply is still to be fully comprehended. Just few outliers (showing high variability) demonstrated a more holistic perspective towards UX by exploring several dimensions, but they seemed more suitable references for the task at hand. Among the most distant from the mean are two methods notably referring to conversational interfaces: SASSI – Subjective Assessment of Speech System Interfaces (Pettersson *et al.*, 2018), 12 dimensions; SUISQ – Speech User Interface Service Quality (Polkosky and Lewis, 2003), 8; the UEQ (Polkosky, 2005), 6; and AttrakDiff (Laugwitz, Held and Schrepp, 2008), 4.

Overall, Meet-AI researchers' evaluations of the consistency of the analyzed UX methods with the subject matter confirmed the baseline hypothesis: no current UX evaluation method is sufficiently comprehensive and not too broad to adequately address AI-infused products. The majority (45) received an average score of 3 (on a 1 to 5 scale), 31 reached a 4, but none has been considered totally suitable to manage AI-infused products assessment. However, two methods have been flagged as particularly interesting for the purpose of the Meet-AI project. The first is the Affective Feedback Loop (Bruns Alonso *et al.*, 2013) which, although not strictly an assessment tool, suggests to design interactive objects in a way that they can sense users' behaviors and receive affective feedbacks as part of the human-computer interaction. This could be particularly valuable for products integrating machine learning (ML) systems – the predominant subset of AI – as they can learn and adapt over time. Translated into an evaluation method, this approach could be implemented in a functional prototype or a market phase of a product to gather rich and unmediated information and to respond with prompt iterations. While the second, Perceptive Sorting (Forlizzi, Gemperle and DiSalvo, 2003), is outside the timespan of the overview but still a relevant reference. It concerns the evaluation of unfamiliar products for target users – as in the case of AI-infused ones – and distinguishes three levels of assessments to infer design directions for the development of new artifacts: using normative qualities to describe the usability of displayed familiar objects, lifestyle indications

for known objects with many unfamiliar models, and affective words for unfamiliar objects which function was explained by the researchers. In the latter, a significant scenario for products using natural language interactions emerged: personality traits and gender have been often attributed to the presented robots, as to define the unknown with more human qualities.

To conclude with the results of this analysis, despites the appealing alternative methods for assessment collection, the questionnaire format is not only the most frequently used, but, probably the most straightforward to elaborate and introduce in the design process of new products integrating AI. Thus, this will be the goal of the Meet-AI project.

### 3.2.3 A deeper dive into AI-infused products: what literature tells us?

As the investigation on existing UX evaluation methods confirmed the initial assumption of the Meet-AI project, a wide-ranging literature review on AI-infused products and AI-related issues seemed essential to identify possible latent but significant UX dimensions before starting the development of a specific evaluation method. Here, we will outline the main findings.

As anticipated, the inquiry started from three main strands: non-human intelligence, to clarify what artificial intelligence is and what are its dominant features; emotions, as they play a prominent role in the UX evaluation, they may help describe human-machine relationship; and meaning, to understand how people make sense or can interpret AI mediated interactions.

The basis of this argumentation retraces AI history to comprehend how this discipline has been conceived and evolved over time. Its very name already suggests a dichotomy in its nature, combining machines and a distinctly human trait: intelligence. In their world-wide recognized textbook on AI Russell and Norvig (2020) try to give an explanation of this concept as foundational feature of this subfield of computer science. They distinguish four connotations that have been attributed to artificial intelligence: intelligence as accurate simulation of *human* performance, as *rationality*, as an internal *thought* process or as an

externalized *behavior*. The combination of such definitions yields four possible representations of AI systems. They can be considered intelligent if they (i) *act humanly* – i.e., they successfully communicate in a human language (natural language processing: NLP), store information they know or hear (knowledge representation), answer questions to draw new conclusions (automated reasoning), or adapt to new circumstances and identify patterns (machine learning: ML); if they (ii) *think humanly* – manifesting human-like thought processes as observed in psychology, cognitive and neural sciences; if they (iii) *think rationally* – applying irrefutable reasoning processes derived by logic or probability; or if they (iv) *act rationally* – meaning that AI systems, and predominantly ML systems, are rational agents that can operate autonomously (with no step-by-step program), perceive their environment and respond within it, adapt to change, improve over time by learning from past experiences, and pursue goals to achieve the best expected outcome. The latter is the prevailing approach in the field; hence, the listed qualities represent a major reference for the study.

Similar features of AI systems emerge from HCI, and ambient intelligence and they are regarded as the disruptive elements that force a change in the way we look at the UX they entail. Products with autonomy, adaptability, reactivity, multifunctionality, ability to cooperate, human-like interaction, personality (Rijsdijk and Hultink, 2009), and that can also be personalized on habits or preferences, context-aware and anticipatory (Aarts and Ruyter, 2009) are no regular products, and their UX cannot be defined by simple usability, utility and interaction aesthetics (Dove *et al.*, 2017). Which is why their *intelligence* (implying all the above) needs to be considered for a specific evaluation method, as it is corroborated by the work of Amershi *et al.* (2019) who have proposed 18 design guidelines for human-AI interaction, justifying them based on the unpredictability of behaviors that AI-infused products have.

