UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCES AND ENGINEERING

UXAmI Framework: User Experience Evaluation in Ambient Intelligence Environments

by

Stavroula Ntoa

PhD Dissertation
Presented in Partial Fulfilment of the Requirements
for the Degree of Doctor of Philosophy

Heraklion, November 2017

UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE
UXAml Framework: User Experience Evaluation in Ambient Intelligence Environments

by **Stavroula Ntoa**

in Partial Fulfilment of the Requirements
for the Degree of Doctor of Philosophy

APPROVED BY:

Author: Stavroula Ntoa, University of Crete

Supervisor: Constantine Stephanidis, Professor, University of Crete

Dimitris Plexousakis, Professor, University of Crete, Director of ICS-FORTH

Dimitris Grammenos, Principal Researcher, ICS-FORTH

Panayiotis Zaphiris, Professor, Cyprus University of Technology

Georgios Papagiannakis, Associate Professor, University of Crete

Stavroula-Evita Fotinea, Principal Researcher, ILSP-ATHENA

Margherita Antona, Principal Researcher, ICS-FORTH

Department Chairman: Bilas Angelos, Professor, University of Crete

Heraklion, December 2017

To my cornerstones and my sunshines,

for all their support and encouragement
that made the pursuit of this work feasible

ACKNOWLEDGEMENTS

Working towards the realization of this thesis has been both a demanding and rewarding process, a route however to which I was not alone. The epilogue of this thesis aims at expressing my gratitude to all those who have helped towards the accomplishment of this work. First and foremost, I am grateful to my supervisor, Professor Constantine Stephanidis, for all the guidance, trust and support he has wholeheartedly provided me during this thesis. Having worked with him for almost twenty years, he has constituted a source of inspiration and motivation through his organized, sharp, hard-working and caring approach to all matters. I am also really thankful to Professor Dimitris Plexousakis and Dr. Dimitris Grammenos, for their insightful recommendations and advice that have effectively helped me to extend the horizons of my work.

I would like to particularly thank Dr. Margherita Antona, for her multilevel assistance, highlighting weaknesses and appraising strengths, and being always there to discuss any considerations I had. I also owe my thanks to Professor Panayiotis Zaphiris, Professor Georgios Papagiannakis, and Dr. Stavroula-Evita Fotinea, for their constructive and substantial suggestions and comments.

Many thanks to friends and colleagues for their support, and especially to Ms. Ilia Adami, Ms. Chryssi Birliraki, Ms. Irini Kontaki, Dr. Maria Korozi, Mr. Vassilis Kouroumalis, and Ms. Eleni Papadaki for their valuable remarks towards improving the UXAmI framework and its tools, as well as to Mr. Achilleas Tsiolkas for implementing the services of the Ambient Intelligence environment, in which user-based experiments were carried out in the context of this thesis.

As a final point, I would like to express my gratitude to all those who have supported me on a personal level. I will always be grateful to my parents, Nikos and Kaiti, for all that they have done and continue doing for me, being my greatest fans and eternal supporters. I also thank my daughters, Joanna and Nikolia, for all their understanding when mom had to be many hours away from home and for brightening those hard-working days. Last, I am especially thankful to my husband George, work and life partner, for all his encouragement and support, being the person I can count on at all times.

ABSTRACT

Keywords: User Experience evaluation, Ambient Intelligence, evaluation framework, evaluation tools, user testing, automated user experience evaluation, tool for working with guidelines, expert-based review, automated guidelines suggestion, user reward scheme, online community

Supervisor: Constantine Stephanidis, Professor,
Computer Science Departiment, University of Crete

UXAmI Framework: User Experience Evaluation in Ambient Intelligence Environments

Ambient Intelligence (AmI) constitutes a new human-centred technological paradigm, where technologically advanced environments that feature interconnected and embedded devices, supported by sensors' network, computer vision, as well as reasoning and adaptation capabilities, are oriented towards anticipating and satisfying the needs of their inhabitants. In this context, and in view of the not distant realization of AmI environments, evaluation becomes of paramount importance.

Evaluation constitutes a central concept in Human-Computer Interaction, exhibiting increased interest and confronting novel challenges, as technology evolves from the desktop paradigm and contexts expand beyond the organizational domain to almost any life activity. To this end, several efforts have attempted to "frame" evaluation and define how it should be pursued in terms of usability, user experience, as well as interaction adaptation and ubiquitousness. Nevertheless, as technology advances, the number of parameters to be assessed becomes too large to be studied through user experiment observators' notes, or evaluation questionnaires to be filled-in by users (a common current practice when evaluating user experience). On the other hand, despite the fact that the notion of Ambient Intelligence exists for more than a decade and the vital importance of evaluation, efforts in the domain have mainly focused in identifying the challenges in the field and advocating the importance of in situ evaluations, while there is a lack of generic and systematic approaches towards user experience evaluation in Ambient Intelligence.

This thesis proposes a novel comprehensive conceptual and methodological framework, named UXAmI, for the evaluation of user experience in AmI environments, aiming to assess a wide range of characteristics and qualities of such environments, taking into account traditional and modern models and evaluation approaches. Adopting an iterative approach, the framework suggests metrics to be assessed through expert-based reviews during the early stages of development, and user-based evaluations for the latter development stages of an AmI system or environment. Taking advantage of the infrastructure of AmI environments, UXAmI framework proposes the automatic assessment of several attributes during user-based evaluation. A combination of automated measurements, user observation, questionnaires and interviews is expected to allow evaluators to gain insight into the composite nature of user experience in

AmI environments, studying issues related to intuitiveness, unobtrusiveness, adaptivity, usability, cross-platform and multi-user usage, implicit interactions, appeal and emotions, safety and privacy, as well as user acceptance. Finally, a number of tools are proposed in the context of the current thesis, aiming to assist UX engineers in carrying out evaluations in AmI environments based on the UXAmI framework. These include a tool for expert-based reviews against guidelines, a tool for aggregating experimental data and analysing the results of user testing experiments, and a professional networking platform for UX engineers, which will act as an information resource and a means for collaboration, integrating the other two tools as a reward to active and loyal community members.

ΠΕΡΙΛΗΨΗ

Λέξεις κλειδιά: Αξιολόγηση εμπειρίας χρήσης, Διάχυτη Νοημοσύνη, πλαίσιο αξιολόγησης, εργαλεία αξιολόγησης, αξιολόγηση με χρήστες, αυτοματοποιημένη αξιολόγηση εμπειρίας χρήσης, αξιολόγηση με εμπειρογνώμονες, εργαλείο για εργασία με οδηγίες, αυτόματη πρόταση οδηγιών, μηχανισμός ανταμοιβής χρηστών, διαδικτυακή κοινότητα

Επόπτης: Κωνσταντίνος Στεφανίδης, Καθηγητής,
Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

UXAmI Framework: Αξιολόγηση της εμπειρίας χρήσης σε περιβάλλοντα Διάχυτης Νοημοσύνης

Η Διάχυτη Νοημοσύνη (ΔΝ) αποτελεί ένα νέο ανθρωποκεντρικό τεχνολογικό παράδειγμα, όπου τεχνολογικά προηγμένα περιβάλλοντα που απαρτίζονται από διασυνδεδεμένες και ενσωματωμένες συσκευές, δίκτυα αισθητήρων, υπολογιστική όραση, καθώς και δυνατότητες συλλογιστικής και αυτοπροσαρμογής, προσανατολίζονται στην πρόβλεψη και ικανοποίηση των αναγκών των κατοίκων τους. Σε αυτό το πλαίσιο και εν όψει της όχι τόσο μακρινής πραγμάτωσης των περιβαλλόντων ΔΝ, η αξιολόγηση λαμβάνει εξέχουσα σημασία.

Η αξιολόγηση αποτελεί κεντρική έννοια στην Αλληλεπίδραση Ανθρώπου – Υπολογιστή, προσελκύοντας αυξανόμενο ενδιαφέρον και αντιμετωπίζοντας νέες προκλήσεις καθώς η τεχνολογία εξελίσσεται πέρα από το υπόδειγμα της επιφάνειας εργασίας (desktop paradigm), ενώ τα πιθανά πλαίσια χρήσης επεκτείνονται εκτός από το εργασιακό περιβάλλον σε οποιαδήποτε σχεδόν δραστηριότητα του ανθρώπου. Προς αυτή την κατεύθυνση, αρκετές προσπάθειες έχουν επιχειρήσει να ορίσουν ένα πλαίσιο για την αξιολόγηση και να καθορίσουν πώς θα πρέπει να επιδιώκεται σε ότι αφορά την ευχρηστία και την εμπειρία χρήσης, καθώς και για αυτοπροσαρμοζόμενα περιβάλλοντα και περιβάλλοντα πανταχού παρούσας υπολογιστικής δύναμης (ubiquitous computing). Ωστόσο, με την πρόοδο της τεχνολογίας, ο αριθμός των παραμέτρων που πρέπει να αξιολογούνται καθίσταται πολύ μεγάλος για να μελετηθεί μέσω σημειώσεων των παρατηρητών μιας συνεδρίας αξιολόγησης, ή μέσω ερωτηματολογίων τα οποία απευθύνονται στους χρήστες (μια συνηθισμένη τρέχουσα πρακτική κατά την αξιολόγηση της εμπειρίας χρήσης). Αφ' ετέρου, παρά το γεγονός ότι η έννοια της Διάχυτης Νοημοσύνης υπάρχει για περισσότερο από μια δεκαετία και παρά την καίρια σημασία της αξιολόγησης, οι προσπάθειες στο πεδίο έχουν επικεντρωθεί στον προσδιορισμό των προκλήσεων και στην προώθηση της σημασίας των επί τόπου αξιολογήσεων (in situ), ενώ παρατηρείται έλλειψη γενικευμένων και συστηματικών προσπαθειών προς την κατεύθυνση της αξιολόγησης της εμπειρίας χρήσης σε περιβάλλοντα Διάχυτης Νοημοσύνης.

Η παρούσα διατριβή προτείνει ένα καινοτόμο, περιεκτικό εννοιολογικό και μεθοδολογικό πλαίσιο, που ονομάζεται UXAmI, για την αξιολόγηση της εμπειρίας χρήσης σε περιβάλλοντα ΔΝ, στοχεύοντας στην αξιολόγηση ενός μεγάλου εύρους χαρακτηριστικών και ιδιοτήτων τέτοιων περιβαλλόντων, λαμβάνοντας υπόψη παραδοσιακά και σύγχρονα μοντέλα και προσεγγίσεις αξιολόγησης. Υιοθετώντας μια επαναληπτική προσέγγιση, το πλαίσιο προτείνει μετρικές που προσδιορίζονται μέσω επιθεωρήσεων από εμπειρογνώμονες (expert-based reviews) κατά τα αρχικά στάδια της ανάπτυξης, και μέσω αξιολογήσεων με χρήστες σε μετέπειτα στάδια της ανάπτυξης ενός συστήματος ή περιβάλλοντος ΔΝ. Επωφελούμενο από την υποδομή των περιβαλλόντων ΔΝ, το προτεινόμενο UXAmI πλαίσιο εισηγείται την αυτόματη ανάκτηση ποικίλων παραμέτρων κατά τη διάρκεια αξιολογήσεων με χρήστες. Ο συνδυασμός των αυτόματων μετρήσεων, της παρατήρησης των χρηστών, των ερωτηματολογίων και των συνεντεύξεων με χρήστες αναμένεται να βοηθήσει τους αξιολογητές να αποκτήσουν επίγνωση των σύνθετων ζητημάτων της εμπειρίας χρήσης σε περιβάλλοντα ΔΝ, μελετώντας παραμέτρους που αφορούν στη διαισθητικότητα, μη παρεμβατικότητα (unobtrusiveness), αυτοπροσαρμογή, ευχρηστία, χρήση σε πολλαπλές πλατφόρμες και από πολλαπλούς χρήστες, συνεπαγόμενες αλληλεπιδράσεις (implicit interactions), ελκυστικότητα και συναισθήματα, ασφάλεια και ιδιωτικότητα, καθώς και αποδοχή από τους χρήστες. Τέλος, στο πλαίσιο της παρούσας διατριβής προτείνονται εργαλεία που στοχεύουν στην υποβοήθηση των μηχανικών εμπειρίας χρήσης (user experience engineers) κατά τη διεξαγωγή αξιολογήσεων σε περιβάλλοντα ΔΝ με τη χρήση του πλαισίου UXAmI. Αυτά περιλαμβάνουν ένα εργαλείο για αξιολογήσεις από εμπειρογνώμονες βάσει οδηγιών (guidelines), ένα εργαλείο για τη συλλογή δεδομένων από αξιολογήσεις με χρήστες και ανάλυση των αποτελεσμάτων τους, καθώς και μια επαγγελματική πλατφόρμα δικτύωσης για μηχανικούς εμπειρίας χρήσης, η οποία θα δρα ως πηγή πληροφόρησης και μέσο συνεργασίας, ενσωματώνοντας τα άλλα δύο εργαλεία ως επιβράβευση των ενεργών και τακτικών μελών της.

Table of Contents

List of Figures

List of Tables

Abbreviations

AAL	Ambient Assisted Living
AmI	Ambient Intelligence
API	Application Programmable Interface
AR	Augmented Reality
AUI	Adaptive User Interface
CA	Cognitive Absorption
CBR	Case-Based Retrieval
CRUD	Create, Read, Update, Delete
DLL	Dynamic-Link Library
DTPB	Decomposed Theory of Planned Behavior
ERP	Enterprise Resource Planning
ESM	Experience Sampling Method
HCI	Human-Computer Interaction
HMD	Head-Mounted Display
HR	Heart Rate
HTA	Hierarchical Task Analysis
IAS	Interactive Adaptive System
IDE	Integrated Development Environment
IDT	Innovation Diffusion Theory
IQ	Information Quality
IS	Information Society
IST	Information Society Technology
IT	Information Technology
ISTAG	IST Advisory Group
Lo-Fi	Low Fidelity
MATH	Model of Adoption of Technology in Households

OC	Online Community
PDA	Personal Digital Assistant
PC	Personal Computer
POI	Point of Interest
PTAM	Pervasive Technology Acceptance Model
QoE	Quality of Experience
QoS	Quality of Service
Q&A	Question and Answer
RFID	Radio-Frequency Identification
SC	Skin Conductivity
SCT	Social Cognitive Theory
SDK	Software Development Kit
SE	Software Engineering
SQ	System Quality
SMS	Short Message Service
SNS	Social Networking Site
TAM	Technology Acceptance Model
TFWWG	Tools for Working with Guidelines
TPB	Theory of Planned Behavior
TRA	Theory of Reasoned Action
UBICOMP	Ubiquitous Computing
UCD	User Centred Design
UEM	Usability Evaluation Method
UGC	User Generated Content
UI	User Interface
UM	User Modelling
US	United States
UTAUT	Unified Theory of Acceptance and Use of Technology

UX	User Experience
VR	Virtual Reality
WoZ	Wizard of Oz
W3C	World Wide Web Consortium
WWW	World Wide Web

1   INTRODUCTION

Ambient Intelligence is an emerging field of research and development, constituting a new technological paradigm. The notion of Ambient Intelligence is becoming a de facto key dimension of the Information Society, since next generation digital products and services are explicitly designed in view of an overall intelligent computational environment (Stephanidis, 2012). Although Ambient Intelligence is a multidisciplinary field, its objective is to support and empower users, therefore the main thrust of research in AmI should emphasize how and whether this goal is achieved, while in this context it is important to consider the implications of user evaluation (Augusto, Nakashima, & Aghajan, 2010).

Evaluation is a core concern in HCI, with the concepts of technology acceptance, usability and user experience (UX) evaluation constituting the focus of many research efforts that aim to provide answers to what makes a technology usable, acceptable, and the entire experience of using it positive. Although the notions of technology acceptance and usability are not novel, it is notable that as technology moves beyond the typical desktop paradigm, they still constitute the objective of active research, through the development of methods, tools, and theoretical frameworks to assess them.

Technology acceptance is defined by two principal factors, namely perceived ease of use and perceived usefulness (Davis, 1985). However, as technology has evolved from the typical personal computer (PC) to smartphones, tablets, and microcomputers hidden in various devices, while its usage has expanded from the typical workplace domain to several contexts (e.g., household, health, learning, AAL), several other factors have been determined to impact the aforementioned two main factors and eventually technology acceptance. Ntoa, Antona & Stephanidis (2017), in a review of 43 relevant models, identified 73 parameters influencing technology acceptance, the majority of which (98.92%) is assessed in the various studies through questionnaires, asking users to self-report their characteristics, attitudes and perceptions.

Usability is also fundamental in HCI and an essential component of UX (Bevan, 2009b). Since the very first definitions of usability until now, several methods have been proposed aiming to assess the usability of a specific product or service, however studies have identified that two methods are most commonly employed in usability evaluations, namely user testing and expert-based reviews (Paz and Pow-Sang, 2014). With the aim to identify how usability should be measured, several frameworks have been proposed in literature, the most recent ones influenced by the UX notion (Hornbæk & Law, 2007) and incorporating attributes such as quality in use, societal impact, aesthetics, usefulness, and usage continuance intentions, resulting in a breadth of parameters that should be studied. User Experience (UX) has recently predominated the usability concept, providing a broader perspective on a user's experience with a product, aiming, according to the related ISO standard, to study "a person's perceptions and

responses resulting from the use and/or anticipated use of a product, system or service", and including all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors and accomplishments that occur before, during and after use of a product or service (ISO 9241-210:2010, 2010). UX methods that go beyond usability evaluation are mainly focused on users' perceptions of system attributes (e.g., aesthetics, playfulness, and fun) as well as on the emotions induced by system usage. In an effort to provide a more systematic approach towards assessing UX several frameworks have been proposed, the majority of which have however remained conceptual.

The prevalence of mobile devices has led to the materialization of Ubiquitous Computing (UbiComp), a term referring to the third era of modern computing, which is "characterized by the explosion of small networked portable computer products in the form of smart phones, personal digital assistants (PDAs), and embedded computers built into many of the devices we own—resulting in a world in which each person owns and uses many computers" (Want, 2010). Evaluation of ubiquitous computing systems is a challenging research area, mainly due to the facts that traditional controlled laboratory testing is no longer appropriate, that a wider range of factors should be evaluated, and that multiple systems participate in the entire experience, making it therefore challenging to define the reasons for a success or failure (Neely et al., 2008; Kim, Kim, & Park, 2003; Schmidt, 2003). Realizing the need for a more systematic approach for the evaluation of UbiComp systems, several research approaches have focused in creating evaluation frameworks and models, the majority of which however does not systematically assist evaluators in deciding which evaluation method to choose, or which exact metrics, while at the same time an unmanageable number of parameters to be assessed is proposed.

Ubiquitous computing has constituted an important paradigm shift, but as we are heading towards the fourth era of modern computing, it is expected that the human–computer experience will be more continuous and seamless than ever before, eliminating references to the distinct number of devices per individual (Abowd, 2012). Such a vision may be fulfilled by Ambient Intelligence (AmI), which incorporates the features of UbiComp environments, but focuses on the human inhabitants of the environment, aiming to elevate the overall user experience. Evaluation in Ambient Intelligence is a challenging objective and a field which has not yet been extensively explored, due to the inherent difficulties it imposes. Stephanidis (2012) highlights that the evaluation of AmI technologies and environments needs to advance traditional usability evaluation in a number of dimensions, concerning both the qualities of the environment to be assessed and the assessment methods. A major concern is that evaluation should go beyond performance-based approaches to evaluation of the overall user experience (Stephanidis, 2012; Gaggioli, 2005), which should be further articulated in the context of AmI environment. Furthermore, evaluation should take place in real world contexts (Stephanidis, 2012; Gaggioli, 2005), which is a challenging task by

itself. Also, a challenging aspect of evaluation in Ambient Intelligence environments is the appropriate selection of methods according to the environment, the context of use and the target users. Additionally, although several frameworks have been proposed in the UbiComp context, none have been explored for Ambient Intelligence environments. AmI as a concept is the direct extension of the concept of UbiComp, but it is much more than this, as AmI systems should be adaptive and responsive to the user's needs and behaviour (Bibri, 2015), therefore it is doubtful whether UbiComp models can be adequate for AmI.

Motivated by the need to define how user experience should be assessed in AmI environments, as well as by the general lack of approaches with practical value in the field of evaluation frameworks, this thesis proposes a novel comprehensive framework, named UXAmI, for the evaluation of User Experience in AmI environments, aiming to assess a wide range of characteristics and qualities of such environments, taking into account traditional and modern models and evaluation approaches. The proposed framework adopts an iterative design approach, suggesting specific evaluation approaches for the different development stages of an AmI environment, system, or application, thus allowing the assessment of the user experience from the early stages of the development lifecycle to the final stages of implementation. UXAmI is a clean-cut conceptual and methodological framework, taking into account the various facets and temporal attributes of UX, providing not only concepts, but also concrete metrics and methods to measure them. Furthermore, taking advantage of AmI environments' architecture and sensors' infrastructure, it advocates the automatic identification of specific metrics, alleviating the need for observers to keep lengthy notes or to address all issues through questionnaire items to be answered by users.

In order to achieve an approach that can be practically adopted by UX engineers, the framework is accompanied by tools facilitating different evaluation approaches, as well as knowledge accumulation and exchange through a professional networking platform. More specifically, UXAmI Observer facilitates analysis of user-based experiments, through automatically calculated metrics, insights and statistics. UXAmI Inspector is a tool assisting evaluations by experts, through the suggestion of guidelines appropriate for the specified evaluation context, supporting the inspection process itself. Guidelines in the Inspector tool are suggested according to tags received through crowdsourcing and structured under predefined categories, following a hybrid taxonomy-folksonomy approach. To this end, the UXAmI Online Community serves the crowdsourcing concept, and also aims to become a knowledge resource and personal repository for UX engineers. As the community is based on User-Generated Content (UGC), it is equipped with an innovative adaptive reward scheme to motivate users towards participating in the community and uploading content of good quality. Both Observer and Inspector interoperate with the UXAmI community and are provided as rewards to loyal and active members.

In summary, the contributions of this thesis are:

- A systematic review of 43 technology acceptance models and 41 evaluation frameworks, resulting in a classification of parameters that influence technology acceptance and that should be assessed in the context of usability and user experience evaluation.

- A comprehensive extensible conceptual and methodological framework for the evaluation of UX in AmI environments, featuring 39 novel metrics addressing the issues of awareness of application capabilities and of the interaction vocabulary, distractions, appropriateness and impact of adaptations, appropriateness of recommendations, cross-platform usability, multi-user usability, implicit interactions and usage of the AmI environment.

- An innovative tool to assist evaluators in analysing user-based experiments carried out in AmI environments, providing automatically acquired metrics, insights and powerful visualizations, without any instrumentation requirements.

- A novel tool supporting expert-based reviews of AmI systems and applications, facilitating the specification of targets in the AmI environment and suggesting relevant guidelines that should be taken into account, without any other explicit input by the evaluator

- A professional network for UX engineers, featuring an innovative adaptive reward scheme to foster high quality contributions and active user participation, interoperating with the aforementioned tools.

The remaining of this thesis is structured as follows:

Chapter 2 introduces the Ambient Intelligence concept and provides an overview of the various definitions that have been provided for AmI environments.

Chapter 3 reviews related work in the domains of technology acceptance, usability and user experience evaluation approaches and frameworks, evaluation of adaptive systems, evaluation approaches and frameworks for ubiquitous computing systems, as well as evaluation approaches for Ambient Intelligence environments.

Chapter 4 introduces the UXAmI framework for the UX evaluation of AmI systems, applications and environments - discussing the attributes of AmI environments that should be evaluated, the evaluation approaches that can be employed - and presents the results of its evaluation.

Chapter 5 describes the UXAmI tools that instrument the concepts advocated by the framework, and in particular UXAmI Observer, UXAmI Inspector, and UXAmI Online Community, and provides a use case to illustrate how they can all be used in the context of a UXAmI compatible evaluation process.

Chapter 6 summarizes the current thesis and its contributions, and also provides an overview of future work plans.

2 AMBIENT INTELLIGENCE ENVIRONMENTS

Ambient Intelligence was initially introduced in 2001 through the elaboration of the IST Advisory Group (ISTAG) Ambient Intelligence scenarios in the near future of 2010 (Ducatel, Bogdanowicz, Scapolo, Leijten, & Burgelman, 2001), commissioned by the European Commission. The following definition is provided:

> The concept of Ambient Intelligence (AmI) provides a vision of the Information Society where the emphasis is on greater *user-friendliness*, more *efficient services support*, *user-empowerment*, *and support for human interactions.* People are surrounded by *intelligent intuitive interfaces* that are *embedded* in all kinds of objects and an environment that is capable of *recognizing* and *responding* to the presence of different individuals in a *seamless, unobtrusive and often invisible way*.

Since then, several definitions for Ambient Intelligence were contributed by the research community. A synopsis of the most popular definitions is provided in Table 1.