This last aspect triggers several unprecedented implications, subject of a lively debate, especially among computer ethicists. The researchers investigated an additional domain and identified a second, essential UX dimension: *trustworthiness*.

As a matter of fact, for an effective user experience of autonomous and continuously evolving systems, certain issues are of utmost importance to gain users' trust. For instance, the values and objectives put

into the machines should be aligned with people's ones (Russell and Norvig, 2020) to guarantee a beneficial impact; humans' role in the development and in the interaction should be well communicated as they are part of the AI systems themselves (Johnson and Verdicchio, 2017); and, above all, people should be able to understand why these systems make their decisions, how they function, what are their capabilities and limitations. Explainability is indeed a crucial aspect from different perspectives: for ethicists (Kulesz, 2018), designers (Yang, 2020) and also computer scientists, in whose discipline the specific branches of explainable AI – XAI (Confalonieri *et al.*, 2021), and interpretable ML (Molnar, 2019) emerged.

A great number of directions for developing beneficial AI systems emerged in a short span of time. For reference Algorithmic Watch (2020) represents the most up to date repository, while Hagendorff (2020) discusses them in his paper. However, the most comprehensive ones are those published by the European Commission (High-Level Expert Group on Artificial Intelligence, 2019). They include seven main guidelines for trustworthy AI: (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.

Looking at ethical issues also from an affective standpoint, they can be relevant for UX assessment as the presence or lack of concerns may affect people's responses, for example a lack of explainability can provoke uncertainty, frustration, doubt, mistrust (Fruchter and Liccardi, 2018).

While the thread of emotion studies dominates the scenario of UX evaluation methods as they fill the gap between people and products (Forlizzi and Battarbee, 2004) influencing their attitudes, behaviors, perceptions and assessments (Scherer, 2005), the researchers also delved into an intertwined matter: *meaning.* As Norman (2004) stated, both affect and cognition are related to an evaluation process: the first to determine the positive or negative impact that things surrounding us may have; the second to make sense of the world. Indeed, attributing a meaning to a product and understanding what it represents for the user is a relevant measure for the UX, even if outlining its defining traits can be tricky. Within the HCI, ethics and design communities, this has been variously debated, both as a cognitive issue (High-Level Expert

Group on Artificial Intelligence, 2019) and as a quality of the human-product interaction that satisfies psychological needs (Dourish, 2001; Hassenzahl *et al*., 2013; Mekler and Hornbæk, 2019) like autonomy, competence, relatedness, popularity, stimulation, security (Hassenzahl, 2011). Not to perpetrate the path of devices showing off their technological novelties but confined to the gadget or toy dimension (Levinson, 1977), the design process should be steered by meaningfulness concepts. AI-infused products, as all human artifacts, might be developed with a predefined purpose, resonate as a personal significance, a shared and/or cultural significance, generate valuable experiences, communicate a symbol or exhibit a temporal quality, thus referring to some kind of meaning at a functional, ritual and/or mythical level (Ajovalasit and Giacomin, 2019).

The research finally closed the circle on possibly relevant UX dimensions for products integrating AI systems by investigating a very peculiar interaction modality that these activate as a result of the first conceptions of intelligence embodiment: *conversational* capabilities. Even though AI is not synonymous with conversational interface, voice assistants are among the most widespread manifestation of this technology, and they fail to be evaluated by traditional UX methods. In fact, as already brought out, specific methods have risen, and additionally, reflections and experiments emerged in HCI literature.

A good overview of the state of user interaction with speech systems has been given by (Clark *et al*., 2019) who highlighted that commonly measured concepts include: 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). Moreover, heuristics (Maguire, 2019) and other experimentations (Bartneck *et al*., 2009; Garcia, Lopez and Donis, 2018) outlined further essential aspects to keep in consideration when dealing with human-like interactions, such as accommodating conversational speech, ensuring high accuracy to minimize input errors and adequate system feedback (also in recovering from errors), but also more human features like anthropomorphism, animacy, likeability, and, above all, the personality of the agent.

## 3.3 Phase 2: Actively exploring the qualities of AI-infused products

### *3.3.1 Overview of the research methods*

Not to limit the investigation to the subjective perspective of the researchers, they opted for introducing a novel element to the comparative review. In addition to the inferences from the literature review, a co-creation of the descriptors to compose the scale has been envisioned as a suitable way to engineer their selection and preserve as much objectivity as possible.