Table 1. Ambient Intelligence definitions

Authors	Definition
Gaggioli (2005)	In AmI, people will be surrounded by a multitude of *interconnected embedded systems*. These devices will be able to *locate* and *recognize* objects and people, as well as people's intentions. The term "intelligence" in this regard refers to the fact that the digital environment is able to *analyse* the context, *adapt* itself to the people and objects that reside in it, *learn* from their behaviour, and eventually *recognize as well as express emotion*.
Remagnino and Foresti (2005)	In AmI, technologies are deployed to make *computers disappear* in the background, while the human user moves into the foreground in complete control of the augmented environment. AmI is a *user-centric paradigm*, it supports a variety of artificial intelligence methods and works *pervasively*, *nonintrusively*, and *transparently* to *aid the user*.
Augusto and McCullagh (2007)	A digital environment that *proactively*, but *sensibly*, *supports people* in their daily lives.

Aarts and Wichert (2009)	Ambient Intelligence (AmI) is about *sensitive*, *adaptive* electronic environments that *respond* to the actions of persons and objects and *cater* for their needs. This approach includes the entire environment – including each single physical object – and associates it with human interaction. The option of *extended and more intuitive interaction* is expected to result in *enhanced efficiency*, *increased creativity* and *greater personal well-being*.
Cook, Augusto, and Jakkula (2009)	The basic idea behind Ambient Intelligence (AmI) is that by *enriching an environment with technology* (e.g., sensors and devices interconnected through a network), a system can be built such that acts as an "electronic butler", which *senses* features of the users and their environment, then *reasons* about the accumulated data, and finally *selects actions* to take *that will benefit the users* in the environment.
Sadri (2011)	Ambient Intelligence is the vision of a future in which environments *support* the people inhabiting them. This envisaged environment is *unobtrusive*, *interconnected*, *adaptable*, *dynamic*, *embedded*, and *intelligent*. In this vision the traditional computer input and output media disappear. Instead, processors and sensors are integrated in everyday objects.

Moreover, in 2009, Aarts & de Ruyter identified the need for complementing the true intelligence of AmI environments with social intelligence, and introduced three attributes of social intelligence into AmI environments: (a) *socialized*, following social rules and commonly accepted manners and social etiquettes (b) *empathic*, demonstrating understanding and helpful behaviour according to the users' inner state of emotions and motives, and (c) *conscious*, exhibiting a consistent and transparent behaviour in their interaction with people.

Summarising all the above definitions, the characteristics and attributes of an AmI environment include - among others - that they are adaptive, embedded, intuitive, intelligent, interconnected, unobtrusive, and supportive for their inhabitants. A word cloud presenting all the AmI attributes and characteristics that have been encountered in the various AmI definitions, according to their frequency of occurrence is presented in Figure 1.

Figure 1. AmI characteristics and attributes[1]

---

[1] Created with WordItOut: http://worditout.com

3 RELATED WORK

This section studies approaches towards evaluation, which are relevant to the current thesis, each from a different perspective. In more details, the topics of this section include technology acceptance, usability evaluation, user experience evaluation, evaluation of adaptive systems, evaluation of ubiquitous computing systems, and evaluation approaches in Ambient Intelligence environments. Having reviewed the related work in each of the aforementioned domains, the section concludes with a discussion of how the current thesis advances state of the art.

3.1 TECHNOLOGY ACCEPTANCE

Determining what would make a technology acceptable by users was widely recognized as a significant field of research since the seventies, when approaches towards defining factors that seem to influence the use of technology have been proposed. Nevertheless, it was in the mid-eighties when researchers concentrated their efforts in developing and testing models that could help in predicting system use (Legris, Ingham, & Collerette, 2003). Several theoretical models have been proposed to this end, with roots in information systems, psychology, and sociology (Venkatesh, Morris, Davis, & Davis, 2003). The following sections introduce the most significant models, which have been successfully applied towards understanding individual acceptance and usage of various technologies.

3.1.1 MODEL OF PC UTILIZATION

Following a different approach, Thompson, Higgins, and Howell (1991) utilized a subset of the theory of human behaviour (Triandis, 1979), which had not been used until then within the Information Society (IS) context, to create a model of personal computer utilization. The theory of human behaviour makes a distinction between beliefs that link emotions to an act and beliefs that link the act to future consequences, and argues that (i) behavioural intentions are determined by feelings people have toward the behaviour, what they think they should do, and by the expected consequences of the behaviour (ii) behaviour is influenced by habit, behavioural intentions and facilitating conditions. Applying this theory to PC utilization implied that the utilization of a PC by a knowledge worker in an optional use environment would be influenced by the individual's feelings toward using PCs, social norms in the work place concerning PC use, habits associated with computer usage, the individual's expected consequences of using a PC and facilitating conditions in the environment conducive to PC use. The conceptual model proposed by Thompson, Higgins, and Howell (1991) introduced the following factors towards predicting the utilization of a PC (Figure 2):

- **Social factors**, defined as the individual's internalization of the reference groups' subjective culture and specific interpersonal agreements that the individual has made with others in special situations, consisting of norms, roles and values.
- **Affect**, which measures the feelings people have toward the behaviour.
- Perceived consequences, which are represented by three dimensions: (i) **complexity**, defined as the degree to which an innovation is perceived as relatively difficult to understand and use (ii) **job fit**, measured as the extent to which an individual believes that using a PC can enhance the performance of his or her job, and (iii) **long-term consequences** of use, described as outcomes that have a pay-off in the future, such as increasing the flexibility to change jobs or increasing the opportunities for more meaningful work.
- **Facilitating conditions**, described as objective factors in the environment that several judges or observers can agree make an act easy to do, and measured by determining the provision of support for users of PCs.
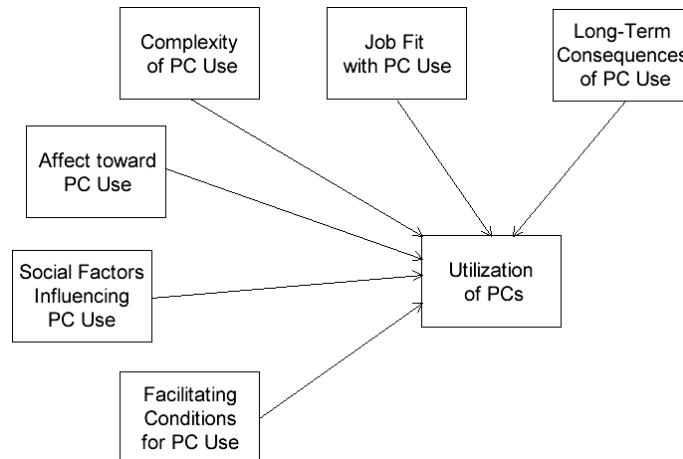
Figure 2. Model of PC utilization

3.1.2 TECHNOLOGY ACCEPTANCE MODEL

One of the most influential models, the Technology Acceptance Model (TAM), has been proposed by Davis (1985) in the context of his PhD thesis, which aimed at improving the understanding of user acceptance process and also to provide a practical user acceptance testing methodology. TAM (Figure 3) defines two components that affect a user's attitude towards using a technology, namely: (i) **perceived usefulness**, described as the degree to which an individual believes that using a particular system would enhance his or her job performance and (ii) **perceived ease of use**, defined as the degree to which an individual believes that using a particular system would be free of physical and mental effort. In order to measure

perceived usefulness and perceived ease of use, Davis proposed two questionnaires, each featuring ten Likert-scale questions.
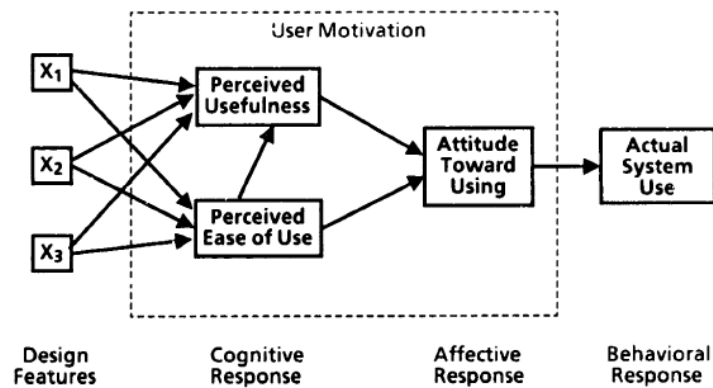
Figure 3. The Technology Acceptance Model

Extending the initial TAM model and taking into account theoretical constructs stemming from social influence processes and cognitive instrumental processes, Venkatesh and Davis (2000) introduced the TAM2 model. More specifically, TAM2 (Figure 4) added seven components to the initial TAM model:

- Subjective norm. A person's perception that most people who are important to him think he should or should not perform the behaviour in question.
- Voluntariness. The extent to which potential adopters perceive the adoption decision to be non-mandatory.
- Image. The degree to which use of an innovation is perceived to enhance one's status in one's social system.
- Experience. The experience gained while using a given technology over time.
- Job relevance. An individual's perception regarding the degree to which the target system is applicable to his or her job.
- Output quality. How well the system performs tasks.
- Result demonstrability. The tangibility of the results using the innovation.
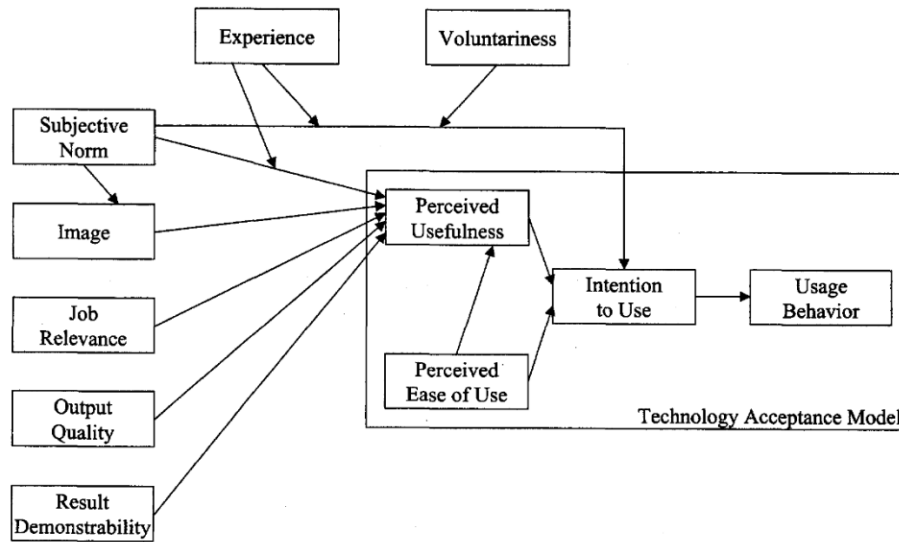
Figure 4. The TAM2 model

TAM has been widely adopted and studied by the research community, resulting in a considerable number of external variables that have been introduced to it, as factors influencing how users perceive the usefulness and ease of use of a technology. These variables include (Lee, Kozar, & Larsen, 2003):

- Relative Advantage. The degree to which an innovation is perceived as being better than its precursor.
- Compatibility. The degree to which an innovation is perceived as being consistent with the existing values, needs, and past experiences of potential adopters.
- Trialability. The degree to which an innovation may be experimented with before adoption.
- Self-efficacy. An individual's convictions about his or her abilities to mobilize motivation, cognitive resources and courses of action needed to successfully execute a specific task within a given context (e.g., computer self-efficacy is defined as an individual judgement of one's capability to use a computer) (Compeau & Higgins, 1995).
- End User Support. Specialized instruction, guidance, coaching and consulting (Igbaria, Guimaraes, & Davis, 1995).
- Objective Usability. A construct that allows for a comparison of systems on the actual level of effect regarding to complete specific tasks.
- Personal Innovativeness. The individual's willingness to try out any new technology.
- Cognitive Playfulness. The individual's cognitive spontaneity when using a technology.
- Social Presence. The degree to which a medium permits users to experience others as being psychologically present.

- Visibility. The degree to which the innovation is visible in the organization.
- Computer Attitude. The degree to which a person likes or dislikes the object.
- Accessibility. Physical accessibility (if someone has physical access to the system) and information accessibility (the ability to retrieve the desired information from the system).
- Management Support. The degree of support from managers to ensure sufficient allocation of resources and act as a change agent to create a more conductive environment for IS success.
- Computer Anxiety. An individual's apprehension, or even fear, when she/he is faced with the possibility of using computers.
- Perceived Enjoyment. The extent to which the activity of using a specific system is perceived to be enjoyable in its own right, aside from any performance consequences resulting from system usage.
- Facilitating Conditions. Resource factors (such as time and money) and technology compatibility issues that may constrain usage.

Addressing the need for defining the determinants of perceived ease of use, TAM3 was proposed (Venkatesh & Bala, 2008) extending TAM2 and including the following determinants: computer self-efficacy, perception of external control, computer anxiety, computer playfulness, perceived enjoyment and objective usability. Perception of external control (or facilitating conditions) is defined as the degree to which an individual believes that organizational and technical resources exist to support the use of the system. Furthermore, TAM3 posits three new relationships, suggesting that experience will moderate the relationships between (i) perceived ease of use and perceived usefulness, (ii) computer anxiety and perceived ease of use, and (iii) perceived ease of use and behavioural intention. Validation of the model through a longitudinal study in four organizations confirmed that experience moderated the effect of perceived ease of use on perceived usefulness such that with increasing experience the effect became stronger and that experience moderated the effect of computer anxiety on perceived ease of use such that the effect became weaker with increasing experience.
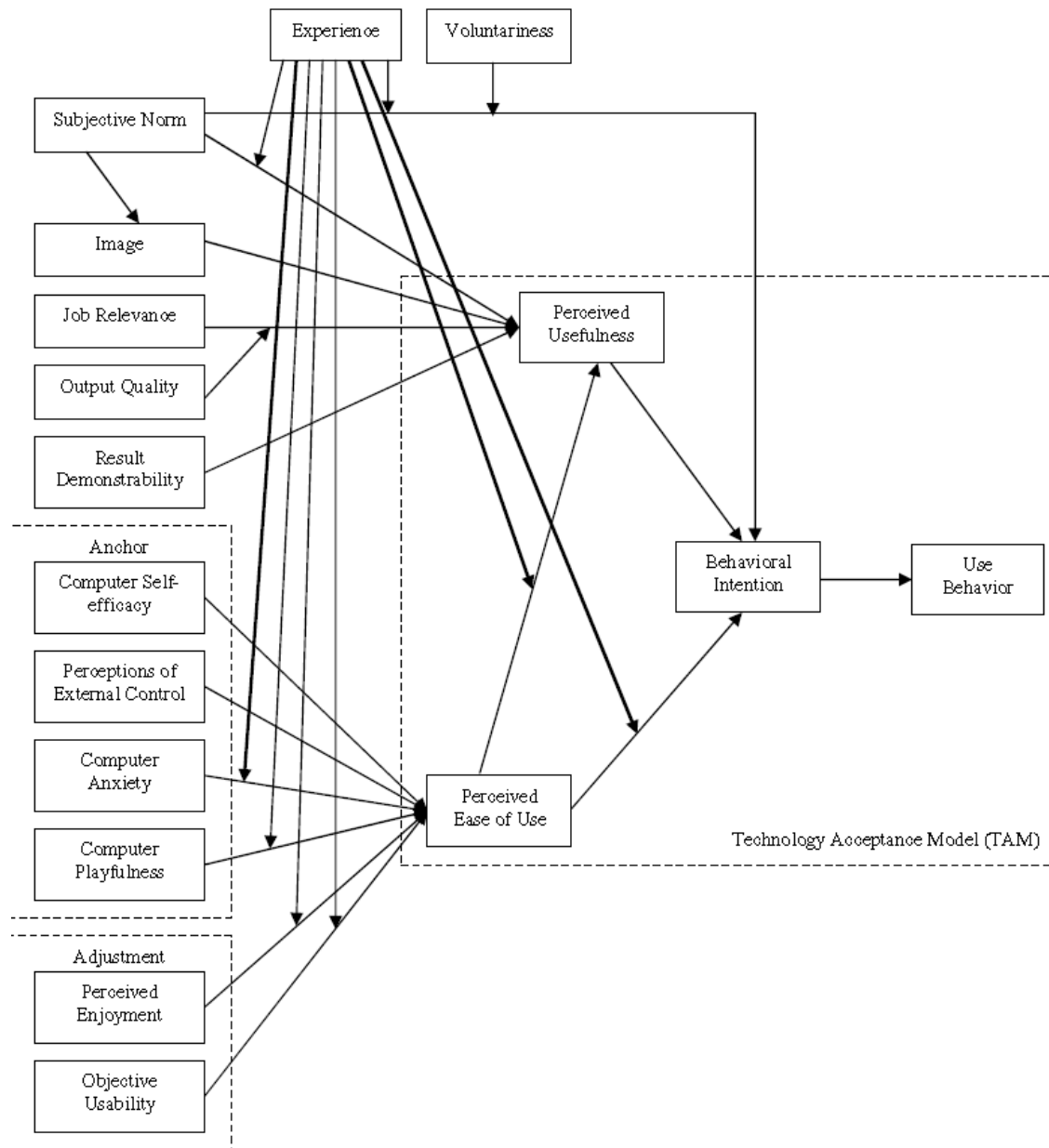
Figure 5. The TAM3 model

3.1.3 THEORY OF PLANNED BEHAVIOR

The theory of planned behaviour (TPB) is based on and extends the Theory of Reasoned Action (TRA) (Ajzen & Fishbein, 1980), which postulates that behavioural intentions are a function of salient information or beliefs about the likelihood that performing a particular behaviour will lead to a specific

outcome. Information or beliefs affect intentions and subsequent behaviour either through attitudes and/or through subjective norms. TPB (Figure 6) extends TRA by adding one more parameter: perceived behavioural control, defined as the perceived ease or difficulty of performing the behaviour determined by the possession of requisite resource and opportunities (Madden, Ellen, & Ajzen, 1992).
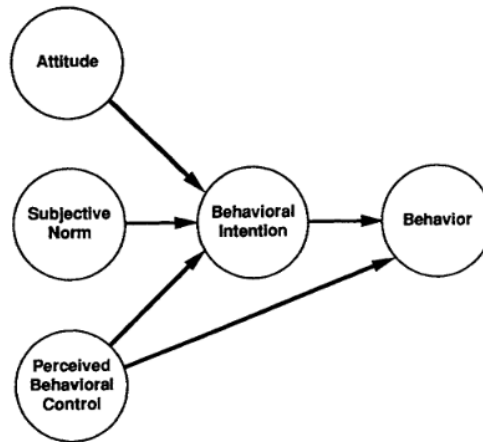
Figure 6. Theory of Planned Behaviour

Taylor and Todd (1995b) proposed a decomposition of the TPB model, incorporating multi-dimensional belief structures and crossover effects to better understand the relationships specified in the model and to improve the explanatory power of the model. The parameters that further decompose the TPB model are (as shown in Figure 7): relative advantages, complexity, compatibility, normative influences, efficacy and facilitating conditions.

Furthermore, highlighting the need for incorporating social and control factors on behaviour into the TAM model, Taylor and Todd (1995a) introduced an augmentation of the TAM model using the TPB model, as shown in Figure 8.
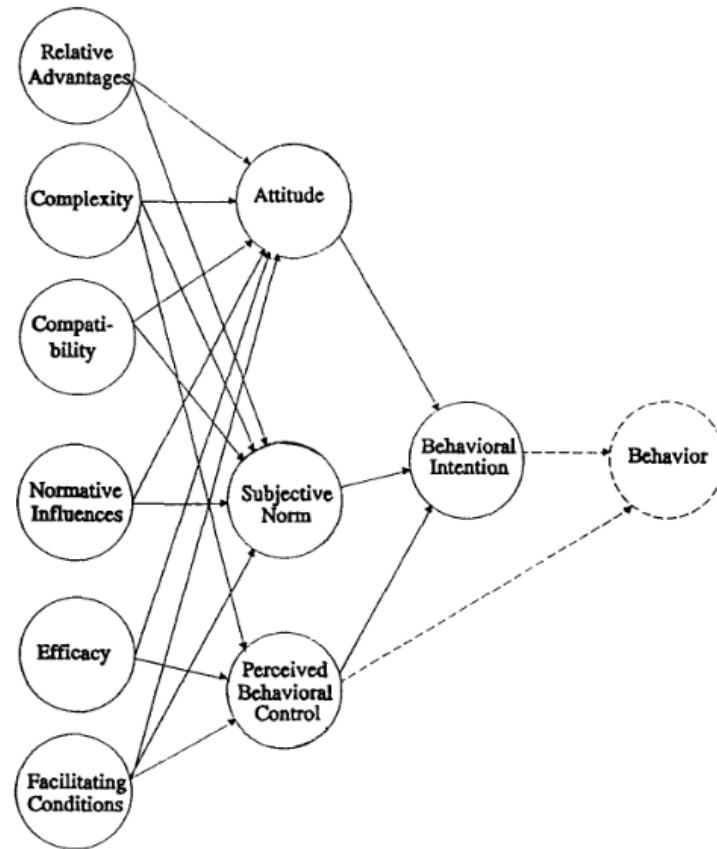
Figure 7. TPB with belief decomposition and hypothesized crossover effects
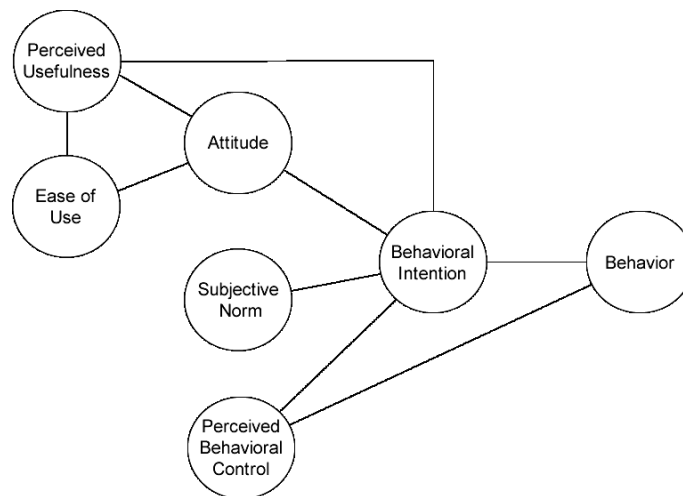
Figure 8. Augmented TAM with TPB

3.1.4 INNOVATION DIFFUSION THEORY

A significant theoretical framework in the area of technology diffusion and adoption was proposed by Rogers (1995), who defined diffusion as "the process by which an innovation is communicated through certain channels over time among the members of a social system". In more details, the innovation-decision process (Figure 9) is described as the process through which an individual passes from first knowledge of an innovation, to forming an attitude toward the innovation, to a decision to adopt or reject, to implementation of the new idea, and to confirmation of this decision (see Figure 9). Therefore, according to this process, besides the perceived characteristics of the innovation there are other factors that determine the adoption of an innovation, including previous practice of the individual, his needs and problems, norms of the social system, the innovativeness of the individual, as well as other socioeconomic characteristics, personality variables and communication behaviour.
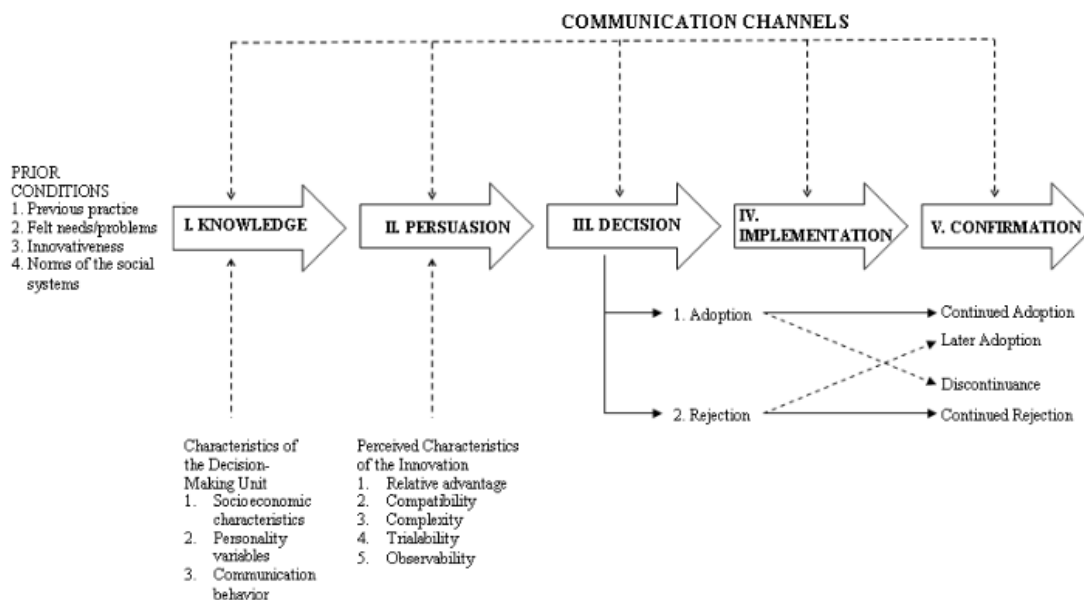
Figure 9. The innovation-decision process

Rogers described the innovation-diffusion process as an uncertainty reduction process and proposed five attributes of innovation, which are important for its rate of adoption: relative advantage, compatibility, complexity, trialability, and observability. Other variables determining the rate of adoption of innovations are: the type of innovation decision (optional, collective, authority), the communication channels used to diffuse an innovation, the nature of the social systems (its norms and the degree to which the communication network structure is highly interconnected), as well as the promotion efforts of change agents.

Moore and Benbasat (1991) adopted the Innovation Diffusion Theory (IDT) and further extended it with two constructs, in order to create an instrument that would measure users' perceptions of adopting an information technology innovation. The two constructs added were image and voluntariness of use. Tornatzky and Klein (1982) carried out a review and meta-analysis of seventy-five articles addressing innovation characteristics and their relationship to innovation adoption and implementation, and identified ten characteristics as the most important and frequent ones, five of which are the attributes of innovation of IDT. The additional five innovation characteristics are:

- Cost. The cost of an innovation is assumed to be negatively related to the adoption and implementation of the innovation; the less expensive the innovation, the more likely it will be quickly adopted and implemented.
- Communicability, defined as the degree to which aspects of an innovation may be conveyed to others. This feature was found to be very similar to that of observability and never rated by the innovation adopters.
- Divisibility, defined as the extent to which an innovation can be tried on a small scale prior to adoption, which is closely related to trialability.
- Profitability, which is the level of profit to be gained from adoption of the innovation.
- Social approval, which refers to status gained in one's reference group, a nonfinancial aspect of reward as a function of adopting a particular innovation.

3.1.5   SOCIAL COGNITIVE THEORY

In 1986, Bandura (Bandura, 1986) proposed the Social Cognitive Theory (SCT), a landmark work in psychology. Social cognitive theory explains psychosocial functioning as a triadic reciprocal causation of the following dynamics: (a) internal personal factors in the form of cognitive, affective and biological events; (b) behavioural patterns, and (c) environmental events, which interact and influence one another bi-directionally. Inspired by SCT and the fact that the model explicitly acknowledged the existence of a continuous reciprocal interaction between the environment in which an individual operates, his or her cognitive perceptions, and behaviour, Compeau, Higgins and Huff (1999) proposed a model to test the influence of computer self-efficacy, outcome expectations, affect, and anxiety on computer usage (Figure 10). The model accounts for the following factors, which have an effect on computer usage:

- Self-efficacy, which reflects an individual's beliefs about his or her capabilities to use computers.
- Outcome expectations (performance), defined as the perceived likely consequences of using computers associated with improvements in job performance (efficiency and effectiveness).

- Outcome expectations (personal), which relate to expectations of change in image or status or to expectations of rewards, such as promotions, raises, or praise.
- Affect, which represents the enjoyment a person derives from using computers.
- Anxiety, representing the feelings or apprehension or anxiety one experiences when using a computer.
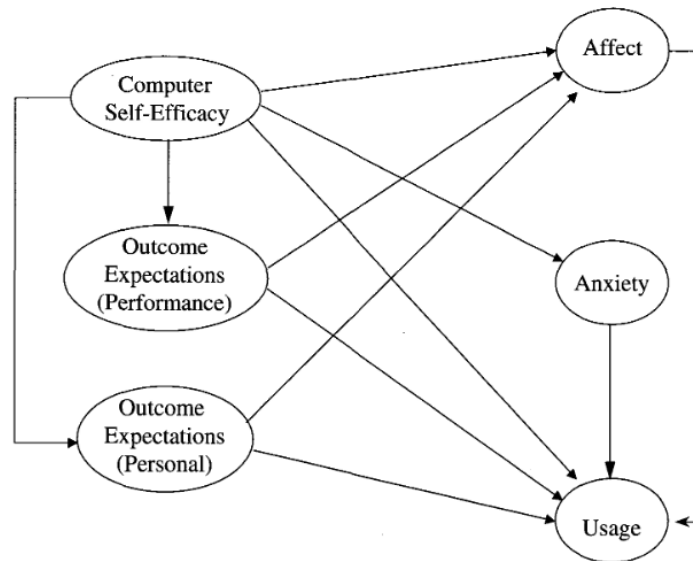
Figure 10. Model of effect of computer self-efficacy, outcome expectations, affect, and anxiety on computer usage

3.1.6   UNIFIED THEORY OF ACCEPTANCE AND USE OF TECHNOLOGY

With the aim to facilitate researchers confronted with a choice among a multitude of models, Venkatesh, Morris, G. B. Davis and F. D. Davis (2003) proposed the Unified Theory of Acceptance and Use of Technology (UTAUT), based on a literature review and a selection of eight prominent models.

According to this theory, four constructs are direct determinants of user acceptance and user behaviour (Figure 11):

- Performance expectancy, defined as the degree to which an individual believes that using the system will help him to attain gains in job performance. This construct is directly related to the following constructs employed by other models: perceived usefulness, extrinsic motivation, job-fit, relative advantage and outcome expectations.

- **Effort expectancy**, which is the degree of ease associated with the use of the system. Constructs from other models that capture the concept of effort expectancy are: perceived ease of use, and complexity.

- **Social influence**, defined as the degree to which an individual perceives that important others believe he or she should use the new system. Constructs from other models relevant to social influence are: subjective norm, social factors and image.

- **Facilitating conditions**, described as the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system. This definition is also referenced by other models as: perceived behavioural control and compatibility.

In addition, four moderators have been identified for the aforementioned determinants, namely: gender, age, experience and voluntariness of use.
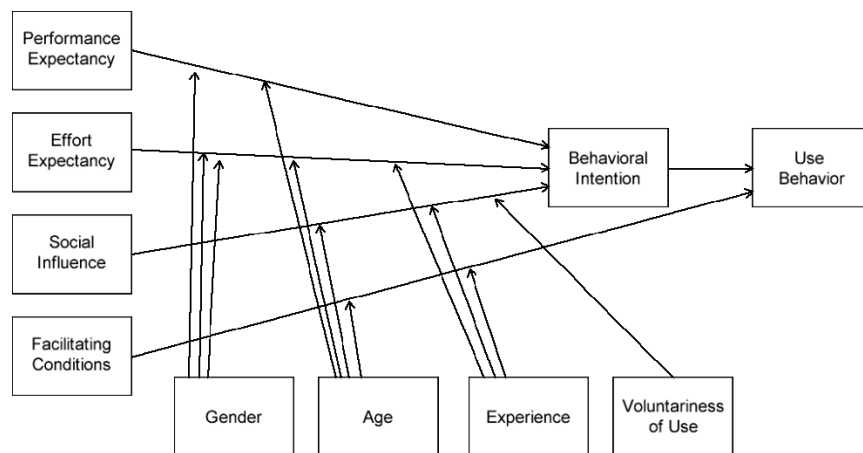
Figure 11. UTAUT model

An extension to the UTAUT model, named UTAUT2 (Figure 12), has been proposed by Venkatesh, Thong, and Xu (2012) to study acceptance and use of technology in a consumer context and incorporates three additional constructs:

- **Hedonic motivation**, defined as the fun or pleasure derived from using a technology.

- **Price value**, which is the consumers' cognitive trade-off between the perceived benefits of the applications and the monetary cost for using them.

- **Habit**, defined as the extent to which people tend to perform behaviours automatically because of learning.
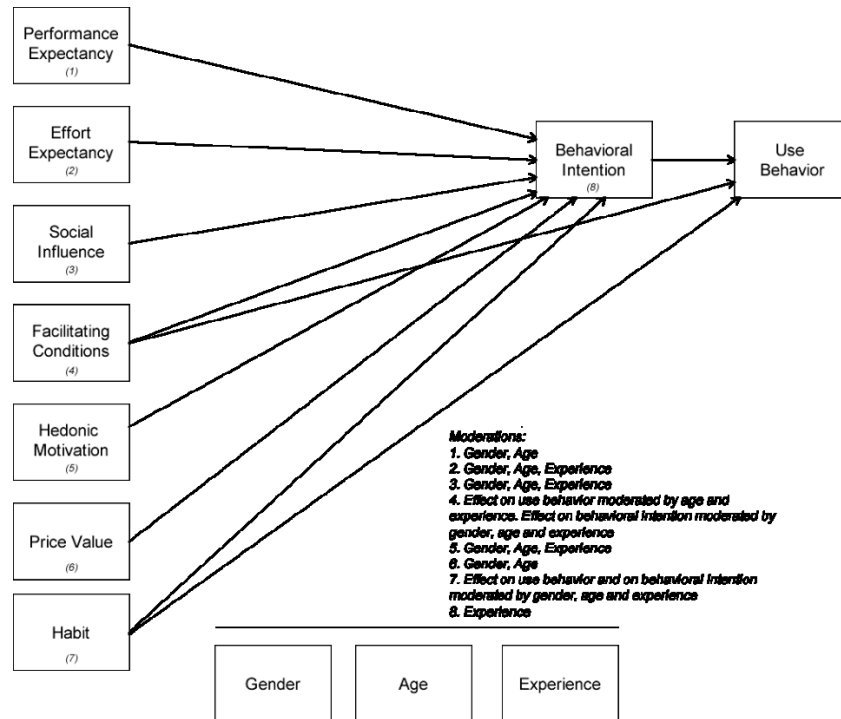
Figure 12. UTAUT2 model

3.1.7 TECHNOLOGY ACCEPTANCE FOR CONTEXTS BEYOND THE WORKPLACE

The majority of the aforementioned fundamental models have initially been applied in organizational settings examining technology adoption in the workplace context, as when they were initially created computers were not used in home or other environments, while technology mostly referred to computer usage. Recent advances of technology have lead however to increased research interest in assessing technology acceptance in a variety of domains. This section reports on the most noteworthy efforts utilising or extending the aforementioned models by adding new variables, towards assessing other contexts or technologies, focusing in the most prevalent contexts and contexts relevant to this thesis (e.g., ubiquitous computing).

Technology Adoption in Households

As a result of studying technology adoption in households Brown and Venkatesh (2005) introduced the Model of Adoption of Technology in Households (MATH), which is presented in Figure 13 and includes the following constructs:

- Utilitarian outcomes, which can be divided into beliefs related to personal use, children, and work.

- **Hedonic outcomes**, defined as the pleasure derived from the consumption, or use, of a product.
- **Social outcomes**, which are described as the "public" recognition that would be achieved as a result of adopting an innovation.
- **Social influence**, which is the extent to which members of a social network influence one another's behaviour and can be further classified to friends and family influences, secondary sources influences, as well as workplace referents' influences.
- **External constraints**, which are characteristics of the PC and its environment and include the rapid change in technology and/or fear of obsolescence, declining cost, and cost.
- **Internal constraints**, reflecting perceptions of the individual's relationship with technology and include the perceived ease of use and requisite knowledge.

In addition, the model defines the following moderators which are related to household life: marital status, age, child's age and income.
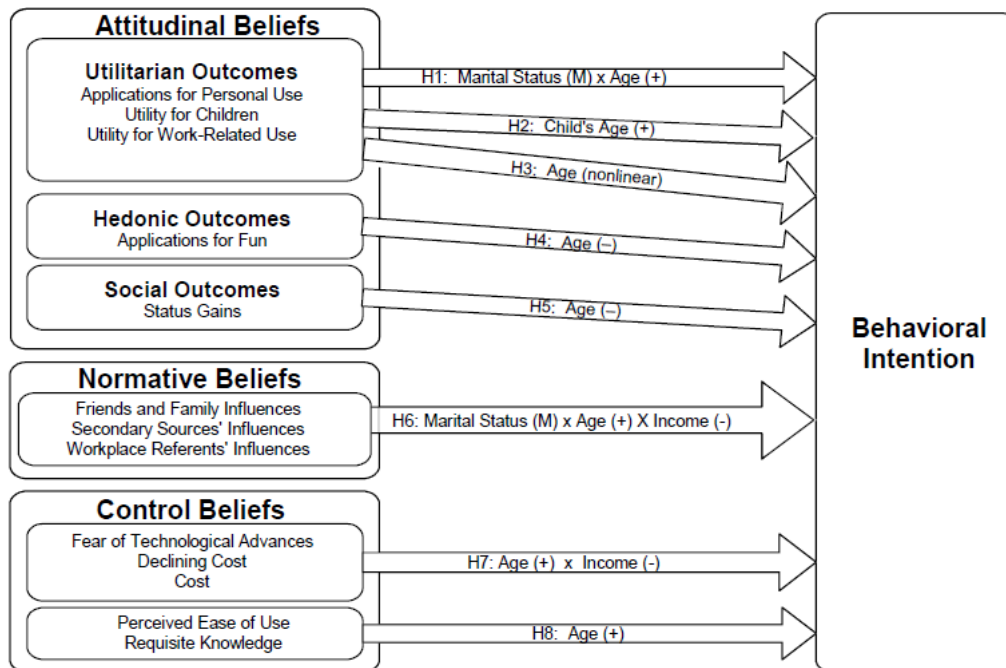
Figure 13. Model of adoption of technology in households

Internet and the World Wide Web (WWW)

Moon and Kim (2001) extended and empirically validated TAM for the WWW context. The results of their study indicate that perceived usefulness, perceived ease of use and perceived playfulness are important determinants of users' perceptions towards using the WWW, but also that playfulness and perceived ease

of use (intrinsic motivations) had a more powerful impact than perceived usefulness (extrinsic motivation) in the case of the WWW. Porter and Donthu (2006) extended TAM to explore specific demographics to explain Internet usage and found that among other factors, education affected the perceived ease of use. The effect of Internet experience and website experience has been studied by Castañeda, Muñoz-Leiva, and Luque (2007), highlighting the positive impact of experience. More specifically, it was found that for users with high experience (a) the influence of perceived usefulness on the process of forming the attitude to the website is substantially greater than for users with low experience, while (b) the influence of perceived ease of use on the attitude towards the website is substantially smaller than for users with low experience.

Gaming and Virtual Worlds

In the domain of WWW, and especially with regard to online games, Hsu and Lu (2004) extended TAM with the constructs of social norms, critical mass and flow experience, and concluded that social norms and flow affect users' intention to play an online game, while critical mass affects users' attitude towards playing an online game, but not intention directly. Focusing on serious games, Yusoff, Crowder, and Gilbert (2010) extended TAM with the concepts of transfer of learnt skills (applying previously acquired skills to other learning), learner control (learners like to explore on their own and pick up skills within the game at their own pace), reward (incentives used to encourage and motivate the learner), as well as situated and authentic learning (using familiar background or common examples in a game's content, relevant to the learner's experience). In the context of virtual worlds, the application of TAM highlighted that communication, collaboration, and cooperation are central in influencing behavioural intention to use and acceptance of the virtual world (Fetscherin & Lattemann, 2008).

Trading, Shopping and Internet banking

The moderating effect of perceived trust has been explored as an extension of TAM in the context of online trading systems (Carlos Roca, José García, & José de la Vega, 2009). Testing the model supported that trust is an important antecedent of user acceptance in this context, and that perceived security affects user's trust. Trust and perceived risk have also been added as extensions to TAM with regard to e-commerce in order to study the user's intention to transact (Pavlou, 2003). Studies that have been carried out to test the extended TAM indicated that trust is positively associated with intention to transact, perceived usefulness and perceived ease of use, and negatively associated with perceived risk. Furthermore, reputation was a significant antecedent of intention to transact, and along with satisfaction with past transactions and web shopping frequency, they were significant antecedents of trust. Trust has been extensively studied by Gefen, Karahanna, and Straub (2003), who introduced a model based on TAM,

investigating how consumer trust along with perceived usefulness and perceived ease of use explain variance in the intended behaviour. The study specifies the following antecedents of trust:

- Personality-based trust, which refers to the individual's tendency to believe or not to believe in others and so trust them, and is relevant for initial trust formation.
- Cognition-based trust, which examines how trust is built on first impressions rather than through experiential personal interactions and is formed via categorization and illusions of control, and is relevant for initial trust formation.
- Knowledge-based trust, and more specifically familiarity with the vendor, which is expected to reduce social uncertainty.
- Calculative-based trust, which is shaped by rational assessments of the costs and benefits of another party cheating or cooperating in the relationship.
- Institution-based trust, which refers to one's sense of security from guarantees, safety nets, or other impersonal structures inherent in a specific context.

Previous experience with the Internet was found to be of significant importance for both initial and repeated purchases, while users who consider that they have more competence and capacity also have better perceptions about e-commerce and, as a consequence, carry out more online purchases (Hernández, Jiménez, & Martín, 2010). E-shopping quality is another factor that was found to be influential (Ha & Stoel, 2009) in perceptions of usefulness, trust, and enjoyment, which in turn influence consumers' attitudes toward e-shopping. In this study e-shopping quality consists of four dimensions, namely web site design, customer service, privacy / security and atmospheric / experiential quality.

The role of perceived risk, as well as that of perceived benefit, have been included in a TAM extension studying user acceptance of internet banking (Lee, 2009). In more details, the results of the study confirmed that perceived benefit has a primary effect on intention to use online banking, as well as that security, financial, time, social and performance risks all emerged as negative factors on the intention to adopt online banking. Risks have been further explored and analysed as a parameter for e-services adoption by Featherman and Pavlou (2003), comprising the facets of performance, financial, time, psychological, social, privacy and overall risk.

eLearning and mLearning

In the context of eLearning, the TAM model has been expanded to include system characteristics, and more specifically: (i) functionality, which refers to the perceived ability of an e-learning system to provide flexible access to instructional and assessment media, (ii) interactivity, which refers to interaction support

between teachers and students, and students themselves, and (iii) response time (Pituch & Lee, 2006). The model also included the user attributes of self-efficacy and internet experience, and studied the impact of the aforementioned factors on perceived usefulness and perceived ease of use, as well as use of the system for supplementary learning and use for distance education. Saadé and Bahli (2005) extended TAM taking into account the moderating effect of cognitive absorption, which in turn is defined by the user's temporal dissociation, focused immersion and heightened enjoyment when using the online learning system. The role of cognitive absorption as well as system attributes has been pointed out in a TAM extension based on the expectancy disconfirmation theory (Roca, Chiu, & Martínez, 2006). The results of the study suggest that continuance intention is determined by satisfaction, which in turn is jointly determined by perceived usefulness, information quality, confirmation, service quality, system quality, perceived ease of use and cognitive absorption.

eLearning self-efficacy, followed by subjective norm, have been emphasized as the most important constructs explaining eLearning technology adoption by university students (Park, 2009). The role of eLearning experience on continuance intention has also been explored by Lin (2011), highlighting that (i) negative critical incidents and attitude are the main determinants of the users' intention to continue using an e-learning system, irrespective of their level of e-learning experience, (ii) the impact of negative critical incidents on perceived ease of use is greater for less experienced users, while the impact of negative critical incidents on perceived usefulness is greater for more experienced users; and (iii) perceived ease of use has a more critical effect on the attitude and continuance intention of less experienced users, whereas perceived usefulness is found to be a stronger determinant of the attitude and behavioural intention of more experienced users. The importance of digital literacy in eLearning use for professional development has been stressed in a study extending the UTAUT model (Mohammadyari & Singh, 2015), which found that digital literacy has an impact on users' performance and effort expectations, which in turn affect continuance intention and eventually performance. On the other hand, in terms of mLearning adoption intention, near-term usefulness, long-term usefulness and personal innovativeness have proved to have significant influence, with the most influential predictor being long-term usefulness (Liu, Li, & Carlsson, 2010).

In summary, eLearning is a domain in which many studies have been carried out in terms of user acceptance. A meta-analysis of eLearning technology acceptance studies (Šumak, HeričKo, & PušNik, 2011) identified that TAM is indeed the most-used acceptance theory in the specific context, but more importantly that the size of the causal effects between individual TAM-related factors depends on the type of e-learning technology.

Mobile Technology

Advances in mobile technology have led to increased interest in exploring adoption intentions and acceptance of services in this domain. Lu, Yao, and Yu (2005) modified TAM to explore the adoption of wireless internet services via mobile technology, and found strong causal relationships between social influences, personal innovativeness and perceptual beliefs—usefulness and ease of use, which in turn impact adoption intentions. A new model has been proposed by Nysveen, Pedersen, and Thorbjørnsen (2005), integrating the motives that are revealed in information systems theories, uses and gratification theory, and domestication theory and examining four mobile services, namely text messaging, contact, payment, and gaming. The model includes the motivational influences of usefulness, ease of use, enjoyment, and expressiveness, attitude towards using the mobile services, normative pressure as a social influence, and behavioural control reflecting resource-related influences such as the user's economy, experience and skills in using a service. The results indicate that attitude towards using the service is moderated by enjoyment, usefulness, and ease of use, while a user's intention to use the service is moderated by attitude towards the service, expressiveness, enjoyment, usefulness, ease of use, normative pressure and behavioural control. Taking into account TAM, as well as other models extending it for e-commerce acceptance, Fang, Chan, Brzezinski, and Xu (2005) propose a new model focusing on mobile commerce identifying the moderating effects of task type on technology acceptance. A study was carried out to test the proposed model, and the results highlight that perceived usefulness and perceived ease of use were important for user intention to perform general tasks that do not involve transactions and gaming on wireless handheld devices, while perceptions of playfulness influence user intention to play games using wireless technology, and user intention to transact on handheld devices is affected by perceived usefulness and perceived security.

The role of context in the user acceptance of mobile systems was highlighted by Mallat, Rossi, Tuunainen, and Öörni (2009) in the application domain of mobile ticketing systems. The results of the study indicated that the context of use has an important effect on intention to use the mobile service, as well as a mediating effect of perceived usability on user intention, while other decision factors, such as ease of use and compatibility, had a direct effect. Considering the mobility context, Zarmpou, Saprikis, Markos, and Vlachopoulou (2012) extended TAM and introduced the concept of relationship drivers as those dimensions that create a relationship between the consumers and the m-services, including for instance the time and location personalization of m-services, their adaptation to the consumers' profile, the consumers' dynamic permission option and the consumers' reward by the use of the m-services. Testing the model highlighted that relationship drivers have an important effect on perceived usefulness and behavioural intention.

Health Technology

Although the success of health Information Technology (IT) certainly goes beyond user (patient or health professional) acceptance, increasing interest in this application domain has raised the importance of theories that predict and explain health IT acceptance and use (Holden & Karsh, 2010). Such theories are based on existing models, such as TAM, while findings of reviews and meta-studies highlight that TAM predicts a substantial portion of user acceptance of health IT, however several additions and modifications have been proposed (Holden & Karsh, 2010).

An alternative approach to extending TAM aimed at identifying barriers to health IT adoption instead of extending it with determinants positively influencing acceptance (Yarbrough & Smith, 2007). To this end, the following barriers have been identified: interruption of traditional practice patterns, lack of evidence regarding the benefits of IT, organizational issues, as well as system-specific issues such as reliability and dependency. An extended TAM model for health IT acceptance suggested information quality and enabling factors as second order constructs which affect perceived usefulness and perceived ease of use (Moores, 2012). In the proposed model, information quality is posited to be determined by accuracy, content, format and timeliness, while computing support and self-efficacy constitute enabling factors. The results of a study carried out to test the model highlight that the quality of the information provided by the system and the extent to which the user feels they have the technical support or skills to make use of the system are both significant. With a focus on attributes of the individual that have an impact on health IT acceptance, IT feature demands and IT knowledge have been proposed as additional TAM constructs, while the physician's specialty has been studied as a moderator (Melas, Zampetakis, Dimopoulou, & Moustakis, 2011). The individual's technological attitude has also been explored with regard to technology acceptance in a study focusing on mobile electronic medical record adoption by nurses (Kuo, Liu, & Ma, 2013), emphasizing the importance of optimism on perceived usefulness and the impact of optimism, innovativeness, insecurity and discomfort on perceived ease of use.