Hence, a protocol combining mixed methods and different steps was established. *Step 0:* the driving force behind the second phase of research is a survey intended to validate the assumptions concerning the dimensions to describe AI-infused products and to solicit a creative contribution on descriptors to expand the non-comprehensive set extracted from the literature review. While the survey provided immediate results on the dimensions (*step 1*), the analysis of the suggested descriptors was more articulated and composes *step 2*. A preliminary homologation of the responses proposing descriptors was independently conducted by two of the researchers and finally confronted to compile a shared list. The descriptors in the list were further analyzed through an affinity map to synthesize repetitions and filter out out-of-context responses. The consequent set of descriptors was then submitted to the researchers of the Meet-AI project for an inter-coder evaluation aimed at assessing the consistency of the descriptors with the related dimension and their relevance for AI-infused products, and finally at extracting the most significant ones. The last step (*step 3*) before the construction of the evaluation scale was an internal workshop within the Meet-AI team to define the elements (dimensions and descriptors) that would form its structure in light of the results of the whole research work.

Because of the articulated configuration of this second phase of the research, further details on the methods will be provided in the next paragraphs, in combination with the results of each step.

### 3.3.2 Step 0: a survey to expand the boundaries of AI-related qualities

To cross-examine the findings from the initial investigation (namely the eight identified dimensions: *pragmatic, aesthetic, hedonic, affective, intelligence, trustworthiness, conversational, meaningfulness*) and to further enrich the spectrum of attributes that might describe the target products, a group of advanced users has been involved through a digital survey. The selected population was composed of 110 students from MSc in Digital and Interaction Design and 47 young researchers from the Design Department of Politecnico di Milano, who are familiar with the type of products being studied and have a developed sensitivity and comprehension of the design of interactive objects. From the total, 42 responded, with a response rate of 26.75%.

The survey was meant to be clear and transparent on its purpose, therefore, this was openly stated in the introduction, and the AI-based smart speakers, learning thermostats and smart cams were presented both to give references of the artifacts to be addressed and to understand the respondents' level of experience with the exemplified objects. Instead, the core inquisitive part was twofold, focusing first on (i) seeking a new set of descriptors and then on (ii) acquiring feedbacks on the UX dimensions aimed at describing products integrating AI systems. (i) To avoid possible misunderstandings, the dimensions according to which the advanced users were asked to suggest new UX qualities were portrayed with a definition before getting to research request. As synthesized in Fig. 3.4, in some cases the respondents had to indicate at least three attributes, in others, a minimum of two positive and two negative features, to encourage heterogeneous answers, (ii) After the explorative contribution, the proposed dimensions have become the subject of critique for the researchers to understand how well they perform in the evaluation of AI-infused products, which are considered the most relevant, and whether may there be missing ones. To gather such information, direct questions have been posed. A profiling section closed the questionnaire.

| 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. |
| ***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. 3.4 – Synthesis of the descriptors requests as they appeared in the survey.*

### 3.3.3 Step 1: UX dimensions of AI-infused products according to advanced users

Moving backwards in the analysis of the survey responses, this paragraph outlines the assumptions related to the proposed dimensions, while more grained and qualitative considerations on the descriptors that the respondents attributed to each of them will be depicted in the following one.

As graphically portrayed in Figure 3.5, the overarching qualities listed in the survey received quite positive ratings. As might be expected, *trustworthiness*, *intelligence, conversational,* and *meaningfulness* – the dimensions stemming from the AI-focused literature review – found a strong consensus among advanced users (underlining their consistency with the target products), while the most frequently used in current methods – *pragmatic, aesthetic, hedonic* and *affective* dimensions – were mainly assessed as "important".

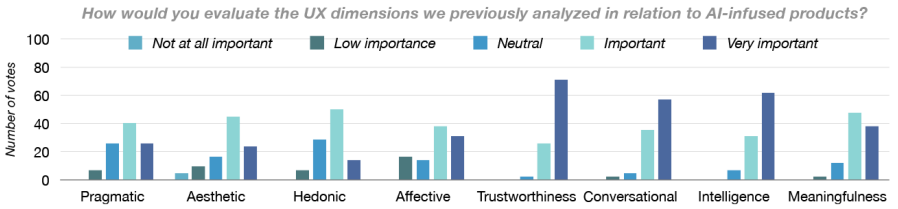



*Fig. 3.5 – Survey results on the evaluation of the proposed UX dimensions for AI-infused products.*

The next question, double-checking the relevance of the dimensions with respect to the UX of AI-infused products (Fig. 3.6), further confirmed these results. With the favor of 76% of the respondents, *trustworthiness* prevailed, and it was followed by *conversational* (59.5%), *intelligence* (50%), and *meaningfulness* (40.5%).