Ambient Assisted Living and Ubiquitous Computing

Assistive technology and robotics is another technological advancement that has led to further exploration of technology acceptance and extensions of existing models. Heerink, Kröse, Evers, and Wielinga (2010) proposed the Almere model (Figure 14), an extension of the UTAUT model, considering the effect of perceived enjoyment, social presence, perceived sociability, trust, and perceived adaptivity. Perceived adaptivity refers to the capability of the system to change over time in order to support the changing conditions and needs of its users. Testing the proposed model identified among others that perceived adaptivity directly affects user attitude and perceived usefulness, perceived sociability affects

perceived enjoyment and social presence, while intention to use is directly influenced by social influences, attitude, perceived usefulness and ease of use, as well as perceived enjoyment.
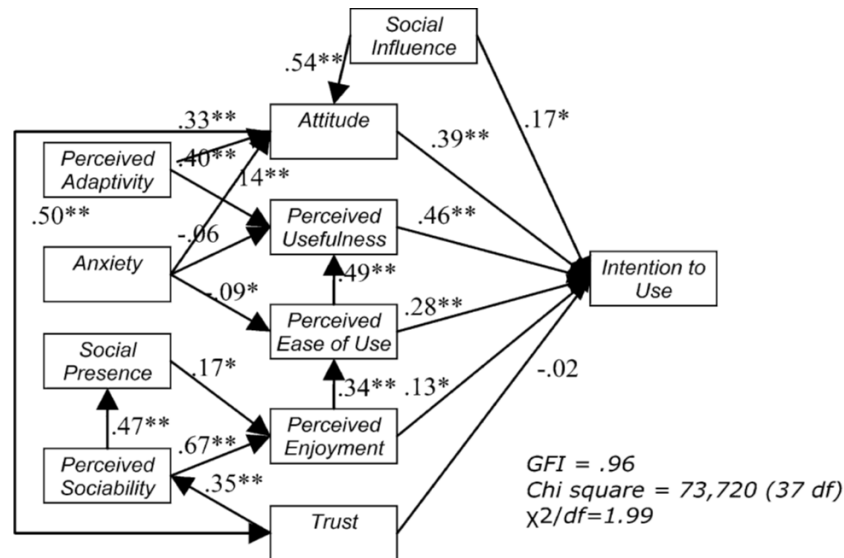
Figure 14. The Almere model

The Ubiquitous Computing Acceptance Model (Shin, 2010) has been proposed to predict whether potential users will accept ubiquitous computing, by studying the relationships among trust, security, privacy, usefulness, ease of use and intention to use a ubiquitous computing technology. In the domain of ubiquitous computing, the Pervasive Technology Acceptance Model (PTAM) (Connelly, 2007) has extended TAM by adding the constructs of trust and integration as direct determinants of behavioural intention, while it adds usage motivation, socioeconomic status, age, gender, and expertise as moderators. Trust is examined in terms of keeping the information collected about the individual as confidential and in terms of trusting the application to behave as expected, given its potential to tailor its behaviour. Integration refers to how well the technology is integrated into the individual's life (e.g., by not distracting them or interfering with their other activities).

The aforementioned research efforts and studies constitute only a part of the literature, which abounds with studies of users' acceptance in wide a variety of domains, such as e-logistics (Tung, Chang, & Chou, 2008), online tax system (Wu & Chen, 2005), hotel office front systems (Kim, Lee, & Law, 2008), Enterprise Resource Planning (ERP) systems (Bueno & Salmeron, 2008), electronic mediated commerce using interactive television (Yu, Ha, Choi, & Rho, 2005), Personal Digital Assistant (PDA) acceptance by healthcare professionals (Mun, Jackson, Park, & Probst, 2006), RFID technology (Hossain & Prybutok, 2008). As already mentioned, this section has reported on studies pertaining to the most major

technological domains (e.g., web, mobile) or domains more relevant to the content of this thesis (e.g., ambient assisted living, ubiquitous computing), focusing on presenting studies involving new parameters or antecedents to existing models and not on providing an exhaustive list of all studies relevant to user acceptance.

It should be noted the majority of the models that have been developed have been mainly tested in the United States and in Canada. However, when tested in other countries, these models have been found to be less predictive (Bandyopadhyay & Fraccastoro, 2007). McCoy, Galetta, and King (2007) tested TAM across several cultures and concluded that it does not hold for certain cultural orientations, and specifically in cultures with low uncertainty avoidance, high masculinity, high power distance and high collectivism. Im, Hong, and Kang (2011) compared the UTAUT model for testing two technologies, the MP3 player and Internet banking, in two different cultures, namely Korea and United States (US) The results reveal that the effects of effort expectancy on behavioural intention and the effects of behavioural intention on use behaviour were greater in the US sample, suggesting that when deciding to adopt a technology, US users seem to take more into account its fundamental characteristics and ease of use than Korean users do, as well as that users in Korea are more influenced by factors other than their own use intentions (e.g., popularity of the technology, trends, social groups). Three models, namely TPB, TAM, and IDT, have been used to analyse technology acceptance of computer technology by Arabian workers and compared to results of these models in Western countries (Hu, Al-Gahtani, & Hu, 2014). The results suggest that perceived behavioural control and subjective norms constitute more important acceptance determinants than attitude, while both perceived usefulness and perceived ease of use remain significant determinants of attitude and intention, however their total effects are comparable in magnitude and statistical significance. Srite and Karahanna (2006) studied the role of espoused cultural values in technology acceptance and found that social norms are stronger determinants of intended behaviour for individuals who espouse feminine and high uncertainty avoidance cultural values. In the context of mobile recommender systems' acceptance, a study has been carried out in three countries, namely China, South Korea, and the United Kingdom (Choi, Lee, Sajjad, & Lee, 2014). The results indicated that two cultural values, collectivism and uncertainty avoidance, moderated the relationships between belief factors and attitudes towards the technology, which may explain why some technologies diffuse more quickly in some countries in which collectivism is more highly valued. Cultural dimensions have also been studied as a moderator of acceptance of mobile banking (Baptista & Oliveira, 2015) indicating that collectivism, uncertainty avoidance, long term cultural values and power distance had an important moderating effect on use behaviour.

3.1.8 TECHNOLOGY ACCEPTANCE AND AMBIENT INTELLIGENCE

In summary, research in the direction of technology acceptance has led to the aggregation of a considerable number of parameters that can be considered as important towards predicting the acceptance of a given technology by its target audience. An important concern is how to practically employ these models in the context of assessing Ambient Intelligence technologies. As AmI environments are equipped with various sensors and monitoring capabilities, privacy and trust become issues of paramount importance for their inhabitants (Cook et al., 2009), while technology acceptance needs to be studied from a new perspective. Ambient Intelligence may be found in any potential daily living environment, such as home, workplace, health care, educational setting, or public space (Friedewald, Vildjiounaite, Punie, & Wright, 2007) embracing any activity carried out in these environments. Therefore, the parameters that may impact user acceptance of an Ambient Intelligence environment definitely extend beyond the parameters suggested in the first models studying computer acceptance in workplace environments.

Towards studying acceptance in AmI environments, a short review of the initial technology acceptance models and their evolution, as well as their adaptations to address different contexts of use has been carried out. Moreover, a classification of the parameters studied in these models is provided, with the aim to assist researchers in identifying parameters that should be included in studying user acceptance of AmI environments, according to the target environment and context of use. Attributes that can be used for this classification include:

- Category of reference: if the metric is used to describe an attribute of the individual, of the social environment, the system under evaluation, or its impact on the individual
- The objectiveness of the metric (subjective or objective)
- Assessment method: which method is employed to find out the value of the specific metric (e.g., questionnaire, observation, automated system measurement)
- The context in which the specific metric can be applied (e.g., workplace, education, health, home environment, public environments)
- Models which include the specific metric.

The tables that follow list all the metrics identified in literature, as follows: Table 2 lists all system-related parameters, Table 3 refers to parameters concerning the individual (user),

Table 4 encompasses attributes describing social influences as well as environment factors, and

Table 5 features parameters describing system impact on the individual.

Table 2. Parameters referring to the system: (S) stands for subjective and (O) for objective method

Metric	Method	Context	References
Perceived usefulness	Questionnaire (S)	All	Davis (1985); Venkatesh & Davis (2000)
Perceived ease of use	Questionnaire (S)	All	Davis (1985); Venkatesh & Davis (2000)
Objective usability	Keystroke model (O)	Computer Software	Venkatesh & Davis (1996)
Complexity	Questionnaire (S)	Consumer product Organizational setting	Taylor & Todd (1995); Thompson et al. (1991); Moore & Benbasat (1991); Rogers (1995)
Functionality	Questionnaire (S)	eLearning (providing access to instructional and assessment media) Mobile services	Pituch & Lee (2006); Zarmpou et al. (2012)
Output quality	Questionnaire (S)	Organizational setting eLearning (information quality, service quality, system quality) Health IT	Davis (1985); Venkatesh & Davis (2000); Roca et al. (2006); Yarbrough & Smith (2007)
Trialability Divisibility	Questionnaire (S)	Organizational setting	Moore & Benbasat (1991); Rogers (1995)
Perceived enjoyment	Questionnaire (S)	Organizational setting WWW usage	Davis et al. (1992); Moon & Kim (2001)
Accessibility	Questionnaire (S)	Electronic communication media	Karahanna & Straub (1999); Karahanna & Limayem (2000)
Perceived adaptivity	Questionnaire (S)	Assistive technology and robotics Mobile services	Heerink et al. (2010); Zarmpou et al. (2012)
Personalization	Questionnaire (S)	Mobile services	Zarmpou et al. (2012)
Response time	Questionnaire (S)	eLearning	Pituch & Lee (2006)

Interactivity	Questionnaire (S)	eLearning (interaction between teachers and students, and students themselves)	Pituch & Lee (2006)
Social presence	Questionnaire (S)	Electronic communication media Assistive technology and robotics	Karahanna & Straub (1999); Heerink et al. (2010)
Perceived sociability	Questionnaire (S)	Assistive technology and robotics	Heerink et al. (2010)
Enabling expressiveness (of emotions, social or personal identity)	Questionnaire (S)	Mobile services	Nysveen et al. (2005)
Communication	Questionnaire (S)	Virtual worlds	Fetscherin & Lattemann (2008)
Collaboration, Cooperation	Questionnaire (S)	Virtual worlds	Fetscherin & Lattemann (2008)
Perceived security	Questionnaire (S)	Online trading system eCommerce Ubiquitous Computing	Carlos Roca et al. (2009); Ha & Stoel (2009); Shin (2010)
Privacy	Questionnaire (S)	Ubiquitous Computing	Shin (2010)
Perceived risk	Questionnaire (S)	eCommerce Internet banking e-services adoption	Pavlou (2003); Lee (2009); Featherman & Pavlou (2003)
Web site design	Questionnaire (S)	eCommerce (e-shopping Quality)	Ha & Stoel (2009)
Atmospheric / experiential quality	Questionnaire (S)	eCommerce (e-shopping Quality)	Ha & Stoel (2009)
Customer service	Questionnaire (S)	eCommerce (e-shopping Quality)	Ha & Stoel (2009)
Reputation	Questionnaire (S)	eCommerce (extended TAM)	Pavlou (2003)
Reliability	Focus group (S)	Health IT	Yarbrough & Smith (2007)
Cost	Questionnaire (S) Data analysis (O)	Household	Tornatzky & Klein (1982); Brown & Venkatesh (2005)
Price Value	Questionnaire (S)	Mobile Internet technology	Venkatesh et al. (2012)

Transfer of learnt skills	Questionnaire (S)	Serious games	Yusoff et al. (2010)
Learner control	Questionnaire (S)	Serious games	Yusoff et al. (2010)
Reward	Questionnaire (S)	Serious games Mobile services	Yusoff et al. (2010); Zarmpou et al. (2012)
Situated and authentic learning	Questionnaire (S)	Serious games	Yusoff et al. (2010)
External constraints (PC & environment characteristics)	Questionnaire (S)	Household	Brown & Venkatesh (2005)
Dependency	Focus group (S)	Health IT	Yarbrough & Smith (2007)
Compatibility	Questionnaire (S)	Consumer product Organizational setting	Taylor & Todd (1995); Moore & Benbasat (1991); Rogers (1995)
End-user support	Questionnaire (S)	Organizational setting Health IT	Lee et al. (2003) Moores (2012)
Integration	Questionnaire (S)	Ubiquitous Computing	Connelly (2007)

Table 3. Parameters referring to the individual: (S) stands for subjective and (O) for objective method

Metric	Method	Context	References
Experience / Self-efficacy / Digital literacy / IT knowledge	Questionnaire (S) System logs (cookies, WAM) (O)	Organizational setting Mobile Internet technology Consumer product WWW E-commerce eLearning Health IT Household Ubiquitous Computing	Davis (1985); Venkatesh & Davis (2000); Venkatesh et al. (2012); Taylor & Todd (1995); Compeau et al. (1999); Castañeda et al. (2007); Hernández et al. (2010); Park (2009); Lin (2011); Mohammadyari & Singh (2015); Melas et al. (2011); Brown & Venkatesh (2005); Connelly (2007)

Affect / Computer attitude / Computer anxiety / Technology anxiety / Anxiety towards the system / Technological attitude	Questionnaire (S)	Organizational setting Self-service technology Assistive technology and robotics Health IT	Thompson et al. (1991); Compeau et al. (1999); Meuter et al. (2003); Heerink et al. (2010); Kuo et al. (2013)
Effort expectancy	Questionnaire (S)	Organizational setting	Venkatesh et al. (2003)
Outcome expectations (performance & personal)	Questionnaire (S)	Organizational setting	Compeau et al. (1999); Venkatesh et al. (2003)
Hedonic motivation Hedonic outcomes	Questionnaire (S)	Mobile Internet technology	Venkatesh et al. (2012); Brown & Venkatesh (2005)
Trust	Questionnaire (S)	eCommerce Online trading systems Assistive technology and robotics Mobile services Ubiquitous Computing	Gefen et al. (2003); Pavlou (2003); Carlos Roca et al. (2009); Heerink et al. (2010); Zarmpou et al. (2012); Shin (2010)
Personal innovativeness	Questionnaire (S)	WWW Mobile services mLearning	Agarwal & Karahanna (2000); Lu et al. (2005); Zarmpou et al. (2012); Liu et al. (2010)
Cognitive playfulness	Questionnaire (S)	WWW	Agarwal & Karahanna (2000); Moon & Kim (2001)
Age	Questionnaire (S)	Organizational setting Mobile Internet Ubiquitous Computing	Venkatesh et al. (2003); Venkatesh et al. (2012); Connelly (2007)
Gender	Questionnaire (S)	Organizational setting Mobile Internet Ubiquitous Computing	Venkatesh et al. (2003); Venkatesh et al. (2012); Connelly (2007)

Education	Questionnaire (S)	Internet usage	Porter & Donthu (2006)
Marital status **Child's age**	Questionnaire (S)	Household	Brown & Venkatesh (2005)
Income **Socioeconomic status**	Questionnaire (O)	Household Ubiquitous Computing	Brown & Venkatesh (2005); Connelly (2007)
Habit	Questionnaire (S)	Organizational setting Mobile Internet	Thompson et al. (1991); Venkatesh et al. (2012)

Table 4. Parameters referring to social influences and influence of the environment: (S) stands for subjective method

Metric	Method	Context	References
Observability **Result demonstrability** **Communicability**	Questionnaire (S)	Organizational setting	Moore & Benbasat (1991); Rogers (1995); Davis (1985); Venkatesh & Davis (2000)
Image **Social approval** **Social outcomes**	Questionnaire (S)	Organizational setting Household	Davis (1985); Venkatesh & Davis (2000); Moore & Benbasat (1991); Tornatzky & Klein (1982); Brown & Venkatesh (2005)
Social Factors **Subjective norm** **Normative influences** **Normative pressure** **Social influence** **Social norm**	Questionnaire (S)	Organizational setting Consumer product Household Online game	Davis (1985); Venkatesh & Davis (2000); Taylor & Todd (1995); Thompson et al. (1991); Venkatesh et al. (2003); Brown & Venkatesh (2005); Hsu & Lu (2004)
Voluntariness	Questionnaire (S)	Organizational setting	Davis (1985); Venkatesh & Davis (2000); Moore & Benbasat (1991); Venkatesh et al. (2003)
Management support	Questionnaire (S)	Organizational setting	Ibgaria et al. (1997)

Facilitating conditions Perceptions of external control	Questionnaire (S)	Consumer product Organizational setting	Taylor & Todd (1995); Thompson et al (1991); Venkatesh et al. (2003); Venkatesh & Bala (2008)
Cultural dimensions	Questionnaire (S)	Mobile recommender systems Mobile banking Usage of Personal Computers	Choi et al. (2014); Baptista & Oliveira (2015); Srite & Karahanna (2006)
Critical mass	Questionnaire (S)	Online game	Hsu & Lu (2004)
Context of use	Questionnaire (S)	Mobile ticketing (TAM and IDT extension)	Mallat et al. (2009)

Table 5. Parameters referring to the impact of the system to the individual: (S) stands for subjective method

Metric	Method	Context	References
Job relevance Job fit	Questionnaire (S)	Organizational setting Organizational setting	Davis (1985); Venkatesh & Davis (2000); Thompson et al. (1991)
Relative Advantage	Questionnaire (S)	Consumer product Organizational setting	Taylor & Todd (1995); Moore & Benbasat (1991); Rogers (1995)
Outcome expectations (performance & personal)	Questionnaire (S)	Organizational setting	Compeau et al. (1999); Venkatesh et al. (2003)
Long-term consequences of use Usefulness Utilitarian outcomes Perceived benefit	Questionnaire (S)	Organizational setting (MPCU) mLearning Internet banking Household Ubiquitous Computing	Thompson et al. (1991); Liu et al. (2010); Lee (2009); Brown & Venkatesh (2005); Shin (2010)
Flow experience	Questionnaire (S)	Online game	Hsu & Lu (2004)
Cognitive absorption	Questionnaire (S)	E-learning	Saadé & Bahli (2005)

Interruption of traditional practice patterns	Questionnaire (S)	Health IT	Yarbrough & Smith (2007)

The purpose of the current review was to emphasize the plethora of parameters that should be taken into account, especially in AmI environments, due to their technological complexity and diversity in context of use. As a result, the review has included studies mostly relevant to AmI and studies of major everyday life domains, with a focus on those that have introduced new constructs in acceptance models. Indeed, the presented review and classification has resulted in 71 parameters of technology acceptance that act as direct determinants, antecedents or moderators of technology acceptance. Also, it is noteworthy that the overwhelming majority of these parameters is assessed in the various studies through questionnaires, asking users to self-report their characteristics, attitudes and perceptions.

Although the self-reporting approach is inevitable in many cases, and the only possible method when the first studies were carried out, this is no longer an ideal solution in the context of AmI environments. On the one hand, the number of questions to be asked to the user may become unmanageable in such environments, if all the relevant aspects are to be assessed. On the other hand, an AmI environment has the capability to provide measurements through its sensors that will reduce the number of questions that need to be asked to the user. The vision of AmI can bring about new perspectives to technology acceptance and evaluation, facilitating not only the environment in adapting itself to better serve the needs of the user, but also evaluators aiming to assess the overall user acceptance of such environments. This potential highlights the need for a user acceptance evaluation model in AmI environments, aiming to assess a wide range of characteristics and qualities of such environments, taking into account traditional and modern models and evaluation approaches.

3.2   USABILITY EVALUATION

Several definitions for usability have been provided in the HCI literature. ISO (ISO/IEC 9241-11, 1998) defines usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. Another fundamental definition was provided by Jakob Nielsen (1994b), who suggested that usability is a quality attribute that assesses how easy a user interface is to use and is defined by five quality components:

- Learnability, which measures how easy it is for users to accomplish basic tasks the first time they encounter the design.

- Efficiency, studying how quickly users can perform tasks once they have learned the design.
- Memorability, which refers to how easily users can re-establish proficiency when they return to the design after a period of not using it.
- Errors, referring to how many errors users make, how severe are these errors, and how easily they can recover from the errors.
- Satisfaction, reviewing how pleasant it is to use the design.

Since usability is an important component of an AmI technology and is critical for the overall experience a user has with a technology, as well as for its overall acceptance, the following sections review the most important usability evaluation methods and present frameworks that have been proposed towards the evaluation of usability. Two categories of tools, directly relevant to the most popular usability evaluation methods of usability and user testing are also studied in details: tools for working with guidelines, as well as tools supporting automated measurements.

3.2.1   USABILITY EVALUATION METHODS

Since the very first definitions of usability until now, several methods have been proposed aiming to assess the usability of a specific product or service. This section reviews the most important methods. Table 6 below provides an overview of the most popular usability evaluation methods.

Table 6. Summary of usability evaluation methods

A/B testing	A technique which allows the comparison of two different versions of a design to verify which one performs statistically better against a predetermined goal (Hanington & Martin, 2012)
Affinity diagramming, KJ technique	It can be used in the context of usability testing in the laboratory to create clusters of user observations, by having the team watching the experiment in the observation room write down the articulated user observations on sticky notes and posting them to a whiteboard, organized in categories. Typically, the categories that have many usability issues will include the largest number of post-it notes (Hanington & Martin, 2012). In the context of iterative evaluation and design, the method can be applied to reach objective group consensus out of a collection of subjective, opinionated data (Spool, 2004).
Card sorting	A method used to evaluate the information architecture of an application, in which users are given a set of cards to group together (Spencer, 2009). Open card sorting: users are asked to group the cards as it makes sense to them and once they are done,

	they are asked to name the groups. Closed card sorting: participants are provided with specific group categories.
Coaching	During a coaching study, the test user who is working with the system under evaluation is allowed to ask any system-related question of an expert coach who will answer to the best of his or her ability (Nielsen, 1994b).
Cognitive walkthrough	One or more evaluators work through a series of tasks and ask themselves a set of questions from the perspective of the user, aiming to assess the system's learnability for new or infrequent users (Wharton, 1994).
Competitive Testing	Competitive usability evaluations are a method to determine how a system performs in relation to competitors. The comparison can be holistic (e.g., ranking by overall usability metrics), or it can be more focused (e.g., comparing features, content, or design elements) (Schade, 2013).
Consistency inspection	Experts review products or projects to ensure consistency across multiple products (Nielsen, 1994c).
Constructive Interaction / Co-Discovery Learning	It involves two test users using a system together, enhancing the naturalness of verbalizing their thoughts (Nielsen, 1994b).
Co-operative evaluation	An end-user and a developer form the evaluation team and together they explore a prototype and develop a critique (Muller, Haslwanter, & Dayton, 1997).
Diaries	A method of understanding participant behaviour and intent. The method attempts to manage the gap between these two by having participants record events as they happen, by answering predefined questions about events (feedback studies) or capturing media that are then used as prompts for discussion in interviews (elicitation studies) (Carter & Mankoff, 2005).
Eye tracking	A promising technique that can be used in usability evaluation, whereby the user's eye movements are measured so that the researcher knows where a person is looking at a given time and the sequence in which the person's eyes are shifting from one location to another (Poole & Ball, 2006)
Feature inspection	Expert evaluators check the interface and list the sequence of features used to accomplish typical tasks, check for long sequences, cumbersome steps, steps that would not be natural for users to try, and steps that require extensive knowledge/experience in order to assess a proposed feature set (Nielsen, 1994c).
Focus groups	A focus group brings together a cross-section of stakeholders in a discussion group format, while views are elicited by a facilitator on relevant topics (Caplan, 1990).

Formal usability inspection	It uses a six-step procedure to combine heuristic evaluation and a simplified form of cognitive walkthrough (Nielsen, 1994c).
GOMS models	A model-based evaluation method, in which the model describes the knowledge of procedures that user must have in order to operate a system. Constructing a GOMS model involves writing out the methods for accomplishing the task goals of interest, and then calculating predicted usability metrics from the method representation (Kieras, 2009). There are different forms of GOMS models that represent the methods at different levels of detail, and whose calculations can range in complexity from simple hand calculations to full-fledged simulations.
Guidelines inspection	An inspection method using published guidelines, which provide evaluators with specific recommendations about the design of an interface (Jeffries, Miller, Wharton, & Uyeda, 1991).
Heuristic evaluation	Heuristic evaluation involves having a small set of evaluators examine the interface and judge its compliance with recognized usability principles (the "heuristics") (Nielsen, 1994a).
Heuristic walkthrough	This method combines the methods of heuristic evaluation, cognitive walkthrough, and pluralistic usability walkthroughs. Evaluators make two passes through a product: one which uses "thought-provoking" questions and requires the evaluators to work through a set of prioritized tasks, and one which requires evaluators to use a set of heuristics to find additional problems (Sears, 1997).
Interviews	A valuable method for exploratory user research that involves one-on-one discussions that help researchers learn about users' attitudes and beliefs regarding a specific system (Nielsen, 2010). Interviews may be structured, un-structured or semi-structured.
Logging actual use	It involves having the computer automatically collect statistics about the detailed use of a system (Nielsen, 1994b).
Observation	It involves observing users as they work with a system (Nielsen, 1994b).
Performance measurement	User performance is measured by having a group of test users perform a predefined set of test tasks while collecting performance metrics (e.g., time and error data) (Nielsen, 1994b).
Persona-based inspection	Personas are descriptions of fictitious users, users who do not actually exist, but are created based on relevant information from potential and real users, and are described in such a way that the reader can believe that the user could exist in reality (Nielsen, 2012). They can be the source of different perspectives and a persona-based

	evaluation would be a variation on the perspective-based inspection, as a guide to how users might interact with a product (Wilson, 2013).
Perspective-based inspection	Each inspection focuses on a subset of usability issues to check, and a specific procedure for conducting the inspection. The inspectors are given the description of the perspectives to focus on, a list of user tasks, a set of questions related to the perspective and a list of heuristics related to the perspective (Zhang, Basili, & Shneiderman, 1999).
Pluralistic walkthrough	A group of usability experts, users, and product developers, review a user interface design by following a task scenario and examining each element of interaction by posing a set of given questions (Bias, 1994).
Question-asking protocol	A variation of the user testing method, during which test participants are prompted by the experimenter to answer specific questions, in order to gain insight to their mental model and where they have trouble in understanding and using the system (Fernandez, Insfran, & Abrahão, 2011).
Questionnaire	Evaluation questionnaires aim to study users' opinions and more specifically how users use systems and what features they particularly like or dislike (Nielsen, 1994b).
Remote Evaluation	A situation where the evaluators are separated in space and/or time from users and may use synchronous and asynchronous methods and may involve end users or experts (Andreasen, Nielsen, Schrøder, & Stage, 2007).
Retrospective testing	If a videotape has been made of a user test session, users review the recording, allowing thus experimenters to collect additional information by the users' comments while reviewing the tape (Nielsen, 1994b).
RITE	RITE (Rapid Iterative Testing and Evaluation) involves user testing with representative users and differs from traditional usability testing by emphasizing extremely rapid changes, as soon as a problem has been verified (even after one single participant) and verification of the effectiveness of these changes (Medlock, Wixon, Terrano, Romero, & Fulton, 2002).
Shadowing	It is a qualitative research technique, where the researcher accompanies the user and observes how they use the product or service within their natural environment, without interfering with them (Interaction Design Foundation, 2017).
Standards inspection	An expert on some interface standard inspects the interface for compliance (Nielsen, 1994c).
Surveys	Surveys are defined as compilations of questions that are implemented via a computer or paper-and-pencil-based environment, that either have quantitative or

	qualitative scales, or are open-ended, and that target at extracting a variety of information from a representative sample of the target population (which is in most cases current or prospective users of a system being evaluated) (Ozok, 2009). Typically, surveys are administered through questionnaires, but their main difference is that in surveys data are gathered and statistically analysed towards reaching specific conclusions, whereas questionnaires constitute the means of collecting data.
Task network models	In task network models, task performance is modelled in terms of a PERT-chart-like network of processes. Each process starts when its prerequisite processes have been completed, and has an assumed distribution of completion times. This basic model can be augmented with arbitrary computations to determine the completion time, as well as what its symbolic or numeric inputs and outputs should be (Kieras, 2009).
Teaching	The test participant, after becoming familiar with the system, demonstrates it to a seemingly her user (a confederate) and describes how to accomplish certain tasks (Vora & Helander, 1995).
Thinking aloud	Users are asked to think aloud, i.e. to verbalize their thoughts, during a user testing experiment and as they move through the interface (Nielsen, 2012).
User testing	User testing with real users is the most fundamental usability evaluation method and is in some sense irreplaceable, since it provides direct information about how people use computers and what their exact problems are with the product being tested (Nielsen, 1994b). Testing can be carried out on mock-ups, a system prototype, or the final product and involves observing users while carrying out tasks with the system.
Web analytics	In the case of websites, web analytics refer to the objective tracking, collection, measurement, reporting and analysis of quantitative Internet data to optimize websites and web marketing initiatives (Kaushik, 2007)
Wizard of Oz	Studies where participants are told that they are interacting with a computer system, though in fact they are not. Instead the interaction is mediated by a human operator (Dahlbäck, Jönsson, & Ahrenberg, 1993).

It is evident that there is a plethora of available evaluation methods, and that new methods emerge over time. Specific attributes and characteristics of each method can be employed for their classification, by taking into account:

- *who* is involved (users, experts, both, system for automated evaluations)
- *where* the method can be carried out (in the laboratory, in the field / in situ)

- *when* the method is applied (formative vs. summative)
- *the type of the method* (empirical, analytical, or inspection)
- *objectiveness* of the method (objective or subjective), as well as
- *the type of results* produced (qualitative or quantitative).

Beyond the methods described in Table 6, as the most prevalent approaches towards usability evaluation, literature thrives with evaluation studies in numerous contexts. In an effort to summarize the approaches that have been applied and to identify the challenges that remain yet to be addressed, meta-reviews of usability studies have focused on how usability is measured in current practices. Hartson, Andre, and Williges (2001) identified that techniques for evaluating and comparing the effectiveness of usability evaluation methods (UEM) are not well established. The considerably large number of 2.116 studies was reviewed by Martins, Queirós, Silva, and Rocha (2014) with the aim to identify, analyse, and classify the methodologies and methods used in the literature for the evaluation of IT systems and technologies. The review pointed out that the most commonly employed methodology was that of inquiry, followed by test, inspection, and controlled experiments. Paz and Pow-Sang (2014) carried out a review of 274 usability studies and reported that the most commonly used methods were usability tests, questionnaires, and heuristic evaluations.

Hornbæk (2006) carried out a thorough review of current practice regarding how usability is measured, by categorizing and discussing usability measures from 180 studies published in core HCI journals and proceedings. In summary, the problems that have been identified include the following: (1) measures of the quality of interaction, for example assessed by domain experts, are used only in a few studies; (2) approximately one quarter of the studies do not assess the outcome of the users' interaction, leaving unsupported any broad claims about usability; (3) measures of learning and retention of how to use an interface are rarely employed, despite being recommended in prominent textbooks; (4) some studies treat measures of how users interact with interfaces as being synonymous with quality-in-use despite an unclear, if not weak, relation between usage patterns and quality-in-use; (5) measures of users' satisfaction with interfaces are in a disarray and most studies reinvented questions to be asked users, ignoring validated questionnaires readily available; and (6) some studies mix together, perhaps even consider synonymous, users' perceptions of phenomena with objective measures of those phenomena. Additionally, the challenges identified are to distinguish and empirically compare subjective and objective measures of usability; to focus on developing and employing measures of learning and retention; to study long-term use and usability; to extend measures of satisfaction beyond post-use questionnaires; to validate and standardize the host of subjective satisfaction questionnaires used; to study correlations

between usability measures as a means for validation; and to use both micro and macro tasks and corresponding measures of usability.

An interesting observation is that a combination of test and inquiry methodologies is commonly found in usability evaluations, probably due to their complementary nature (Martins et al., 2014). The need for combining usability evaluation methods had also been highlighted by Nielsen (1994b) who identified that usability evaluation methods should be combined to achieve better results, as for instance user testing and heuristic evaluation can be alternated during the evaluation of a system since they have been shown to find fairly distinct usability problems. Also, interviews that can be used to gain insight on specific points identified via questionnaires mailed to a large number of users. The flexibility of method use, involving a variety of methods and adapting them during a usability evaluation, has also been pointed out by a survey with 155 usability practitioners on the analysis of their latest usability evaluation (Følstad, Law, & Hornbæk, 2012). The survey results highlighted also the need for a taxonomy of method components and an identification of their strengths and weaknesses for typical evaluation contexts.

With the aim to address the widely recognized need for a classification of UEMs, an analysis of 23 usability engineering methods in various contexts is attempted in Gulati and Dubey (2012), based on the criteria of immediacy of response, intrusiveness, expensiveness, location, development stages that the method can be applied, usability issues covered (effectiveness, efficiency, satisfaction), as well advantages and disadvantages. From a different point of view, recognizing that UEMs beyond their practical application in the IT industry are an active research topic, Freiberg and Baumeister (2008) studied how these methods have been applied in PhD and MA thesis in terms of the expertise and number of participants, time required, evaluation goal, and application evaluated. In the context of this research, eight sets of heuristic guidelines used in the studied literature are introduced, as well as a collection of the most frequently applied usability metrics. Hornbæk and Law (2007), identified that literature in HCI offers little help in selecting the appropriate measures of usability. In an effort to address this issue they carried out a meta-analysis of correlations of effectiveness, efficiency and satisfaction measures among usability measures and identified among others that the process of determining what constitutes an error is not well defined, that measures of users' perceptions of phenomena are generally not correlated with objective measures of the phenomena.

In terms of objectivity, but from a different perspective, the impact of the individual evaluator on the evaluation results is another important concern that has been reported in literature (Hertzum & Jacobsen, 2001). More specifically, it is emphasized that different evaluators evaluating the same system with cognitive walkthrough, heuristic evaluation, or think aloud study detect substantially different sets of

usability problems in the system. In the case of expert-based reviews, this is expected and anticipated by the methodologies that suggest the involvement of at least three to five evaluators. However, in observation protocols it would be expected that the evaluation results would not be subject to the evaluators themselves. To this end, Hertzum and Jacobsen (2001) suggest that evaluators should be explicit on goal analysis and task selection, that an extra evaluator should be included, and that a reflection on the evaluation procedures and problem criteria is needed to adjust one's practices and try to polish them.

Hornbæk and Law (2007) also identified that the UX movement has had an impact in the notion of usability, finding however mixed results on how these notions are correlated (i.e., whether UX broadens or narrows the usability notion). Recently, in this respect, Bevan, Carter, and Harker (2015) published a work regarding the revision that the ISO-9241-11:1998 standard is undergoing, in an effort to recapitulate the lessons that have been learnt ever since regarding usability. According to their analysis, the relation between usability and user experience is one of the issues that need to be addressed by current usability definitions. Also, new metrics of usability beyond effectiveness, efficiency and user satisfaction need to be studied. Finally, it should be studied how usability contributes to the avoidance of negative outcomes from using a product or service. Last, the authors clarify the difference between usability and UX, by explaining that usability typically deals with goals shared by a user group, while UX is concerned with individual goals, which can include personal motivations, including needs to acquire new knowledge and skills, to communicate personal identity and to provoke pleasant memories.

The relationship between usability and UX goes far beyond the purpose of this thesis, however it is clear that usability is important and should be considered in any IT system or environment, in any context. It is also evident that there is a large number of usability methods that one should consider, each with specific aims, scope, and benefits. Yet, there is still a challenge as to which method should be employed and how, according to the specific research or practical evaluation purposes. In this context, this thesis proposes an online community for UX practitioners, which can be used as a knowledge-base of current usability evaluation research and practice, acting both as a repository and guidance tool for usability and UX researchers and practitioners.

Finally, a challenge that is often reported in literature is that of the objectivity of recordings and how the individuals who participate in the evaluation have an impact on the outcomes. To address this issue, several approaches have targeted towards automated measurements (Section 3.2.4). Embracing the need for supporting evaluators with objective metrics, and taking advantage of the sensing capabilities of AmI environments, the UXAmI Framework incorporates automated measurements, while the accompanying

UXAmI Observer tool assists evaluators by automatically identifying potential user interaction errors and providing descriptive statistics per user session, user group, and experiment / system evaluated.

3.2.2   USABILITY EVALUATION FRAMEWORKS

An early framework for evaluating user-computer interaction was proposed by Sweeney, Maguire and Shackel (1993), according to which evaluations are categorised based on three dimensions: (i) approach to evaluation, which may be user-, expert- or theory-based; (ii) type of evaluation, which classified evaluations into three basic types, diagnostic, summative and certification; and (iii) time of evaluation, which reflects the temporal location in the product life cycle at which the evaluation is conduced. The framework discusses the three dimensions and their relationship, provides a classification of usability evaluation methods and studies several evaluation methods, analysing their advantages and disadvantages. Another classification approach, studying usability in relevance to HCI and Software Engineering (SE), was proposed by Ferre, Juristo and Moreno (2005). The framework aimed at offering developers who have the objective of integrating usability practices into their software process, a tool that characterizes 35 selected HCI techniques in relation to six relevant criteria from a SE viewpoint, and organizes them according to the kind of activities in the development process where they may be applied, and to the most appropriate time of application in an iterative life cycle. The techniques are organised according to the HCI activity they can be used for (requirements elicitation and analysis, requirements specification, requirements validation, interaction design, and usability evaluation) and are characterised according to the following criteria: user participation, training needs, general applicability, as well as proximity to SE, usability improvement/effort ratio, and representativeness. Furthermore, each technique is mapped to the stage where it is intended to be applied (initial cycles, central cycles, or evolution cycles).

Despite the large number of evaluation methods, only a small number of methods are typically used in usability evaluations, as discussed in the previous section. Hence, the emphasis of evaluation frameworks has mainly shifted towards defining what should be measured rather than how to measure it, which also constitutes the main focus of this section. Moving beyond the notion of usability towards that of UX, the concept of quality has had a pivotal role in models and frameworks. Bevan (2009a) proposed a theoretical framework taking into account quality in use, which is measured by the usability, flexibility, and safety of the product under evaluation. Usability is further analysed in effectiveness, efficiency and user satisfaction, which in turn includes the constructs of likability, pleasure, comfort and trust. Flexibility is determined by three constructs, namely context conformity, context extendibility and accessibility. Finally, safety is further decomposed into commercial damage, operator health and safety, public health and safety, as well as environmental harm.

A framework also taking into account the concept of quality in use is QUIM – Quality in Use Integrated Map (Seffah, Kececi, & Donyae, 2001), which brings together different factors, criteria, metrics and data defined in different HCI and SE models. More specifically, QUIM is a hierarchical model with four levels: factors, criteria, metrics and data (Figure 15). The model includes the following factors: effectiveness, efficiency, satisfaction, productivity, safety, internationability, and accessibility. Criteria of the model include: attractiveness, consistency, minimal action, minimal memory load and completeness. By analysing other models, more than 100 metrics have been identified and integrated into QUIM, organized under the criteria categories.
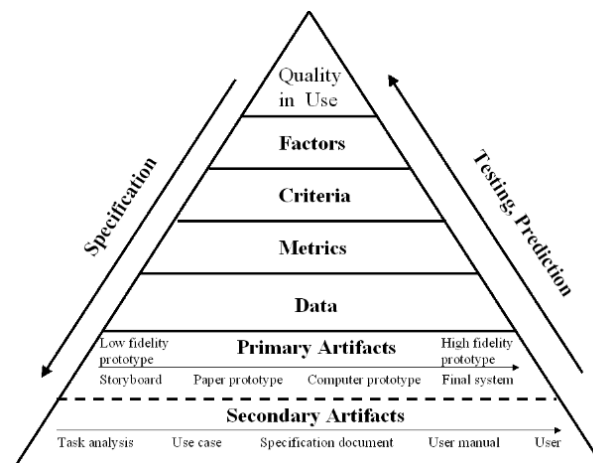
Figure 15. Quality in Use Integrated Map

A consolidated version of the QUIM model (Seffah, Donyaee, Kline, & Padda, 2006) includes ten usability factors, namely: efficiency, effectiveness, productivity (the level of effectiveness achieved in relation to the resources consumed), satisfaction, learnability, safety (whether a software product limits the risk of harm to people or other resources), trustfulness, accessibility, universality, and usefulness (whether a software product enables users to solve real problems in an acceptable way). Factors that are likely to be included in future versions of QUIM are identified to be portability, adaptability and comprehension. Each factor of the consolidated QUIM model is broken down into 26 measurable criteria, while each criterion is directly measurable via at least one specific metric. The consolidated QUIM criteria are the following: time behaviour, resource utilization, attractiveness, likeability, flexibility, minimal action, minimal memory load, operability, user guidance, consistency, self-descriptiveness, feedback, accuracy, completeness, fault tolerance, resource safety, readability, controllability, navigability, simplicity, privacy, security, insurance, familiarity, load time, and appropriateness.

The quality model (Kurosu, 2015) studies the artefact quality and the quality in use and encompasses the concept of usability as part of the artefact quality. Both quality aspects are further analysed into subjective and objective constructs (Figure 16), including the following attributes: usability, functionality, performance, reliability, safety, compatibility, cost, maintainability, attractiveness, productivity, freedom from risk, and meaningfulness.
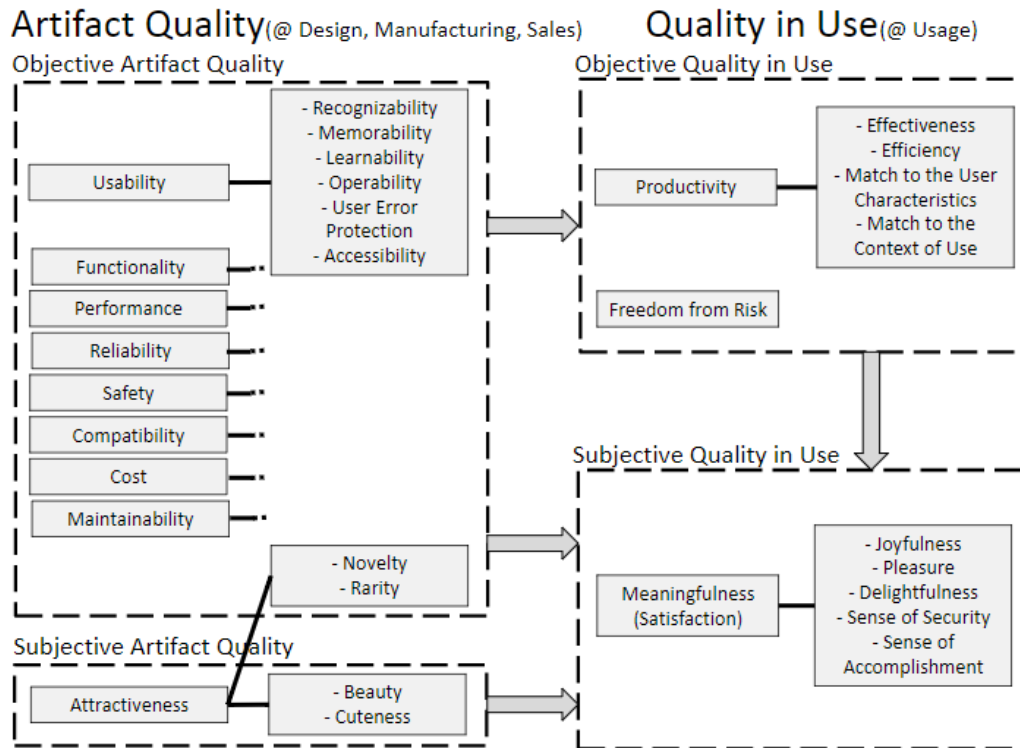
Figure 16. The quality model

An evaluation framework for specifying and measuring the user-orientation (and thus acceptability) of interactive products, emphasizing at the same time the issue of accessibility is proposed by Mourouzis, Antona, Boutsakis, and Stephanidis (2006). According to the framework, user-orientation (and thus system acceptability) is measured by the extent to which: the product is made visible to non-users (visibility), non-users are motivated to gain a personal experience of the system (perceived usefulness & ease of use), actual users find it easy and acceptable to reach the product (availability/approachability), actual users find it useful, easy and acceptable to interact with the product (quality of interaction experience), previous users are motivated to become long term users (relationship maintainability and subjective usefulness & ease of use), product users are not offered more promising and satisfying

alternatives (competitiveness). In addition, beyond the aforementioned product qualities, accessibility is raised as a ubiquitous issue in the user experience lifecycle.

Another framework influenced by the UX movement, taking therefore into account the purchase and continuance intentions of users, is the one proposed by Pu, Chen, and Hu (2011) for the evaluation of recommender systems. According to the framework the following attributes of a recommender system should be evaluated (Figure 17): (i) user perceived qualities, including the quality of recommender items, interaction and interface adequacy (ii) user beliefs as a result of these qualities in terms of ease of use, usefulness and control, (iii) subjective user attitudes, and (iv) the users' behavioural intentions.
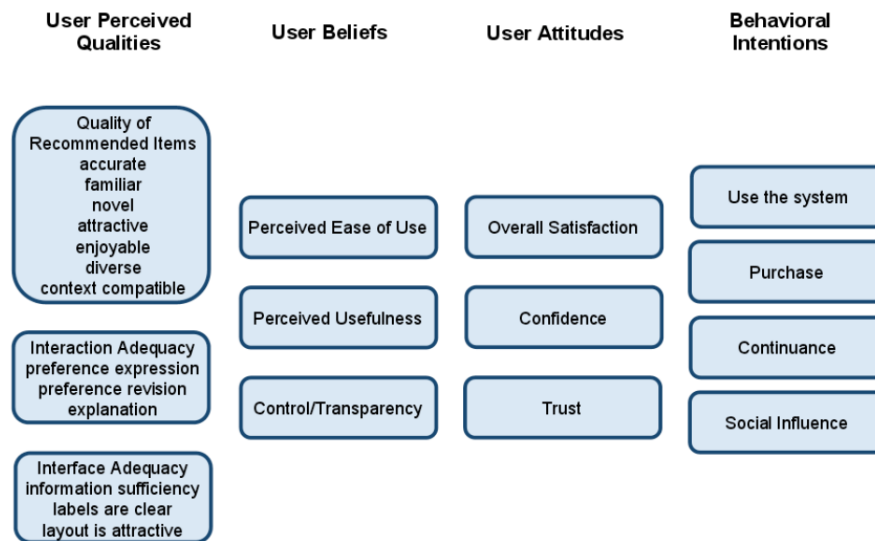
Figure 17. The ResQue evaluation framework for recommender systems

Several other frameworks have been proposed with a focus on specific application domains. An example is a framework for evaluating the usability of clinical monitoring technology, which takes into account direct testing, such as thinking aloud, question asking, co-discovery, performance and psychophysiological measurement, indirect testing methods, such as questionnaires and interviews, observation and ethnographic studies, as well as self-reporting logs (Daniels, Fels, Kushniruk, Lim, & Ansermino, 2007). In the clinical context, a unified framework for Electronic Health Records systems usability is TURF, which features four basic components: task, user, representation, and function (Zhang & Walji, 2011). TURF defines usability as how useful, usable, and satisfying a system is for the intended users to accomplish goals in the work domain by performing certain sequences of tasks, and provides a set of measures for each of the usability dimensions (Figure 18). The authors indicate that TURF can be used: (1) for describing, explaining, and predicting usability differences; (2) for defining, evaluating, and measuring usability

objectively; (3) for designing built-in good usability; and (4), once fully developed, for developing HER usability guidelines and standards.
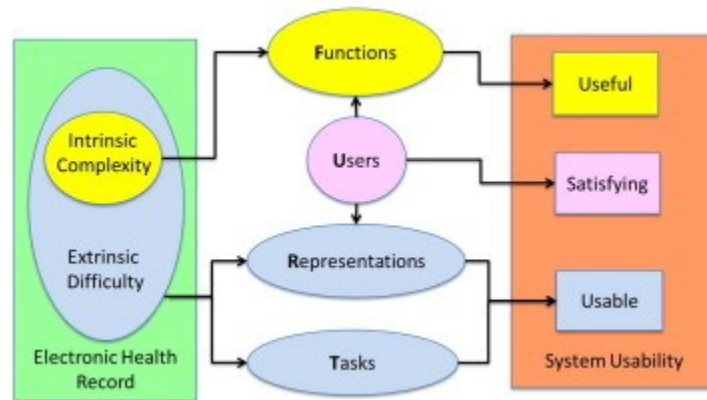
Figure 18. The TURF framework for HER usability

In the context of learning, and more specifically mobile learning, Vavoula and Sharples (2009) identify six challenges and propose a three-level evaluation framework. The identified challenges in evaluating mobile learning include: capturing and analysing learning in context and across contexts, measuring mobile learning processes and outcomes, respecting learner/participant privacy, assessing mobile technology utility and usability, considering the wider organisational and socio-cultural context of learning, and assessing in/formality. The evaluation framework that is proposed to address the aforementioned challenges includes three levels of evaluation: (i) micro level, which examines the individual activities of the technology users and assesses the usability and utility of the educational technology system; (ii) meso level, which examines the learning experience as a whole, to identify learning breakthroughs and breakdowns, and assesses how well the learning experience integrates with other related activities and experiences; and (iii) macro level, which examines the impact of the new technology on established educational and learning practices and institutions. Cota, Díaz, and Duque (2014) stipulate that the evaluation of an m-learning application can be decomposed to evaluation of pedagogical usability and user interface usability. User interface usability is further decomposed in six constructs, which in turn consist of sub-constructs, as follows: (i) operability, defined by the ease of use, navigation, orientation, flexibility, and functionality, (ii) user error protection, analysed in error prevention, freedom, error tolerance; (iii) aesthetics, including the criteria of attractiveness, presentation, consistency, and understandability; (iv) feedback, determined by the attributes of progress, alerts, encouragement, help, support, precision, and system status; (v) accessibility, consisting of adaptability, links, search, as well as input/output support; (vi) motivation, determined by game-based learning, competitiveness, engagement, immersion, intrinsic and extrinsic motivations, and convenience.

Regarding mobile phones and their applications, the GQM framework (Hussain & Kutar, 2009) includes metrics organized under six guidelines categories (simplicity, accuracy, time taken, features, safety, and attractiveness) further clustered in the three main dimensions of usability: effectiveness, efficiency and satisfaction. The specific metrics proposed are built around the following questions:

- Simplicity: Is it simple to key-in the data? Does the application provide virtual keyboard? Is the output easy to use? How easy is it to install the application? Is the application easy to learn?
- Accuracy: Is the application accurate? How many tasks are successful in the first attempt? How many tasks are successful in a given time?
- Time taken: How much time does it take to complete a given task? How much time does it take for the application to respond? How much time does it take for the user to learn?
- Features: Does the application provide appropriate help? Does the application provide appropriate menu buttons for touch screen? Does the application provide voice assistance? Does the application provide automatic update?
- Safety: Is there any effect to the user while using the application? How do users feel when using the application?
- Attractive: Are users happy with the interface? Are users familiar with the interface?