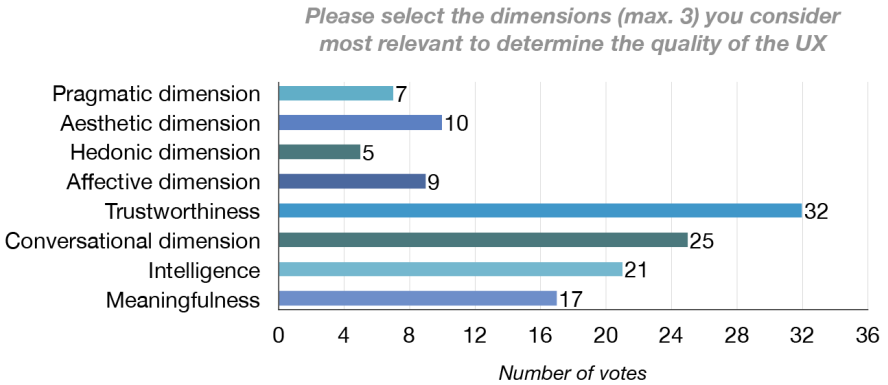



*Fig. 3.6 – Survey results highlighting the most relevant UX dimensions for AI-infused products.*

In contrast, the direct solicitation for additional dimensions essential to the project goal was not met with useful replies. In fact, the few comments received confirmed the proposed selection, contained features better identifiable as descriptors, or were off topic.

Other parameters may be looked at to find a confirmation of the results above and the researchers' initial hypothesis. The substance and

the way in which people answered to the request for descriptors for each dimension is indeed significant.

For instance, the personal contribution, appropriateness, and coherence of the responses reveal the advanced users' perceptions towards the different dimensions. In some cases, they were straightforward (e.g., *pragmatic* dimension), in others, a proper assessment could be more difficult. *Hedonic*, *affective* and *meaningfulness* dimensions, in fact, received mostly incoherent and long-winded responses, showing respectively high subjectivity, shortcomings and both these issues merged, with some respondents openly stating their inability to answer at all. The inconsistency was also proved by valuable attributes provided within these categories, but that were more appropriate in other contexts.

The prevailing richness of the responses – in terms of amount (for *conversational* and *intelligence* dimensions), and articulation (*trustworthiness*) – their suitability with the overarching factors and the subjects of evaluation, and the pervasiveness of the related contents throughout the questionnaire made the appreciation of these qualities even more explicit.

*Aesthetics*, though, was the one dimension with poor-quality data. A low perceived relevance in relation to the research goal transpired (as confirmed in the explicit evaluation), with responses bearing some level of superficiality by addressing specific characteristics of the products on the market.

### 3.3.4 Step 2: insights from an intertwined analysis of AI-related descriptors

To properly derive all facets from the suggested descriptors and infer useful information for the construction of a new evaluation method for AI-infused products, the raw responses needed some preparation and further assessment by the Meet-AI research group.

Preliminarily, to make the survey responses consistent with the initial request, two of the researchers redacted a homogeneous list, translating sentences and Italian answers in single English words. The resulting lists of one-word descriptors were then confronted to compile a uniform one (Spallazzo and Sciannamè, 2021).

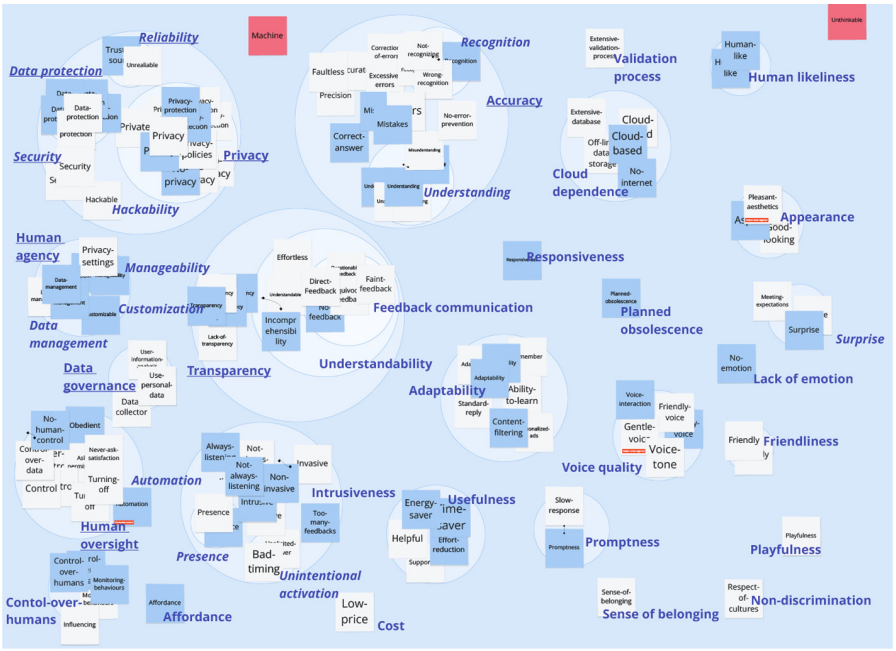



*Fig. 3.7 – Affinity map of the descriptors from the Trustworthiness dimension.*

Subsequently, all the entries were collected in a Miro board (Fig. 3.7), differentiated according to the study dimensions. For each, an affinity map was built to visualize semantic concentrations, extrapolate descriptors in a univocal way – blue words outside post-its (also referring to the language encountered in literature – underlined), and flag (in red) any incomprehensible or patently out-of-context concept (e.g., *lights* or *function* as descriptors of the *affective* dimension). Blue post-it notes highlight the terms modified by the two researchers to meet the one-word (English) format.