Heo, Ham, Park, Song, and Yoon (2009) introduce a conceptual framework to assist experts in the evaluation of mobile phones, supporting task-based and task-independent evaluation. The framework proposes that the evaluation should focus on three different interface types: Logical User Interface (e.g., menu and navigation structure), Physical User Interface (e.g., keypad and microphone), and Graphical User Interface (e.g., icons and fonts). Independently of the interface type assessed, five usability indicators can be used to guide the expert's review, namely visual support of task goals, support of cognitive interaction, support of efficient interaction, functional support of user needs and ergonomic support.

Physical design factors constitute an important aspect of the framework proposed by Jin, Ji, Choi and Cho (2009), which evaluates the relationship between consumer sensation and usability among the physical design factors of a product (in the specific case dishwashers). To this end, the method of Quality Function Deployment was used, which is a technique that evaluates the ideas of key stakeholders for developing a product that better addresses customer needs. Four aspects of the evaluation model were analysed with this technique, namely overall sensation factors, detail sensation factors, usability evaluation factors, and physical design factors of products.

In the context of haptic systems, based on the ISO standard for guidance on haptic and tactile interaction, Khan, Sulaiman, Said, and Tahir (2011) keep up with the fundamental notions of usability, and namely

efficiency, effectiveness, satisfaction, and learnability, while they add one more evaluation objective, that of safety, and propose a list of metrics and conceptual measures for each. The metrics proposed (Figure 19) are operability, time behaviour, accuracy, navigability, consistency, flexibility, familiarity, simplicity, user guidance, and resource safety. For each metric, the framework suggests a number of measures, including well-established usability measures (e.g., completion time) as well as measures that specifically pertain to haptic interaction, such as: control of speed and force, accidental activation, appropriate size of haptic objects, controllable force, etc.
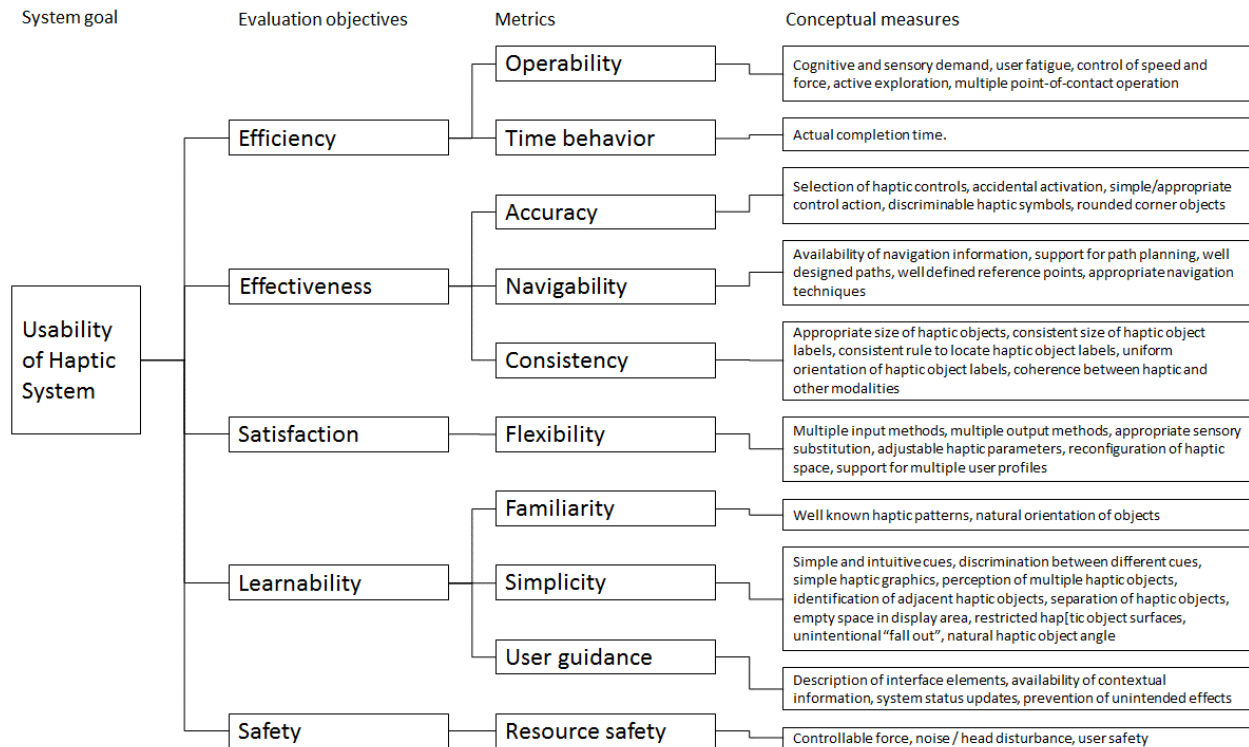
Figure 19. Usability evaluation framework for haptic systems

Finally, an interesting framework is USUS (Weiss, Bernhaupt, Lankes, & Tscheligi, 2009), which addresses usability, social acceptance, user experience, and societal impact of humanoid robots used in collaborative tasks and describes the methodological approach to perform the evaluation of human-robot interaction, including expert reviews, user studies, questionnaires, physiological measurements, focus groups and interviews. It is important to note that the framework is one of the few that go beyond a conceptual model and establish a methodological model as well, describing how to evaluate the constructs involved, an approach which is in line with the current thesis.
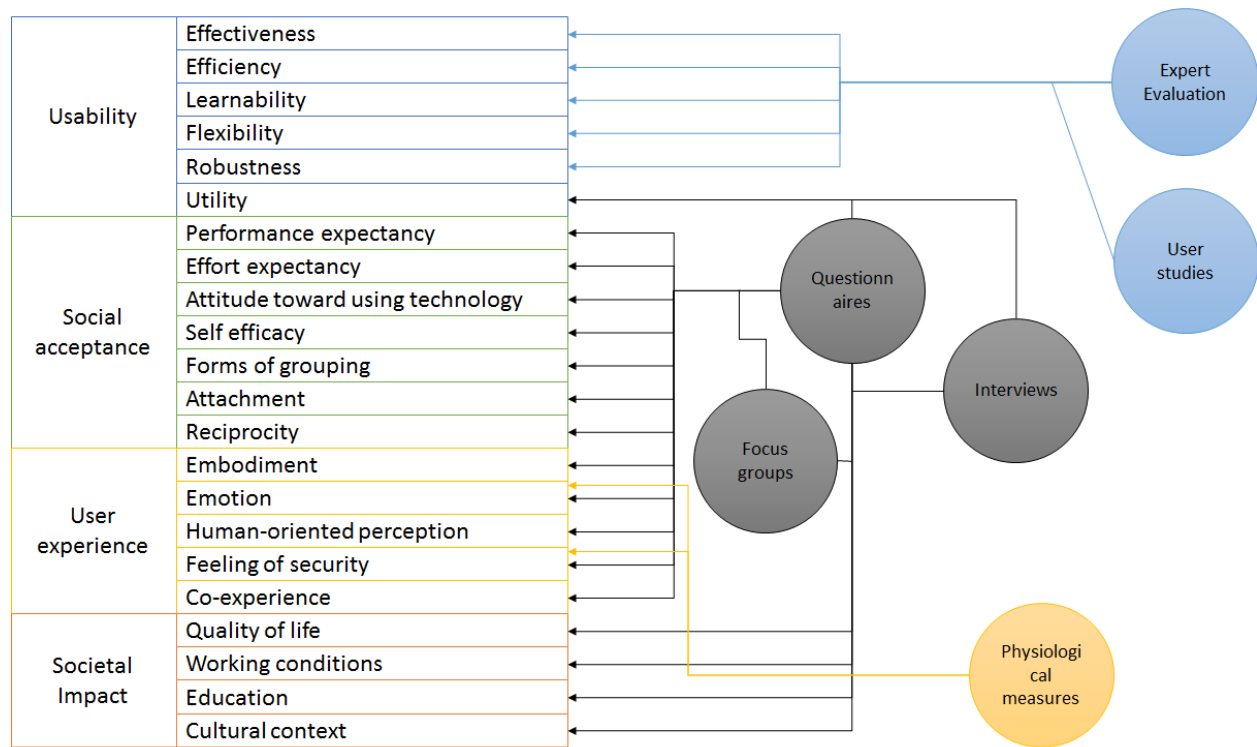
Figure 20. The USUS model for Human-Robot Interaction

Table 7 lists the main attributes proposed in the aforementioned evaluation frameworks. In summary, 97 attributes are recorded, organized under 20 main categories.

Table 7. Characteristics included in usability evaluation frameworks

Category	Attribute	References
Effectiveness	Consistency, Feedback, Accuracy, Orientation, Navigability, Understandability	Cota et al. (2014), Khan et al. (2011), Seffah et al. (2006)
Efficiency	Time behaviour, Resource utilization, Minimal action, Minimal memory load, Operability, Capacity	Cota et al. (2014), Hussain and Kutar (2009), Khan et al. (2011), Kurosu (2015), Seffah et al. (2006)
Flexibility	Context conformity, Context extendibility, Accessibility	Bevan (2009), Cota et al. (2014), Kurosu (2015)
Learnability	Familiarity, Ease of use, Easy to learn, User guidance	Cota et al. (2014), Khan et al. (2011), Seffah et al. (2006), Zhang and Walji (2011)

Functional suitability / Usefulness	Functional Completeness, Functional correctness, Functional appropriateness, Match to the user characteristics, Match to the context of use	Kurosu (2015), Seffah et al. (2006), Zhang and Walji (2011)
Safety	Commercial damage, Operator health and safety, Public health and safety, Environmental harm, Freedom from risk, Resource safety, Insurance	Bevan (2009), Cota et al. (2014), Hussain and Kutar (2009), Khan et al. (2011), Kurosu (2015), Seffah et al. (2006)
Security	Confidentiality, Integrity, Non-repudiation, Accountability, Authenticity	Kurosu (2015)
Reliability	Maturity, Availability, Fault tolerance, Recoverability	Kurosu (2015), Seffah et al. (2006)
Compatibility	Co-existence, Interoperability	Kurosu (2015)
Universality	Self-descriptiveness, Readability, Controllability, Simplicity, Privacy	Hussain and Kutar (2009), Seffah et al. (2006)
Maintainability	Modularity, Reusability, Analysability, Modifiability, Testability	Kurosu (2015)
Motivation	Engagement, Immersion, Convenience, Intrinsic and extrinsic motivations	Cota et al. (2014)
Physical User Interface	Ergonomics, Position and Manipulation	Heo et al. (2009)
Objective artefact quality (other)	Cost, Recognisability, Memorability, Error protection, Error tolerance, Novelty, Rarity, Robustness,	Cota et al. (2014), Kurosu (2015), Weiss et al. (2009), Zhang and Walji (2011)
User orientation	Visibility, Competitiveness	Mourouzis et al. (2006)
User satisfaction	Likability, Attractiveness, Pleasure, Joyfulness, Delightfulness, Comfort, Trust, Sense of security, Sense of accomplishment, Sensation satisfaction	Bevan (2009), Cota et al. (2014), Hussain and Kutar (2009), Jin et al. (2009), Kurosu (2015), Seffah et al. (2006), Zhang and Walji (2011)
User beliefs	Perceived qualities of recommendations, Perceived ease of use, Perceived usefulness, Control / transparency	Pu et al. (2011), Zhang and Walji (2011)
Behavioural intentions	Use the system, Purchase, Continuance, Social influence	Mourouzis et al. (2006), Pu et al. (2011)

Social acceptance	Performance expectancy, Effort expectancy, Attitude toward using technology, Self-efficacy, Attachment, Reciprocity	Weiss et al. (2009)
Societal impact	Quality of life, Working conditions, Education, Cultural context	Weiss et al. (2009)

In summary, usability evaluation frameworks have attempted to adopt a broader perspective and move beyond the notion of usability, however it is a fact that usability and its sub-constructs constitute important parameters in all the frameworks. Influenced by the user acceptance theories and the UX movement, several frameworks have studied quality in use, societal impact, aesthetics, usefulness, as well as the influence of a system beyond its usage, encompassing therefore purchase and continuance intentions. Additionally, the concepts of trust and safety have been adopted in several frameworks, which become highly important in contexts such as mobile and health IT. Finally, extending the evaluation concept to domains beyond human-computer interaction has drawn attention to other factors, such as the physical design of an artefact. It is noticeable however, that the majority of efforts constitute only conceptual frameworks, describing the parameters that should be studied, but not the methods to study them, an issue that is anticipated in the context of the UxAmI framework, which establishes both a conceptual and a methodological framework towards the evaluation of UX in AmI environments.

3.2.3   TOOLS FOR WORKING WITH GUIDELINES

Guidelines list well-known principles for user interface design which should be followed in the development of a system (Nielsen, 1992). They are widely used in HCI to assist designers in their decision-making process, providing ready-to-use knowledge and recommendations. Guidelines are widely used during the early phase of design of an interactive system (Grammenos, Akoumianakis, & Stephanidis, 2000), as well as in the context of expert-based usability evaluations (Jeffries et al., 1991; Nielsen, 1994a). Nevertheless, a number of problems have been identified in their use, including their huge number, variations across contributing disciplines and conflict (Vanderdonckt, 1999). An approach towards resolving several of the shortcomings involved in the process of using guidelines is the development of tools for working with guidelines. Tools for working with guidelines (TFWWG) can be broadly classified in two main categories, namely tools for access and retrieval of guidelines and tools for automatically or semi-automatically evaluating user interface layout representations (Tran, Ezzedine, & Kolski, 2013).

In the first category, SDISelect (Vanderdonckt, 2001) is a small knowledge-based system to assist designers in selecting an appropriate interaction style for a particular context of use. For this purpose, guidelines for selecting the appropriate style are provided on the basis of parameters specified regarding the task (e.g., minimal/moderate/maximal prerequisites, low/moderate/high productivity), the user (e.g., elementary/regular/rich task experience, elementary/regular/rich system experience), and the environment (processing type and capacity). Other approaches include the tool PROKUS (Zülch & Stowasser, 2000), which assists experts in evaluating a system based on a catalogue of questions, the GUIDE tool (Henninger, 2000), which organizes guidelines under fundamental questions / problem characteristics (e.g., what navigation strategies will be used) and provides support for design in a specific organization by adopting "organizational memory". Another approach with targeted scope is the usability assistant for the heuristic evaluation of interactive systems (Pribeanu, 2009), a tool which aims at facilitating the heuristic evaluation process, focusing therefore on specific guidelines.

In the second category, Sherlock (Grammenos et al., 2000) supports both manual and automated inspection of guidelines. The tool is compatible with the Microsoft Visual Basic Integrated Development Environment (IDE), as a popular development environment of the time it was developed. Automatic evaluation is carried out by parsing a Visual Basic project and creating a textual description of the user interface. Inspection is carried against the rules selected by the administrator and an evaluation report is produced, explaining each error and providing details regarding the guidelines that are violated, as well as possible solutions. Rules are loaded as external libraries, which makes the system easily extensible. Towards automatic evaluation, but with a restricted focus, ErgoSim (Bouzit, Calvary, Chêne, & Vanderdonckt, 2016) is a software tool that can automatically evaluate the design of menu bars, pull-down menus, and sub-menus of a graphical user interface by reviewing usability guidelines related to menu design.

An alternative approach to automated usability inspection based on guidelines is proposed by Charfi, Trabelsi, Ezzedine and Kolski (2011), suggesting the implementation of User Interfaces (UIs) with customized controls that evaluate themselves. The implemented tool, ISUTI (Charfi, Trabelsi, Ezzedine, & Kolski, 2013), extends Microsoft Visual Studio 2010 controls and has been used in the context of implementing a network transportation system, providing early support to the application developers.

In any approach followed in the TFWWG context, tools are required to perform a multi-step procedure including the collection of guidelines, their organization, as well as incorporation into the approach, operationalization and usage (Vanderdonckt, 2001). To this end, several classification schemes have been proposed among the various tools. For instance, guidelines may fall into one of the three categories:

generic standards and guides, platform-specific style guides, product-line or corporation-specific style guide (Parush, 2000). Extensible solutions proposed more sophisticated schemes for representing guidelines. For example, a Rule entity in the Sherlock system (Grammenos et al, 2000) consists of several fields, including its title, class (describing depicts the full path from the root of a guidelines collection tree to the leaf containing the specific guideline), description, reference, known conflicts, author, and source DLL. With a focus on providing a solution towards tracing conflicting guidelines, Masip et al. (2012) suggest that a guideline should feature the following attributes: description, source, application domain, UI components, keywords, factors, criteria, factors' importance, and pointers to potential conflicting guidelines.

Finally, several approaches have focused on web evaluation and the list of tools in this context is considerably larger. Potential reasons for this is on the one hand, the wide penetration of the World Wide Web in any life domain, and on the other hand the strict syntax of the HTML language employed for the development of web sites, which facilitates automated inspection. One of the first and more influential efforts in this domain is WebTango (Ivory & Hearst, 2001; Ivory & Hearst, 2002; Ivory, 2013), which developed 157 page- and site- level measures, as well as an analysis tool encompassing several statistical models for assessing Web page and site quality. Other approaches include, but are not limited to, WebRemUSINE (Paganelli & Paternò, 2002), AWUSA (Tiedtke, Märtin, & Gerth, 2002), KWARESMI (Beirekdar, Vanderdonckt, & Noirhomme-Fraiture, 2003), EvalIris (Abascal, Arrue, Fajardo, Garay, & Tomás, 2004), the USEFul framework (Dingli & Mifsud, 2011), and AWebHUT (Rukshan & Baravalle, 2011). Many efforts have also focused on web content accessibility evaluation. Indicative of the large number of automated evaluation tools, is the list of 93 tools[2] provided by the World Wide Web Consortium (W3C) to assist in web page accessibility evaluation.

In summary, automated evaluation of a UI against specific guidelines is a challenging task that can only be partially supported. Two major constraints are identified in this domain: (a) automated evaluation tools are often platform specific (e.g., Visual Basic, Visual Studio, HTML), an approach that cannot be viable in AmI environments and (b) only a limited set of guidelines can be automatically tested. Regarding the latter, it is apparent that only the guidelines that can be operationalized can be automatically checked (e.g., that a label is provided for a button). However, human intervention is always required to evaluate a UI with reference to higher level guidelines (e.g., "Map between system and the real world" meaning for instance that the label text should be appropriate for the task at hand). Such early automated inspection tools were successful in providing guidance, as often UI design was carried out by developers themselves. Recent advances in HCI however turned user interface/experience design as a highly influential professional field (Shneiderman, 2017), therefore the target audience of these tools is no longer developers, but UX specialists. Consequently, it is questionable whether the effort spent by a UX expert

---

[2] https://www.w3.org/WAI/ER/tools/

in preparing, reading, and assessing automated evaluation reports will be less than the time required to locate by oneself the corresponding major usability issues (e.g., misaligned elements in a UI, poor colour choice, etc.). What's more, a good UX is not just a matter of following recipes regarding distances, alignments, and colour contrasts. It is mainly a balance of all the above, with good and meaningful information structure, appropriate functionality, and aesthetic design. Finally, the ubiquity of digital applications and the new era of digital natives has made obsolete early conventions and design guidelines – the days of relying on blue underlined links are gone (Schlatter & Levinson, 2013). Therefore, the need for evaluation from UX experts and interpretation of the guidelines cannot be overcome, sometimes not even partially supported anymore through automated measurements.

It has been early recognized that a problem with using guidelines is their constantly increasing volume (Vanderdonckt, 1999), however all previous approaches do not alleviate the need for selecting which sets of guidelines the system or the expert should take into account. Considering the evolution in the HCI domain, the volume of literature is quite larger today (Shneiderman, 2017). In the context of designing and evaluating applications and systems in AmI environments, the diversity of technologies, the pervasive nature of applications, the multimodality provided, the variety in contexts, all dictate the need for researching and considering an eventually unmanageable number of guidelines. To this end, this thesis proposes a tool that automatically suggests guideline sets that should be taken into account when evaluating an AmI application or system, based on the devices and services that the application uses. Furthermore, it proposes an easy way for adding guidelines, avoiding long and complicated input procedures, while it takes advantage of crowdsourcing and online community assets to facilitate adding new high quality content and retrieve best practice examples.

3.2.4 AUTOMATED MEASUREMENTS

Automated usability evaluation methods are promising complements to traditional UEMs, assisting evaluators to identify potential usability problems (Ivory & Hearst, 2001). A classification of UEMs with a focus on their support towards automated measurements is provided in Ivory and Hearst (2001), whereby methods are organized according to their class (testing, inspection, inquiry, analytical, simulation), type (e.g., log file analysis, guideline review, survey), automation type (none, capture, analysis, critique), and effort level (minimal, model development, informal use, formal use). The study that was carried out explored automation support organized in four main categories: usability testing methods, inspection methods, inquiry, analytical modelling and simulation.

Another study of automation support classifies approaches for extracting usability information from UI events, according to the supported techniques (Hilbert & Redmiles, 2000), and namely:

- Synchronization and searching, allowing the synchronization and cross-indexing of UI events with other sources of data, such as video recordings and observation logs
- Transforming event streams, targeted to selecting, abstracting and recording event streams to facilitate human and automated analysis
- Analysis, involving (a) performing counts and summary statistics, (b) detecting sequences, by allowing investigators to identify occurrences of concrete or abstractly defined target sequences within source sequences of events, (c) comparing source sequences against target sequences and indicating the extent to which they match, (d) characterizing sequences by creating an abstract model to summarize or characterize their interesting features
- Visualization, presenting the results of transformation and analyses so as to facilitate the evaluators in interpreting the results
- Integrated evaluation support, facilitating the flexible composition of various transformation, analysis and visualizations.

The classification and examples of tools and environments discussed are based on their technical capabilities, however it is pointed out that there is very little data published regarding the relative utility of the surveyed approaches in supporting usability evaluations. Moreover, it is highlighted that more advanced methods require the most human intervention, interpretation, and effort, while the more automated techniques tend to be least compelling and most unrealistic in their assumptions. In the context of the current thesis, tools to support usability testing automation through logging will be discussed, as the ones more relevant to the current work.

A common approach in such tools is the generation of statistics and automatic calculation of usability metrics. DRUM (Macleod & Rengger, 1993) supports management of evaluation data, task analysis, video mark-up and logging (real-time and retrospective logging of events), analysis of logged data and calculation of metrics. In more details, DRUM provides the following automatically calculated metrics, based on logged data: task time; snag, help and search times; effectiveness; efficiency; relative efficiency, compared with experts or with the same task on another system; productive period. AIDE (Sears, 1995) is a metric-based tool assisting designers in creating and evaluating layouts for a given set of interface controls. More specifically, AIDE includes five metrics: efficiency, evaluating how far the user must move a cursor to accomplish their tasks; alignment, assessing how well objects are aligned; horizontal and vertical balance, calculating how balanced is the screen in the two axes; and constraints, providing a quick overview of the status of any designer-specified constraints. USINE (Lecerof & Paternò, 1998) is a tool that takes as input the task model of the system describing the user's interactions with the system, as well as a log-task table, created with information from the task model and one log file which contains all the

possible actions, mapping logged actions with tasks in the model. The tool supports the evaluator by providing the accomplished, failed and never tried tasks, number of user errors, the time that user errors occurred, time to complete each task, and sequences of accomplished tasks that occur more than once.

In the context of analysing the user's behaviour and problem solving process, AMME (Rauterberg, 1993) employs Petri nets to reconstruct the user's interaction with the system, facilitating qualitative analysis and identification of specific types of pattern from the evaluator. The symbolic representation of a machine system consists of objects, operations and states and AMME postulates that given a finite action space, each state corresponds to a system context, and each transition to a system operation. AMME also provides quantitative metrics for the task solving and the behavioural process. Metrics in the first category include task solving time, number of all used transitions, and number of all occurring states. The quantification of the complexity of the behavioural process is achieved by calculating the difference of the total number of connections and the total number of states.

With the aim to facilitate recording of user behaviour, TRUE (Kim et al., 2008) propose an approach that combines log files with attitudinal data, received from polling at specific intervals users themselves. The innovative aspects of TRUE include logging sequences of events and not simply events, as well as collecting event sets that collect both the event of interest as well as the contextual information needed to make sense of that event. Automation support is provided in terms of synchronizing the video that is captured with the logged events, as well as providing visualizations of the recorded events. The enhanced log files with event sets and sequences, allow evaluators to drill down to specific events and determine the causes of the identified problems. The tool has been applied for evaluating the user experience in serious games, however the fact that it constitutes a custom development build in each application that needs to be evaluated (Heilbrunn, Herzig, & Schill, 2014), makes it inappropriate for use as a generic all-purpose usability evaluation tool.