Ultimately, a synthetic list of unambiguous descriptors was then prepared and shared among the research group for crossed-evaluation, seeking intercoder agreement (Creswell, 2014) for each feature based on two parameters: (i) the consistency with the dimension, and (ii) the relevance for AI-infused products. Each descriptor was presented within the dimension for which it was originally proposed by the respondents, along with the frequency of its occurrences. To complete the picture, also descriptors from the literature review (L)

were added in different sheets. The researchers, in the role of judges, had to assess all of them on a 1 (not consistent/relevant at all) to 4 (highly consistent/relevant) scale. The results are public at (Spallazzo, Ajovalasit, *et al*., 2021), and they have been analyzed by calculating the mean and z score for each descriptor and according to each parameter (consistency and relevance). Then, they have been organized into quartiles to easily spot the most significant. Finally, a comprehensive overview (Fig. 3.8) was obtained by comparing the relevance z scores of the descriptors altogether. Also dividing these into quartiles, the segment >75% counts 134 descriptors. Hence, special consideration has been given to the 36 receiving the maximum score from all the researchers (the "golden" ones), and they have been operationalized in the last step (described in the next paragraph).

After this processing, an overview of which is offered in Fig. 3.9, some informed inferences could be derived. For the sake of synthesis, they are here discussed in relation to their overarching dimensions.

*Pragmatic dimension*. It was probably the easiest for the respondents, who answered consistently (one of the highest overall consistency scores from the researchers' evaluation), with one-word descriptors as requested, and collected a total of 46 descriptors with 134 submitted items (number of words originally suggested by the respondents and then synthesized based on their semantic affinity) – only two of which did not reach the list to be assessed. It did not present major innovations compared to the literature, yet some new aspects directly related to AI-infused objects emerged, e.g.: *smartness, customization, responsiveness, adaptability, connectivity, unobtrusiveness,* and different concepts linked to *trustworthiness*. In terms of relevance, it marked the second-best score in both the mean of evaluations and the overall "golden" descriptors, probably underlining that this basic dimension for evaluating UX is still essential or, at least, most of the relevant qualities of AI-infused products have been attributed to this dimension by the respondents.

*Aesthetic dimension*. Respondents' answers in this category presented a high influence of currently adopted practices in the industry of AI-infused products. Often, they were too specific in indicating characteristics of products on the market (e.g., white color, small size, rounded shapes, etc.) and a great work of generalization was necessary in order to submit the descriptors to the judges. In the end, these

| SOURCE | DESCRIPTOR | R1 EV | R2 EV | R3 EV | R4 EV | R5 EV | R6 EV |
|---|---|---|---|---|---|---|---|
| CONV-L | Voice naturaleness | 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 |

| SOURCE | DESCRIPTOR | R1 EV | R2 EV | R3 EV | R4 EV | R5 EV | R6 EV |
|---|---|---|---|---|---|---|---|
| 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 |

*Fig. 3.8 – List of the "golden" descriptors with the related dimensions.*

| 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*** | 155 (S) | 49 (S) | 2 | 2.24 | 2.39 | / | 1 |

(S) from survey | (L) from literature

*Fig. 3.9 – Descriptors performances, synthesized according to the related dimension, in the various steps of the analysis*

received the lowest consistency, the lowest relevance mean of evaluation (even with a negative connotation: score of 1.76 out of 4), and they were the least represented among the overall most relevant with no one resulting as a "golden" descriptor. What stands out (with a relevance mean of 3.8), though, is a quality that diverge from the most traditional conception of aesthetic, merging with studies in the affective perception of products: *personality.* It is followed by the concept of *mimesis* (dear to the field of ubiquitous computing) which records a mean of 3 but occurs also in the nuances of *invisibility* and *unobtrusiveness.*

*Hedonic dimension.* Despites being at the second place for number of descriptors (55) from the questionnaire, its performances were among the lowest, with slightly sufficient thresholds both in the consistency and relevance evaluations. Here, qualities directly linked to AI-infused products emerged (e.g., *multifunctionality, responsiveness, voice interaction*) but most of them were considered not particularly significant or just more appropriate for other dimensions. This is also manifest in the overall ranking of relevant descriptors, where the hedonic dimension gives just a little contribution. Particularly noteworthy is instead the value of *empathy* which receives here the highest accreditation from the judges, immediately followed by *adaptability* (even if both occur in 6 over 8 dimensions).