In the context of visualizing user activity, QUIP (Helfrich & Landay, 1999) provides automatic analysis of usage trace data obtained from real users running an instrumented version of a target application, where instrumentation refers to recording each action that the user performs. The system produces a directed graph illustrating users' action traces, compared to the designer's "ideal" action trace.

An important concern in the development of usability evaluation automation tool refers to instrumenting the software to collect usage data. In summary, five main methods are reported in literature for instrumenting systems, and namely (Bateman, Gutwin, Osgood, & McCalla, 2009): manual instrumentation, by adding logging instructions to the code of the system; toolkit instrumentation, during which the toolkit used for the presentation and handling of the UI is instrumented; system-level

instrumentation that uses logging at the operating-system level; and aspect-oriented instrumentation. Following the latter approach UMARA (Bateman et al., 2009) is an interactive usability instrumentation tool, which allows evaluators to specify what actions to log, by clicking on interface elements in the application itself (e.g., select a text field of interest and then decide which events to monitor for this element, including mouse events, windows events, keyboard events, focus events, etc.).

An automated usability evaluation tool running as a service in windows environment and supporting data collection, metrics and data analysis is described by Chang and Dillon (1997). The data collected include messages that the user sends to the application being tested, messages sent by the system to the user, keystrokes and mouse clicks. The system calculates the number of windows opened (total and per window), number of times a menu is selected, and number of times a button is pressed. Finally, the system produces charts to illustrate mouse density, mouse travel pattern, and keystrokes. AppMonitor (Alexander, Cockburn, & Lobb, 2008) is a windows-based tool that has been designed to record low-level and high-level events for two specific windows applications, Microsoft Word and Adobe Reader. The tool runs in Microsoft Windows XP platform, and listens to events exchanged between the applications and Windows through an event-hooking Dynamic-Link Library (DLL). The evaluators can select the windows events they wish to be monitored and logged, while the output of the tool is a file listing all the events that have been captured. Ma et al. (2013) propose a usability evaluation toolkit for mobile applications that implements a Software Development Kit (SDK) which can be used by the applications with minor modifications in their source code. The toolkit logs view events, dialog and menu events, system keys, and unhandled events that cannot be classified under the previous three event types. It also features an automated metric discovery model based on comparing the ideal sequence of events towards accomplishing a task, as carried out by an expert, with users' sequences of events. Then several usability indicators can be calculated, such as the number of backtracks, correct flow ratio, or the number of users who failed to accomplish the task. EISEval (Tran et al., 2013) is a tool extending usability evaluation automation by capturing data concerning not only the interactions between users and the UI, but between agents themselves as well, thus supporting the evaluation of UIs' dynamic behaviour. EISEval then performs data analysis on the collected data through measurements and statistics (e.g., frequencies, times, successes and failures), and generates PetriNets to visually reproduce the activities of the user in the target system. Evaluators are also supported by an open and modifiable list of criteria, and are facilitated to record their observations for each of the criteria. Before actually being used in the context of an evaluation, EISEval requires the evaluator to specify information about the tasks that can be performed with and by the system, as well as information about agents and other configuration settings. EISEval has been used in the context of the environment proposed by Assila, Oliveira and Ezzedine (2016)

to combine objective and subjective metrics, complemented by a questionnaire generating tool, and a guidelines inspector tool. Objective and subjective results acquired through the aforementioned modules are visualized in scatter plot charts, organized under specific usability indicators (e.g., information density) to facilitate the evaluator in their interpretation. In an effort to provide a more generic evaluation framework that supports usability testing in real production environments and can be applied on arbitrary software application, Muhi, Szőke, Fülöp, Ferenc, and Berger (2013) introduce UEF. UEF uses XML files for describing meta-information about the system and providing concrete usage data (logs), while a validator component checks the log files according to predefined syntactic and semantic rules. Then the data are evaluated according to specific usability model, which can vary for different systems, while the framework also supports subjective evaluation through questionnaires. In the context of developing the UEF framework, Muhi et al. (2013) also developed a list of requirements for general usability evaluation frameworks, as follows: (i) support of real-life production environments, (ii) detection of patterns, (iii) detection of usability bugs, (iv) transparency, (v) automation, and (vi) wide applicability.

Shifting the focus from instrumenting the software to user-based instrumentation, Christensen and Frøkjær (2010) propose DUE, a technique for collecting and evaluating usability data based on users to report them. More specifically, DUE supports recording video from the user's screen, as well as voice recordings. When the user detects a usability problem they press a button to report it, record an explanation, and rate its severity.

An interesting approach, alleviating the need for any instrumentation and event logging is scvRipper (Bao et al., 2017), a tool which uses computer vision scrapping to automatically extract time-series data (software used and application content accessed/generated) from screen-captured videos, enabling thus the creation of quantitative metrics. Although no instrumentation is required, it should be noted that a sampling process is required once for each application to define the application windows, during which each window is defined through collecting sample images of its visual cues. Along the same lines, an early approach is IBOTS (Zettlemoyer, Amant, & Dulberg, 1998) which logs predetermined low-level events and by using image processing techniques it can recognize specific UI elements in a screen. The system does not create extended log files, instead it can replay user interactions as they occurred. By being able to associate events with specific UI components as they have been recognized, the system provides support for identification of patterns of behaviour, however their interpretation, visualization, and analysis are part of future work.

Commercial tools on the other hand avoid any instrumentation, and support a variety of features, such as data capture of low-level events (keystrokes, mouse clicks, system events), metrics (e.g., time or activity),

screen video capture, logging of observational comments, event definitions allowing the association of hot keys with the defined events (Howarth, Smith-Jackson, & Hartson, 2009). In addition, they rely on experiment observers to identify task success and user errors (Morae, 2017), while eye tracking plugins[3,4] are supported to combine actions of users with their eye-gaze trails, as well as connection with physiological measurement systems [5] to acquire physiological data, such as blood pressure, skin conductance etc.

Using logs in the context of web sites is a more complicate and challenging process, as these logs are not associated with observation, therefore the metrics logged, their analysis and visualization should point to deeper comprehension of the user experience. User activity with a web site can be logged in the user's own computer (client-side logging), on the server, or using a proxy server (Kellar, Hawkey Inkpen, & Watters, 2008).Typically, techniques for analysing and visualizing data include (Menez & Nonnecke, 2014): frequency analysis, looking at the frequency of user interactions with the potential to reveal users' preferences and behaviour; time analysis, which looks at the amount of time spent on the website in a single visit and can indicate proficiency, higher cognition demands, or that the content is interesting to the user; exit analysis, exploring the reasons that a user left a web page (e.g., bounce rate); and pattern analysis that aims to discover patterns within the data and may include pairing analysis to determine how often certain actions are performed in a sequence, clustering to discover similar groups of user based on preferences or usage, and path analysis used to understand common navigation patterns.

Despite the fact that several tools and frameworks towards usability evaluation automation have been proposed in literature, it is noteworthy that most usability testing is done in a very manual, labour-intensive way (Norman & Panizzi, 2005). A possible explanation for this is that although statistical information is calculated, the results are often not useful for the evaluator, as the data logged leave out the user's goals and intentions and much of the user's focus of attention when not actually clicking on a button or typing in a field (Norman & Panizzi, 2005). Efforts towards capturing users' goals require instrumentation of the process, however the complexity of today's systems makes successful instrumentation a challenging task (Kim et al., 2008). An important challenge regarding the instrumentation solution is that it should easily map between: UI events and application features; lower level and higher level events (e.g., typing, deleting, and moving); and events and the context in which they occur (Hilbert & Redmiles, 2000). Furthermore, challenges that logging approaches should address are to be designed to focus high-level user actions, capture provenance of all events, observe intermediate user actions, obtain the analysed data's metadata and statistics, and collect user goals and feedback (Alspaugh, Ganapathi, Hearst, & Katz, 2014). Moreover, Au et al. (2008) highlight the requirements that an automated usability testing tool should meet, including: capturing a range of inputs, performing analyses

---

[3] https://www.techsmith.com/morae.html#
[4] http://www.noldus.com/human-behavior-research/accessories/eye-trackers
[5] http://www.noldus.com/human-behavior-research/accessories/daq-systems

on different aspects of usability, presenting results clearly, being simple and flexible to use, and being able to be used throughout development.

In the light of the above, an innovative usability automation tool is proposed in the context of the current thesis, aiming to (i) emphasize on metrics and statistical information that is meaningful and useful to the evaluator, (ii) ensure that information is provided with a minimum level of effort for the evaluators, the designers, and the software engineers involved and (iii) provide valuable UX data enriched with context information acquired through the AmI environment's sensors and agents.

3.3   USER EXPERIENCE EVALUATION

User Experience (UX) has recently become a popular concept widely adopted by the HCI community, and has predominated the usability concept. It is intriguing however, that although UX has been so widely accepted, it was not clearly defined or well understood until recently (Law, Roto, Vermeeren, Kort, & Hassenzahl, 2008). As a result, a considerable number of definitions and viewpoints on UX exist (Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009), while several research efforts have aimed at clarifying the concept (Law et al., 2008, Hassenzahl & Tractinsky, 2006).

For instance, Hassenzahl and Tractinsky (2006) identified three main perspectives of UX (Figure 21):

- Addressing the human needs beyond the instrumental, which studies the interactive product in a more holistic approach beyond the achievement of user's behavioural goals in work settings and includes user needs such as aesthetics, personal growth, increase of knowledge and skills, self-expression, interaction with relevant others, self-maintenance and memories.
- Affective and emotional aspects of the interaction, which considers emotions as consequences of product use and as antecedents of product use and evaluative judgements.
- The nature of experience, which emphasizes two aspects of technology use: its situatedness and its temporality.
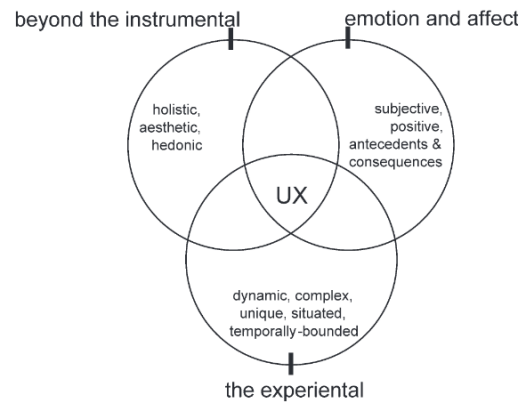
Figure 21. The facets of UX
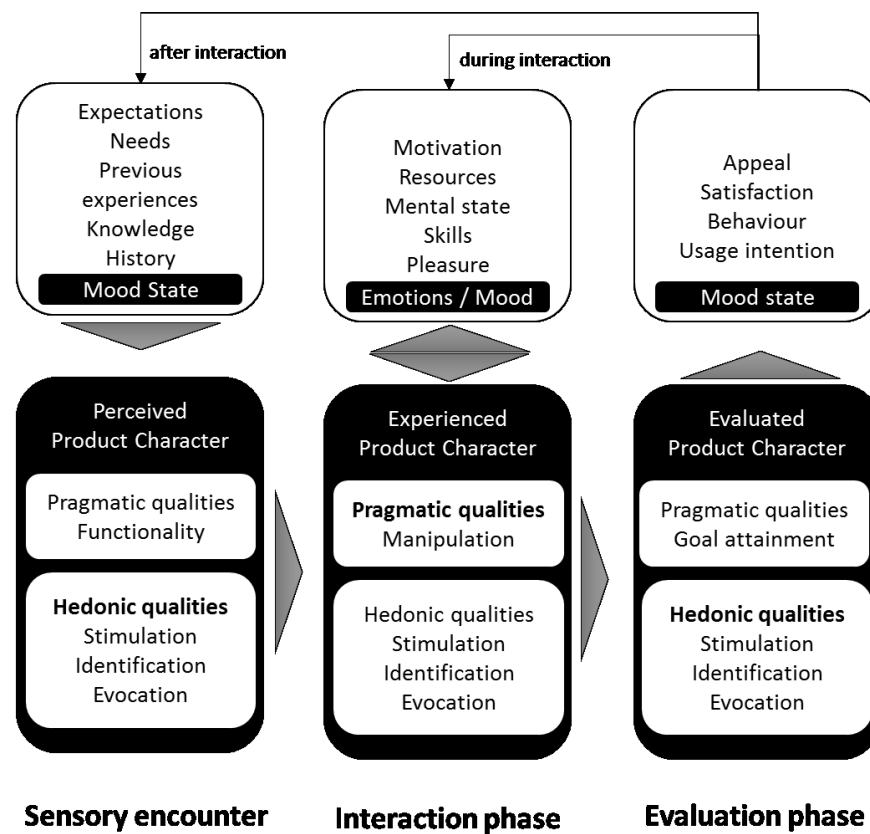
Figure 22. User Experience Framework (Zimmermann, 2008)

Zimmermann (2008) proposed a UX framework organized according to the three phases of user's encounter with a product: sensory encounter when the user first encounters the product, interaction

phase, and evaluation phase which takes place after the interaction or during breaks in the interaction phase (Figure 22). At the beginning of the user experience process, the user perceives the product's features (e.g., layout, content, functionality, interaction capabilities). These are combined with the user's expectations, needs or standards to form the perceived product character. The product character consists of pragmatic and hedonic qualities, the relative importance of which can change over the course of the experience.

A framework studying how UX changes over time entails three phases in the adoption of a product - orientation, incorporation, and identification – and explores how the temporality of experience represented by the constructs of familiarity, functional dependency, and emotional attachment motivate the transition between the three different adoption phases (Karapanos, Zimmerman, Forlizzi, & Martens, 2009). A five-week ethnographic study that was carried out prior to the conceptualization of the framework identified that early experiences seemed to relate mostly to hedonic aspects of product use, while prolonged experiences became increasingly more tied to aspects reflecting how the product becomes meaningful in one's life.

A clarifying definition of UX was provided by ISO 9241-210:2010, which described UX as a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service, and includes all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use. In addition, ISO 9241-210:2010 notes that user experience is a consequence of brand image, presentation, functionality, system performance, interactive behaviour and assistive capabilities of the interactive system, the user's internal and physical state resulting from prior experiences, attitudes, skills and personality, and the context of use.

It is noteworthy and indicative of the dynamic of the UX notion the fact that even after the ISO definitions, new conceptual UX frameworks, studies and definitions continued to emerge. In an effort to illustrate that user experience is a complex construct, al-Azzawi (2014) identified three main categories under which UX processes may be classified, namely interaction, construction, and evaluation and discusses user experience from a variety of perspectives using the theories of experience as qualities, experience as consequences, and experience as processes. The role of time and the resulting different facets of UX have been studied by Marti and Iacono (2016) in a four-week research. Four types of UX were compared: anticipated referring to the period of time before first use, momentary concerning any perceived change during the interaction at the moment it occurs, episodic which is an appraisal of a specific usage episode extrapolated from a wider interaction event, and remembered referring to the memory the user has after

having used the system for a while. The results of the study confirmed that UX changes over time, while issues related to usability and reliability of data become dominant to users' perceptions of the product after four weeks of use.

In the process of defining a framework to understand UX of mobile video, Song, Tjondronegoro, and Docherty (2012) carried out a review of existing frameworks concluding that the following parameters are considered in total by all the seven reviewed frameworks (but not by each and every one): (i) user attributes of emotion, needs, prior experiences, perceptions, expectations, motivation, profile, physical resources, (ii) system or service attributes and namely product appearance or system complexity, functionality, usability, aesthetic quality, interactivity, and (iii) context attributes, and more specifically, context of use or physical context, social context, cultural context, temporal and task context. The authors proposed their own framework for describing UX of mobile video featuring user, system and context attributes related to user and system. User attributes include: Audio/Visual system and perception, motivations, profiles (age, sex, preferences, prior experiences and technology background), needs, emotion (pleasure, enjoyment), and expectations. System attributes are further classified in three subcategories: device characteristics (screen size, display resolution, CPU, memory, battery lifetime, user interface), network characteristics (network bandwidth, channel performance, data cost), and video service characteristics (usability and interactivity, content availability, bit rate, video/audio quality, codec, delivery strategy, commercial plan). Finally, four context types are described which are inter-related: physical (where and when a user is using the mobile video, available network), social and cultural (sharing or solitary use of mobile video, selections of video content, voting popularity), temporal (user's available time and willingness, system's battery consumption, network switch, and video length), and task (other tasks of the user, and other usages of the device).

Park, Han, Kim, Oh, and Moon (2013) introduced a total of 22 dimensions including overall UX and three more elements of UX (usability, affect, and user value), as well as their 18 sub-elements to evaluate UX with a commercial tablet PC. More specifically, the eighteen sub-elements were clustered as follows:

- Usability: simplicity, directness (degree of user's perception of directly controlling the user interface of a product), efficiency, informativeness, flexibility, learnability, user support.
- Affect: colour (degree to which the colour used in a product is likable or vivid), delicacy (degree to which a product is elaborate, or finely and skilfully made), texture (degree to which a product's texture or touch appeals to the users), luxuriousness (degree to which a product is luxurious or looks expensive and superior in quality), attractiveness, simplicity.

- User value: self-satisfaction (degree to which a product gives a user satisfaction with oneself or one's achievements), pleasure, customer need, sociability, attachment (ability for the user to attach subjective value to a product).

In the context of the study that was carried out, all metrics were acquired through users' rating in questionnaires. The results were studied using several quantification models, while it turned out that directness, flexibility, colour, texture, simplicity, and attachment seem to be important sub-elements of UX, because these are involved in all the "best models", which were selected based on the results of the quantification models.

Olsson (2014) proposed an early framework towards understanding users' expectations, which in turn play an important role in the overall user experience with a product or service. The framework suggests four layers of expectations, namely desires, experience-based assumptions, social and societal norms, as well as must-be expectations. Expectations stemming from experience-based assumptions reflect what people are habituated to and how own and important others' experiences shape their conceptual models regarding how technology should perform, behave and evolve. Must-be expectations also stem from users' past experiences, but they represent requirements based on negative experiences with other products or services.

Besides comprehending the UX notion, an important concern relates to measuring UX. To this end, ISO 9241-210:2010 states that usability, when interpreted from the perspective of the users' personal goals, can include the kind of perceptual and emotional aspects typically associated with user experience, therefore usability criteria can be used to assess aspects of user experience. Nevertheless, it is clear that UX goes far beyond usability and therefore additional methods and metrics are required.

Bevan (2009b) discussed the difference between usability and UX evaluation methods and identified that many people in industry appear to have subsumed usability within user experience. In contrast, researchers working in the field consider user experience to be entirely subjective. In summary, Bevan concluded that methods for UX evaluation can be categorized as methods to evaluate: (i) the hedonic goals of stimulation, identification, and evocation and associated emotional responses, and (ii) the user's perception of achievement of pragmatic goals associated with task success. The second category of methods have already been analysed in the previous section referring to usability evaluation. This section will mainly focus on methods that go beyond traditional usability evaluation.

MacDonald and Atwood (2013) present how evaluation has changed perspectives from 1940 until now, shifting from the system reliability phase (1940-1950), the system performance phase (1950-1960), the

user performance phase (1960-1970), the usability phase (1970-2000), to the current UX phase (2000-present). The authors identify that a major challenge in UX evaluation is the lack of a shared conceptual framework for UX, despite the fact that many models have been proposed. Towards the evaluation of future systems, they highlight five main research areas relevant to creating a more holistic vision for UX evaluation, where hedonic and pragmatic feedback will be seamlessly integrated in order to assess whether systems are useful, usable and desirable. Finally, the need for learning from evaluation as applied by practitioners is stressed, so as to better understand the purpose and role of evaluation in practical settings.

Although the field of UX is relatively new, there is already a plethora of evaluation methods. Allam and Dahlan (2013) provided a taxonomy of the different types of UX evaluation methods according to a variety of criteria. The various UX study types are field studies, lab studies, online studies and questionnaires. With regard to the development phase of the product, scenarios, early prototypes, functional prototypes, or products on market can be the target of evaluation. The studied period of experience, may be before usage, snapshots during interaction, an experience of a task or activity, or long-term UX. Finally, concerning the evaluator, they may be UX experts, one user at a time, groups of users, or pairs of users.

In an effort to elucidate the field Roto, Obrist, & Väänänen-Vainio-Mattila (2009) investigated 30 UX evaluation methods and presented a categorization along their applicability for lab tests, field studies, online surveys, or expert evaluations without actual users, ending up in the following method categories: lab studies with individuals, lab studies with groups, short-term field studies, long-term field studies, surveys, expert evaluation methods and mixed methods. Vermeeren et al. (2010) studied 96 methods and provided a much more sophisticated categorization for characterizing the methods, according to: their origin (academia or industry), type of collected data (quantitative, qualitative, or both), information sources (single users, group of users, experts), location (lab, field, online), period of experience (momentary UX, UX of single episodes and test sessions, before usage, long-term usage), development phase (later stage when a prototype is available, earlier stages when concept ideas or non-functional prototypes are available). On the other hand, carrying out a critical analysis of 66 empirical studies of user experience, Bargas-Avila and Hornbæk (2011), found that (i) context of use and anticipated use, often named key factors of UX, are rarely researched (ii) emotions, enjoyment and aesthetics are the most frequently assessed dimensions (iii) the methodologies used are mostly qualitative, and known from traditional usability studies, and that (iv) many studies use constructive methods and self-developed questionnaires with unclear validity. Finally, a recent review of UX evaluation studies (Maia & Furtado, 2016) identified that the majority of studies (80%) collect data manually, 12% do it in mixed form and only 8% do it automatically. In addition, it is reported that 76% of the studies involved a single measurement

while only 24% over-the-time or repeated measurements. With regard to the tools and techniques used to evaluate use experience, 84% of the studies employ questionnaires, 16% interview, 8% online survey, 8% the UX-Curve tool, while observations, reaction cards, video recordings, brain-computer interface, eye tracking, and face recognition were employed in only 4% of the studies.

From the above it is evident that there is not a clear and established methodology for evaluating user experience as a whole. On the contrary, there is a vast number of methods aiming to evaluate specific UX aspects[6]. Compiling a detailed list of all the UX-related methods is beyond the scope and the aims of the current thesis. However, with the aim of studying how these methods can be embedded in the proposed evaluation framework, an indicative sample of methods is presented, providing for each method a short description and a reference for further information. Methods are organised under the UX aspect measured, while for each UX aspect, the most influential or innovative methods have been selected.

3.3.1   AESTHETICS

Aesthetics measurement (Lavie & Tractinsky, 2004). An instrument initially developed for measuring web sites aesthetics as perceived by their users, along two dimensions, namely classic aesthetics and expressive aesthetics. In more details, the following aesthetics attributes are explored: (i) classic aesthetics: aesthetic, pleasant, clear, clean, and symmetric design, and (ii) expressive aesthetics: creative, fascinating, original and sophisticated design, design using special effects. Each of the aesthetic dimensions is measured by a five-item scale.

3.3.2   AFFECT AND EMOTION

Affect grid (Russell, Weiss, & Mendelsohn, 1989). It is a scale designed as a quick means of assessing affect along the dimensions of pleasure-displeasure and arousal-sleepiness. In more details, users are provided with a 9x9 grid, featuring the following four feelings: stress (top-left corner), excitement (top-right corner), depression (bottom-left corner) and relaxation (bottom-right corner). Then, they are asked to mark on the grid how they feel.

Emocards (Desmet, Overbeeke, & Tax, 2001). Cards depict eight distinct emotional expressions, which vary on the basis of the dimensions "pleasantness" and "arousal": excited emotions come with high levels of arousal, while calm emotions come with low levels of arousal. Two cards are used to represent each emotion as cartoon graphics, one illustrating a female face and one a male face. A "circumplex of emotions" is divided in eight parts, one for each distinct emotion, while each card pair is placed in the corresponding octant of the circumplex. The eight emotions represented are: excited neutral, excited pleasant, average pleasant, calm pleasant, calm neutral, calm unpleasant, average

---

[6] An online resource with 86 UX methods is: http://www.allaboutux.org/, visited: October 6, 2017

unpleasant, excited unpleasant. Users are asked to indicate which card better expresses their feeling, after using a product.

Positive Affect and Negative Affect Schedule – PANAS (Watson, Clark, & Tellegen, 1988). This scale consists of 20 words that describe different feelings and emotions. Users have to provide a rating from 1 to 5, indicating to what extent they experience the specific feeling. The words describing positive affect are: interested, excited, strong, enthusiastic, proud, alert, inspired, determined, attentive, and active. On the other hand, the words describing negative affect are: distressed, upset, guilty, scared, hostile, irritable, ashamed, nervous, jittery, and afraid.

PrEmo (Desmet, 2005). PrEmo is a non-verbal self-report instrument that measures 14 emotions that are often elicited by product design. Of these 14 emotions, seven are pleasant (i.e., desire, pleasant surprise, inspiration, amusement, admiration, satisfaction, fascination), and seven are unpleasant (i.e., indignation, contempt, disgust, unpleasant surprise, dissatisfaction, disappointment, and boredom). Instead of relying on the use of words, respondents can report their emotions with the use of expressive cartoon animations. In the instrument, each of the 14 measured emotions is portrayed by an animation by means of dynamic facial, bodily, and vocal expressions.