*Affective dimension.* In stark contrast to the analyzed UX evaluation methods (counting 96 different descriptors and 219 occurrences from literature), the analysis depicted a lot of confusion among the respondents. The difficulty of correctly expressing one's emotions clearly emerges. Lots of articulated sentences (even if just single words were requested) appeared and most of the answers were aiming at the cause of emotions in the interaction and not at the affective responses themselves. For this reason, it recorded the highest, impressive number of exclusions for manifest inconsistency even before the judges' evaluation (around 1/3 of the descriptors coming from the responses were discarded). In the end, the affective descriptors from the questionnaire did not perform badly in the judges' opinions (also marking the best score for consistency), and actually relevant qualities for AI-infused products can be highlighted, like the empowering *feeling in control* and *feeling understood*, right before those emerging from direct interaction with such devices: *attraction, challenge, disappointment, frustration,* and *satisfaction*. However, the affective dimension was only second to the aesthetic one in terms of the least number of descriptors in the >75%, with no "golden" items as well, and the traditional qualities (coming from literature) proved not to be valuable when considering the UX of AI-infused products, with a negative mean of 1.64 (out of 4).

*Trustworthiness.* It marks its commonly agreed relevance in qualitative and quantitative ways. Firstly, even in this dimension, answers were quite articulated, but mostly they didn't highlight a misunderstanding or a difficulty in answering, but rather a desire for a better explanation. Secondly, qualities referring to this category emerged throughout all other dimensions, underlining their pervasive importance. As expected,

this was also manifest in numbers: the amount of reported descriptors was rich, the judges' evaluations of trustworthiness descriptors were among the highest and this dimension is the one that contributes in the largest part to the top >75% overall relevance ranking with exactly 1/3 of the "golden" descriptors pertaining to it. The most successful ones quite echo the European guidelines and concern *accuracy, data management, data protection, reliability,* and *transparency.*

*Conversational dimension.* With 60 descriptors and 160 submitted items, it was the most prolific in absolute terms. It presented a good set of precise responses (maybe not from a literary point of view but with enough granularity), marking its perceived significance in relation to AI-infused products. The judges' evaluations also reflected this position in both consistency and relevance average values, as well as in the overall relevance ranking, where a lot of conversational descriptors appear in the >75% and "golden" ones count (only following *trustworthiness* and *pragmatic* dimensions). As stressed by the specificity of some of its descriptors (like *NLP quality, accent & dialect recognition, voice quality, character*, etc.), the conversational dimension is mostly relatable to a part of AI-infused products. Tough, others can also be generalized to a more comprehensive behavior. It is the case of *accuracy, context awareness, understanding, feedback quality*, but also *fluidity* and *naturalness.*

*Intelligence.* Even though it is undeniably difficult to define intelligence, highlighting the qualities that characterize a perceived intelligent behavior in AI-infused products proved to be a less heavy task. The responses in this dimension were satisfying: of the 141 submitted items, none was discarded in the first round of analysis (preceding the judges' evaluation). They also performed quite well, placing themselves in an average position in terms of evaluated consistency and relevance of the proposed descriptors, as well as in the overall relevance ranking. Here, again, characteristics like *accuracy*, *adaptability*, *context awareness,* and *understanding* stood out, in a mixed context that presents some traits reminding human capabilities (e.g., *learning, understanding needs, companionship*), and others strictly linked to the machine dimension (e.g. *data elaboration, connectivity*).

*Meaningfulness.* In conclusion, this was undoubtedly the toughest dimension to depict, and it is not by chance that this had the smallest number of items proposed (115). Nonetheless, respondents tried to answer according to the request – without long-winded digressions – but

some of them openly expressed their inability to answer at all. Probably the difficulty to determine what belongs to this domain did not help to encounter some preferred qualities uniquely belonging to it. In fact, attributes with fuzzy boundaries recurred, like *trustworthiness, multi-purposeness, personality, empathy,* and *understanding.* Yet, those evaluated as most interesting (*usefulness*, *being beneficial*, and *helpfulness*) mostly appeal to the human-computer/product relationship.

### 3.3.5 Step 3: a summarizing workshop

Once all the preliminary research activities (from the literature review and the UX evaluation methods mapping to the survey submission and analysis) resulted in a synthetic portrait of the most relevant descriptors for the assessment of AI-infused products, a workshop within the research group seemed the most suitable way to collectively discuss the reached outcome and to pave the way to the construction of a specific evaluation method.

As anticipated, the so-called "golden" descriptors were extrapolated to understand their possible role within the UX assessment of products integrating AI systems. They were displayed on post-it notes on a Miro board with their related dimension. Then, the researchers categorized them according to their perceived likelihood of being part of the scale to be built.

Some (*data protection, quality of data, unfair bias avoidance, trustworthiness,* and *non-discrimination*) were labelled as "not usable" because of the lack of information and difficult measurability for a proper assessment. Instead, repeated descriptors pertaining to multiple and less coherent dimensions were considered "better to keep out", while weaker and too general attributes were left in the "could be in" category to leave space to the "must be in" ones. The ultimately selected descriptors are depicted in Fig. 3.10. While *empathy, understanding*, and *usefulness* refer to human-related qualities; *helpfulness, intuitiveness, reliability, accuracy, adaptability,* and *context awareness* are attributes properly belonging to the system itself. Lastly, *customization, human oversight, data management, privacy, transparency,* and *reliability* (as ethical concern) configure the product as a sociotechnical ensemble, merging human needs and system properties.

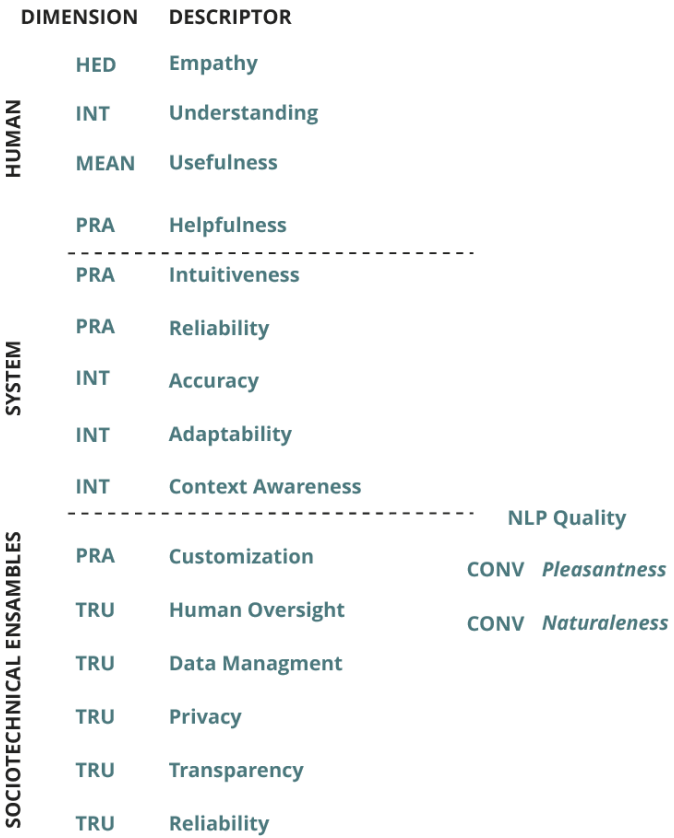



*Fig. 3.10 – Ultimately selected descriptors to build a UX evaluation method for AI-infused products.*

## 3.4 Conclusions, limitations, and future directions

The chapter describes all the phases and steps of the research conducted within the Meet-AI project that anticipate the construction of a UX evaluation method for AI-infused products. Moving from the initial assumption that current methods cannot frame the complexity and peculiarities of this novel products representing an opportunity for UX design, a first phase of the research resulted in eight possibly suitable dimensions to describe their UX: *pragmatic, aesthetic, hedonic, affective, intelligence, trustworthiness, conversational, meaningfulness.* After a second phase, starting from a survey to include perspectives

external to the research team and comprehending subsequent steps of analysis, a final list of seventeen relevant descriptors (Fig. 3.10) was extracted as a basis on which to build the evaluation scale.

Indeed, the research presents limitations in terms of the number and context of people involved, as well as the subjectivity of methods, decisions and evaluations conducted by the researchers. However, future developments should balance the qualitative work here presented.

Specifically, the next steps should include a solid elaboration of a scale that will need to be tested by a large number of smart speakers' users (as they are the most widespread concrete products integrating the technology under study) to gain statistically valuable information for a definitive validation of a UX evaluation method for AI-infused products. After achieving quantitatively robust results, the method should be generalizable and disseminated to support the design and consequent assessment of devices or services integrating AI systems.

## References


Aarts, E. and Ruyter, B. (2009). "New research perspectives on Ambient Intelligence". *JAISE*, 1, pp. 5-14. doi:10.3233/AIS-2009-0001.

Ajovalasit, M. and Giacomin, J. (2019). "Meaning of artefacts". *Conference Proceedings of the Academy for Design Innovation Management*, 2(1). doi:10.33114/adim.2019.02.266.

Algorithmic Watch (2020). *AI Ethics Guidelines Global Inventory.* Available at: https://inventory.algorithmwatch.org/ (Accessed: 27 July 2021).

Amershi, S. *et al.* (2019). "Guidelines for Human-AI Interaction". In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery (CHI '19), pp. 1-13. doi:10.1145/3290605.3300233.

Bartneck, C. *et al.* (2009). "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots". *International Journal of Social Robotics*, 1(1), pp. 71-81. doi:10.1007/s12369-008-0001-3.

Bruns Alonso, M. *et al.* (2013). *Measuring and adapting behavior during product interaction to influence affect*. Berlin: Springer. Available at: https://doi.org/10.1007/s00779-011-0472-3 (Accessed: 28 March 2020).

Clark, L. *et al.* (2019). "The State of Speech in HCI: Trends, Themes and Challenges". *Interacting with Computers*, 31(4), pp. 349-371. doi:10.1093/iwc/iwz016.

Confalonieri, R. *et al.* (2021). "A historical perspective of explainable Artificial Intelligence". *WIREs Data Mining and Knowledge Discovery*, 11(1), p. e1391. doi:10.1002/widm.1391.

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

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

Dove, G. *et al.* (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*. New York, NY: ACM (CHI '17), pp. 278-288. doi:10.1145/3025453.3025739.

Forlizzi, J. and Battarbee, K. (2004). "Understanding Experience in Interactive Systems". In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques. DIS '04 Designing Interactive Systems 2004*, Cambridge, MA: ACM Press.

Forlizzi, J., Gemperle, F. and DiSalvo, C.F. (2003). "Perceptive sorting: a method for understanding responses to products". In *DPPI '03*. doi:10.1145/782896.782922.

Fruchter, N. and 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*. Montreal, QC: Association for Computing Machinery (CHI EA '18), pp. 1-6. doi:10.1145/3170427.3188448.

Garcia, M.P., Lopez, S.S. and Donis, H. (2018). "Voice activated virtual assistants personality perceptions and desires: comparing personality evaluation frameworks". In *Proceedings of the 32nd International BCS Human Computer Interaction Conference*. Belfast: BCS Learning & Development Ltd. (HCI '18), pp. 1-10. doi:10.14236/ewic/HCI2018.40.

Google PAIR (2019). *People + AI Guidebook*. Available at: https://design.google/ai-guidebook (Accessed: 5 July 2021).

Hagendorff, T. (2020). "The Ethics of AI Ethics – An Evaluation of Guidelines". *Minds and Machines* [Preprint]. doi:10.1007/s11023-020-09517-8.

Hassenzahl, M. (2011). "User Experience and Experience Design". In Soegaard, M. and Dam, F.R. (eds.). *The Encyclopedia of Human-Computer Interaction*. 2nd edition. The Interaction Design Foundation.

Hassenzahl, M. *et al.* (2013). "Designing Moments of Meaning and Pleasure. Experience Design and Happiness". *International Journal of Design*, 7(3), pp. 21-31.

Hekkert, P. (2006). "Design Aesthetics: principles of pleasure in design". *Psychology Science*, 48, pp. 157-172.

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

Johnson, D.G. and Verdicchio, M. (2017). "Reframing AI Discourse". *Minds and Machines* [Preprint]. doi:10.1007/s11023-017-9417-6.

Kinsella, B. (2018). *61% of Alexa Skills Still Have No Ratings and Only 1% Have More Than 100 – Voicebot*. Available at: https://voicebot.ai/2018/10/05/61-of-alexa-skills-still-have-no-ratings-and-only-1-have-more-than-100/ (Accessed: 26 February 2019).

Kulesz, O. (2018). *Culture, platforms and machines: The impact of Artificial Intelligence on the diversity of cultural expressions. Information Document*. Paris: UNESCO. Available at: https://en.unesco.org/creativity/sites/creativity/files/12igc_inf4_en.pdf.

Laugwitz, B., Held, T. and Schrepp, M. (2008). "Construction and Evaluation of a User Experience Questionnaire". In Holzinger, A. (ed.). *HCI and Usability for Education and Work*. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 63-76. doi:10.1007/978-3-540-89350-9_6.

Levinson, P. (1977). "Toy, Mirror, and Art: The Metamorphosis of Technological Culture". *Learning Cyberspace*, 34(2), pp. 151-167.

Maguire, M. (2019). "Development of a Heuristic Evaluation Tool for Voice User Interfaces". In Marcus, A. and Wang, W. (eds.). *Design, User Experience, and Usability. Practice and Case Studies*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 212-225. doi:10.1007/978-3-030-23535-2_16.

Mekler, E.D. and 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*. Glasgow: Association for Computing Machinery (CHI '19), pp. 1-15. doi:10.1145/3290605.3300455.

Molnar, C. (2019). *Interpretable Machine Learning*. Available at: https://christophm.github.io/interpretable-ml-book/ (Accessed: 31 May 2021).

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

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

Polkosky, M.D. and 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), pp. 161-182. doi:10.1023/A:1022390615396.

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

Russell, S. and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. 4th edn. Hoboken, NJ: Pearson.

Scherer, K.R. (2005). "What are emotions? And how can they be measured?". *Social Science Information*, 44(4), pp. 695-729. doi:10.1177/0539018405058216.

Spallazzo, D., Ajovalasit, M. *et al*. (2021). "Assessment of Descriptors for UX Evaluation of AI-infused Products". *figshare*. doi:10.6084/M9.FIGSHARE.14387468.V1.

Spallazzo, D., Sciannamè, M., *et al*. (2021). "UX Evaluation Methods Mapping". *figshare*. doi:10.6084/M9.FIGSHARE.14350553.

Spallazzo, D. and Sciannamè, M. (2021). "UX Descriptors for AI-infused Products". *figshare*. doi:10.6084/M9.FIGSHARE.14345498.V1.

White, R.W. (2018). "Skill Discovery in Virtual Assistants". *Communications of the ACM*, 61(11), pp. 106-113.

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