Self-Assessment Manikin (SAM) (Lang, Bradley, & Cuthbert, 1997). It is a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli. SAM varies from a frowning, unhappy figure to a smiling happy figure, when representing the valence dimension. For the arousal dimension, SAM ranges from a relaxed, sleepy figure to an excited, wide-eyed figure. Finally, for the dominance dimension, SAM ranges from a small figure (dominated) to a large figure (in control). The subject can select any of the five figures comprising each scale.

3.3.3 HEDONIC QUALITY

AttrakDiff (Hassenzahl, 2004). A questionnaire aiming to assess user-perceived usability (i.e., pragmatic attributes), hedonic attributes (e.g., stimulation, identification), goodness (i.e., satisfaction), and beauty of interactive products. The questionnaire consists of twenty-one 7-point items with bipolar verbal anchors, as follows: (i) hedonic quality-identification: isolating-integrating, amateurish-professional, gaudy-classy, cheap-valuable, noninclusive-inclusive, takes me distant from people-brings me closer to people, unpresentable-presentable, (ii) hedonic quality-stimulation: typical-original, standard-creative, cautious-courageous, conservative-innovative, lame-exciting, easy-challenging, commonplace-new, (iii) pragmatic quality: technical-human, complicated-simple, impractical-practical, cumbersome-direct,

unpredictable-predictable, confusing-clear, unruly-manageable, and (iv) evaluational constructs: ugly-beautiful, bad-good.

Hedonic / Utility scale (Voss, Spangenberg, & Grohmann, 2003). A scale that measures the hedonic and utilitarian dimensions of consumer attitudes toward product categories and different brands within categories. It includes ten semantic differential response items, five of which refer to hedonic dimension (not fun/fun, dull/exciting, not delightful/delightful, not thrilling/thrilling, enjoyable/unenjoyable) and five of which refer to the utilitarian dimension (effective/ineffective, helpful/unhelpful, functional/non-functional, necessary/unnecessary, practical/impractical) of consumer attitudes.

3.3.4   PSYCHOPHYSIOLOGICAL

Affective Diary (Ståhl, Höök, Svensson, Taylor, & Combetto, 2009). The Affective Diary consists of a mobile phone (with camera), body sensors, and a Tablet PC. During the day, the sensor armband collects sensor data indicating movement and arousal levels. The mobile phone logging system logs activities on the mobile phone such as Short Message Services (SMSs) sent and received, photographs taken and Bluetooth presence of other mobile phones in the vicinity. Once the person is back at home they can transfer the logged data into their Affective Diary application on the Tablet PC, which produces an aggregated view on a timeline. Sensor data is presented as somewhat ambiguously shaped and coloured figures. To help users reflect on their day, the representations invite interpretation and can be altered: changing the posture or colour of the figures, or scribbling diary-notes on top of the materials. In summary, the tool aims at illustrating how bodies and embodied experiences can shape our recollections and the ways we emotionally reflect on them.

FaceReader (Den Uyl, & Van Kuilenburg, 2005). FaceReader constructs a model of the face and classifies the emotional expression shown on a face, in one of the categories: happy, angry, sad, surprised, scared, disgust, or neutral.

Psychophysiological measurements for evaluating game UX (Mandryk, Inkpen, & Calvert, 2006). The study recorded users' physiological, verbal and facial reactions to game technology, and applied post-processing techniques to correlate an individual's physiological data with their subjective reported experience and events in the game. The physiological measures recorded were: galvanic skin response, which is a linear correlate to arousal and reflects both emotional responses as well as cognitive activity; cardiovascular measures, reflecting emotional activity and stress; respiratory measures, related to emotional arousal or relaxation; and electromyography on the jaw detecting tension. The study was mainly exploratory towards understanding how the body physically responds to enhanced interaction and the authors

recognize the need for a continuation of similar benchmark studies in order to have a valid methodology for objectively evaluating user experience with entertainment technologies.

Physiological responses to different web page designs (Ward & Marsden, 2003). Skin conductivity (SC), blood volume and heart rate (HR) of participants were monitored in various loosely controlled computer-based situations with the aim of obtaining prototypical data to indicate the range and magnitude of the psychophysiological changes that occur in response to HCI events, with the following general observations: (i) at rest, HR slows, there is a steady decrease in SC indicating diminishing activity of the eccrine sweat glands, and an increase in finger blood volume indicating dilation of the peripheral blood vessels, all suggesting lowered levels of arousal; (ii) during non-contentious computer-based activities such as browsing the web, HR, SC and finger blood volume tend to show considerable fluctuation but remain around the same general level, suggesting maintained levels of arousal; (iii) following an unexpected HCI event, participants tend to exhibit increases in HR and SC together with lowered peripheral blood volume, suggesting a sudden increase in arousal typical of an orienting response; and (iv) when using software in more realistic situations, physiological readings are similar to those occurring in non-contentious activities, except that there appear to be more fluctuations. Based on the above a model was proposed, according to which HCI situations can be categorised in relation to the kinds of stress stimuli they present, and this would appear to be reflected in the prototypical SC traces they produce.

Cowley et al. (2016) identify the following categories of psychophysiological methods: internal signals, external signals, and combined signals featuring multimodal signal classification. Internal signals can be measurements of the cardiovascular system, skin conductance, respiration, electromyography, oscillatory electroencephalography, and event-related electroencephalography. External signals can be acquired through pupillometry, eye tracking, video, and audio analysis. Although technology advancements have made possible the recording of psychophysiological metrics and their interpretation, there are still many challenges that need to be addressed as the authors highlight, including to move from linking physiology with certain levels of one state (e.g., high arousal) to multiple states.

Indicative of the challenging task of including psychophysiological measures in HCI research is the review carried out by Maia and Furtado (2016b), which concluded that these measures are still difficult to apply, that each psychophysiological measure represents various emotions, and each emotion is measured by various psychophysiological measures. In summary, the majority of psychophysiological sensor UX studies referred to games or web applications, while the most preferred sensors were the ones located in fingers or chest, and the most common evaluation goal was to assess users' arousal or emotional state.

3.3.5   PLAYFULNESS AND FUN

Pleasure framework (Costello & Edmonds, 2007). The method has been developed for measuring user experience with interactive art. Users, after having used the system, are asked to fill-out a survey sheet with a list of thirteen pleasure categories, and to give a single tick for a category if they had felt mild pleasure and a double tick if they felt strong pleasure. They were also told to cross anything that they felt caused them displeasure. The pleasure categories are: creation, exploration, discovery, difficulty, competition, danger, captivation, sensation, sympathy, stimulation, fantasy, camaraderie, subversion.

The fun toolkit (Read & MaFarlane, 2006). The toolkit has been used to assess children's experience with computer applications. The toolkit comprises the following tools: smileyometer, a discrete Likert type scale with five items represented by smiley faces, ranging from awful (sad smiley) to brilliant (happy smiley); funometer, which is similar to the smileyometer but uses a continuous scale, and has seldom been used due to its resemblance with the smileyometer; again-again table, which asks children to indicate whether they would do an activity again; and the fun sorter, which allows children to rank items against one or more constructs.

3.3.6   GENERIC UX

Experience sampling (Csikszentmihalyi & Larson, 1987). The Experience Sampling Method (ESM) aims at collecting experiences "in situ" and immediately. When prompted, participants get the ESM questions to answer, which may vary according to the study or experiment. Questions may include open questions about location, social context, primary and secondary activity, content of thought, time at which the questionnaire is filled out and a number of Likert scales measuring several dimensions of the respondent's perceived situation including affect, activation, cognitive efficiency, and motivation.

Day reconstruction method (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004). The method combines a time-use study with a technique for recovering affective experiences. More specifically, respondents first revive memories of the previous day by constructing a diary consisting of a sequence of episodes. Then they describe each episode by answering questions about the situation and about the feelings that they experienced, as in experience sampling. The goal is to provide an accurate picture of the experience associated with activities (e.g., commuting) and circumstances (e.g., a job with time pressure). Evoking the context of the previous day is intended to elicit specific and recent memories, thereby reducing errors and biases of recall.

Repertory grid (Hassenzahl, & Wessler, 2000). It is a technique for eliciting and evaluating people's subjective experiences of interacting with technology. In a first step, an individual is presented with a

randomly drawn triad from a group of artefacts that populate the design space and is asked to produce their own constructs, identifying in what way two of the three artefacts are similar to each other and different from the third. The process is repeated until no further novel constructs arise. Then, after having provided their own individual, qualitative constructs, the participant is asked to rate the degree to which each artefact relates to each bipolar construct according to some scale (typically a binary or Likert-type scale).

UXCurve (Kujala, Roto, Väänänen-Vainio-Mattila, Karapanos, & Sinnelä, 2011). The method aims at assisting users in retrospectively reporting how and why their experience with a product has changed over time. Users are given a template for drawing a curve describing how their relationship towards the product has changed from the first time they used it until the current day. The template includes an empty two-dimensional graph area and lines for briefly describing (in text) the reasons for the changes in the curve. The horizontal axis represents the time dimension from the beginning of use to the current moment and the vertical axis represents the intensity of the users' experience. In the middle of the graph area there is a horizontal zero line dividing the area into a positive upper part and a negative lower part. The vertical axis is labelled accordingly with + and – signs. In order to facilitate users in reflecting about the product, the tool asks them to draw one curve for each one of the following: general experience, perceived attractiveness, ease of use, utility, and degree of usage of the product.

UXGraph (Hashizume & Kurosu, 2016). Adopting the approach of UXCurve, UXGraph is a tool to record user satisfaction on a time scale, starting from the use of a product. The differences with UXCurve is that UXGraph is only drawn in terms of satisfaction as a generic measure of quality characteristics, the order of drawing is reversed (first a point is marked indicating user satisfaction, and then the lines are drawn), and also that expectations before use and future predictions were added as episodes.

Experience Recollection Method (Kurosu, Hashizume, Ueno, Tomida, & Suzuki, 2016). It is a qualitative method to measure the dynamic change of satisfaction, whereby participants are asked to indicate what they expected from a given product during the various phases starting from before purchase and ending in the near future and provide a rating from -10 to 10, indicating their satisfaction during each phase. They are also asked to draw the curve of frequency of use since they started to use the product until the near future.

3.3.7 OTHER

Human-Computer trust (Madsen & Gregor, 2000). A psychometric instrument designed to measure human-computer trust, through five constructs: perceived reliability, perceived technical competence,

perceived understandability, faith, and personal attachment. Each construct comprises five questions, to which participants have to indicate their agreement or disagreement.

Personal meaning maps (Blythe, Robinson, & Frohlich, 2008). Users are asked to list as many words as they can which they associate with a key word or phrase. It has been used for pre and post museum visit interviews and has been useful for getting people to articulate their perceptions of particular subjects.

3.3.8 UX EVALUATION FRAMEWORKS

Although UX is a relatively young field and a challenging subject to define and understand, several frameworks have already been proposed for its evaluation. However, most of the proposed frameworks in literature are mainly conceptual, or focus on specific contexts and application domains.

A conceptual framework that has been proposed as a medium to design and evaluate UX is that of Hellweger and Wang (2015). According to the framework, UX is affected by six prime elements, namely context, usability, product properties, cognition, needs, and purpose. UX produces the following six prime elements that should be pursued and assessed: memorability, ubiquity, perception, emotional state/mood, engagement, educational value. For each of the twelve prime elements the framework describes sub-elements that should be taken into account, resulting in 86 attributes, which however are not further defined as to how they can be assessed (e.g., efficiency, behaviour patterns, perceived quality, etc.).

Adikari, McDonald, and Campbell (2011) introduced a model oriented towards UX assessment according to which UX is measured by pleasure, usability/comfort, and trust, which are further analysed into eight main usability attributes, namely: satisfaction, functional correctness, efficiency, error tolerance, memorability, flexibility, learnability, and safety. To this end, the authors have employed the following measurements in an experimental study comparing a reference system to a newly developed one: visual appeal, pleasure in interaction, meeting expectations, less frustration, less confusing terminology, overall experience of using the system, completing tasks correctly, available facilities to meet user needs, completing tasks quickly, achieving expected outcome, completing tasks easily, causes fewer errors, clear error messaging for invalid conditions, error messages that inform which actions to take, easiness to remember task steps, needing to memorise task steps, needing to access help documents, alternative ways to perform tasks, navigating back/forward between task steps, user ability to cancel an operation, ease of learning system operation, clarity of system status, knowing what to do next during navigation, fewer keystrokes, security measures to protect personal information, and security measure to protect user transactions.

Miki (2014) proposed an integrated evaluation framework of usability and UX, according to which usability and UX can be measured during use and after use. During use measures are further analysed to objective measures of usability and more specifically effectiveness and efficiency, as well as subjective measures of UX, namely perceived quality, perceived value and satisfaction. After use measures pertain only to UX and include complaints and customer loyalty.

Recognizing the problem that elements of conceptual frameworks are challenging to capture from a data collection perspective, Thayer and Dugan (2009) introduced a five stages methodological framework that can be used in a usability study. The first stage aims to measure the "anticipating" UX element through a pre-experienced interview and pre-test questionnaire to obtain rich, qualitative details about the participant's current experience with similar products or product categories. The second stage, to assess the "anticipating, "connecting", and "interpreting" UX dimensions involves post-task questionnaire to gather baseline quantitative data about participant expectations and to assess expectations for one or two competing products as well. The third stage, targeting the "interpreting", and immediate reflecting" UX attributes is addressed through performance data, think-aloud protocol and evaluator observations in order to gather feedback on the specific features or areas of the product that relate to the experience goals. The fourth stage aiming to measure the "immediate reflecting", "future reflecting" and "recounting" UX elements involves post-test questionnaire and satisfaction questionnaire to gather participant feedback about the total user experience. Finally, the fifth stage related to the "appropriating" UX dimension involves post-experience interview to gather as much relevant information as possible about what might motivate participants to purchase or use the product. Although the proposed framework moves beyond the conceptual to a methodological approach, it does not provide concrete metrics that should be pursued at each stage.

A practitioner-oriented framework for analysing user experience was proposed by McCarthy, Wright, and Meekison (2005). The framework identifies four threads of experience and six sense-making processes. The compositional thread is concerned with relationships between the parts and the whole of an experience. The sensual thread of experience is concerned with the user's sensory engagement with a situation, which orients towards the concrete, palpable, and visceral character of experience. The emotional thread refers to judgments that ascribe to other people and things important to the user's needs and desires. The spatio-temporal thread of an experience may distinguish between public and private space, and recognise comfort zones and boundaries between self and other, or present and future. The sense-making processes are: anticipation, connecting, interpreting, reflecting, appropriating and recounting. The proposed framework has been used for assessing internet shopping, involving three data collection procedures: user interview, participants' own notes of their experiences and debriefing. The

framework was given as a tool to participants, asking them to organize their notes according to the framework. As mentioned by the authors, the specific framework is open and not prescriptive, mainly providing the basis for developing the methods to proceed to UX assessment.

Lachner, Naegelein, Kowalski, Spann, and Butz (2016) developed the QUX tool to support a common organizational understanding of a product's UX and the selection of further in-depth UX evaluations. The authors identified 28 consolidated UX characteristics under seven main clusters, as follows:

1. Emotion: satisfaction, pleasure.
2. Design: interface, aesthetics.
3. Content: information, effectiveness.
4. Technology: Efficiency, functionality, ease of use, performance, usability, utility, security, control, learnability.
5. Result: quality of outcome, error-free.
6. Further Disciplines: brand history, advertisement, price, user expectation, user customization, user self-realization, group affiliation, social connectivity.
7. Environment: memorability, time context, location context.

This list of UX characteristics was further analysed and clustered under the categories of look, feel, and usability and arrived at 9 UX dimensions, each explored through three questions that a user will have to answer. The UX dimensions studied were: appealing visual design (look), communicated information structure (look), visual branding (look), mastery (feel), outcome satisfaction (feel), emotional attachment (feel), task effectiveness (usability), task efficiency (usability), stability and performance (usability).

Focusing on cross-platform user experience Wäljas, Segerståhl, Väänänen-Vainio-Mattila, and Oinas-Kukkonen (2010) also proposed an evaluation framework. The framework conceptualizes a structured set of distinct, designable characteristics of cross-platform systems that essentially influence UX, and the respective main elements of cross-platform service user experience. According to the framework, central elements of cross-platform service UX include fit for cross-contextual activities, flow of interactions and content, and perceived service coherence.

More recently, user experience in complex systems has been studied across three dimensions: instrumental, which refers to the system's ability to have an effect on the environment for which it has been designed, psychological, referring to the effect of the system on the user, and communicative, which refers to the effect that the system has on the community of the users (Savioja, Liinasuo, & Koskinen, 2014). The approach was applied in three studies of the UX of control room systems, and the various

dimensions of UX were assessed by users through a questionnaire, featuring five questions for each category. The results of the studies highlighted that UX is a significant indicator of quality in use when professional users' experiences, which are embedded in the inner characteristics of the work and not always observable by external evaluators, can be brought to the process of designing new systems. Also it was observed that the three operationalisations of UX seemed to work well, yet future work in the field should move beyond simple questionnaires.

A framework for measuring user experience of interactive online products was introduced by Schulze and Krömker (2010). The framework suggests three main components of user experience, namely emotion, motivation, and reflection, and furthermore identifies influencing factors organised under two categories: basic human needs and product qualities (Figure 23). Human needs factors include relatedness, influence/popularity, stimulation, competencies, security and autonomy. Product qualities on the other hand include utility, usability, visual attraction, and hedonic quality. The framework has been used for the evaluation of a new web-community concept, collecting need fulfilment data through likert scale questionnaire, observation notes, and interviews, as well as product qualities data through semantic differential scales, and momentary and emotion data through observation notes, interviews and retrospective questionnaires.
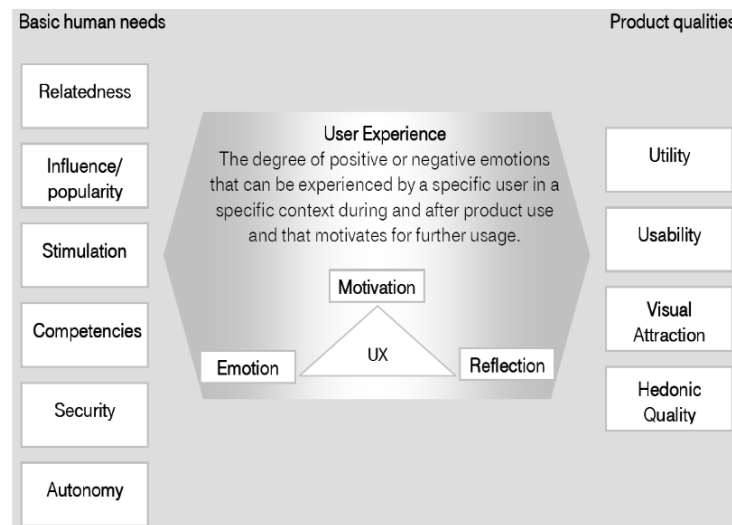
Figure 23. Framework for measuring user experience of interactive online products

In the context of m-learning environments, the MUUX-E framework emphasises usability and user experience in mobile educational contexts (Harpur & De Villiers, 2015). The framework (Figure 24) identifies the following distinct categories: general interface usability, web-based learning, educational usability, m-learning features, and user experience.
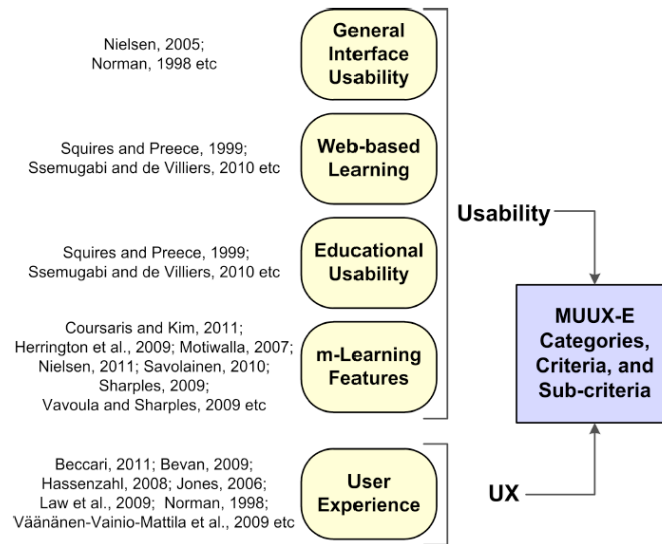
Figure 24. The MUUX-E framework for usability and user experience evaluation in mobile educational contexts

Moreover, it includes 31 categories of criteria and specific criteria under each category for evaluating features of m-learning environments, along the aforementioned general categories, as follows:

A. General Interface Usability

    1. Visibility of system status

    2. Match to the real world

    3. Learner control and freedom

    4. Consistency: adherence to standards

    5. Prevention of usability-related errors

    6. Recognition rather than recall

    7. Aesthetics and minimalism in design

    8. Recognition and recovery from errors

    9. Help and documentation

B. Web-based learning

    10. Simple, well-organised navigation

    11. Relevant pedagogical site content

    12. Information easily accessible

    13. Suitable course content of a high quality

    14. Easy-to-use system, called easiness

    15. Excellent video and digital media

C. Educational usability

16. Clarity of goals, objectives and outcomes

17. Effectiveness of collaborative learning

18. Error recognition, diagnosis and recovery

19. Feedback, guidance and assessment

D. M-Learning Features

20. Handheld devices and technology

Optimum technology, device capability, interface, input mode, and system capabilities; provision of communication channels

21. Contextual factors (pragmatic)

Physical, visual, and auditory environment; nature of the task or activity; fixed or adjustable goals; characteristics of the working environment; context awareness

22. User-centricity (pragmatic)

Support for personalised learning, customisation, experimentation and exploration; specification of user requirements; self-sufficiency, ownership and control; clear student-centric material; longer time for doing tasks; encouragement of active learning

23. Flexibility

An adaptable environment; lesson information viewable in any order; system can be used anytime and anywhere

24. Interactivity

Navigational fidelity; multimedia components with high quality lessons and exercises; synchronous and asynchronous communication and collaboration; simple and easy to use system

E. User experience

25. Emotional issues

Affect; excitement; interest; attitude; joy; well-being; fun; beliefs

26. Contextual factors (hedonic)

User knowledge; user experience and goals; flexibility; time; situation; individual needs

27. User-centricity (hedonic)

Support for personal approaches to learning; personalised learning format; ability to customise material; personal growth potential

28. Social value

Social self-expression; media sharing; synchronous and asynchronous interaction

29. Needs

Autonomy; competence; relatedness; stimulation; security; competition

30. Appeal

New impressions; curiosity; insights; visual power; audio interactivity; aesthetic factors

31. Satisfaction

Pleasure; cognitive likeability; trust; achievements; motivation, goals

MUUX-E has been used for the evaluation of a mobile learning application for a Software Engineering module, involving expert-based reviews and a questionnaire survey.

A preliminary framework was recently proposed for designing and evaluating UX in Mobile Augmented Reality Applications (Irshad & Rambli, 2015). The framework identifies the following important UX components of such applications: information content, functionality and performance, presentation, and interaction. Prerequisites of good information content is to be rich, lively, and relevant, to ensure the stability of 3D content, and provide user with control over the 3D content. In terms of functionality and performance, the system should be reliable and useful, provide user support, be relevant to the reality, use embedded content, and personalization of specific content and versatile features. Good presentation involves user friendly menus, ease of use, straightforward interaction, and ease of learning. Finally, interaction should be positive, simple and intuitive.

In addition, the concept of Quality of Experience (QoE) has constituted a research area for developing frameworks. Being linked very closely to the subjective perception of the end user, QoE is supposed to enable a broader, more holistic understanding of the qualitative performance of networked communication systems and thus to complement the traditional, more technology-centric Quality of Service (QoS) perspective (Schatz, Hoßfeld, Janowski, & Egger, 2013). QoE stems from engineering and reflects the fact that quality is a fundamental property in the evaluation of a system, service or application (Raake & Egger, 2014). Although QoE and UX are similar and related concepts, several differences exist among them (Wechsung & De Moor, 2014): (i) coming from different origins (Telecommunications vs. HCI), UX does not take into account economic aspects, whereas QoE is more close to the concept of customer experience and addresses economic factors; (ii) the driving force of UX is human-centred, while QoE is mainly system- and technology-centred; (iii) different evaluation approaches are pursued, with QoE measurements mainly based on quantitative quality evaluations and numerical expressions (iv) the evaluation target of UX is the overall experience, while QoE is mainly focused on the perception of quality.

To this end, Wu et al. (2009) proposed a framework for measuring the quality of experience in distributed interactive multimedia environments, aiming to model, measure, and understand quality of experience and its relationship with the traditional QoS metrics (Figure 25). In summary, the framework identifies three main components: (1) environmental influences, which include the variables of interactivity speed,

interactivity range, interactivity mapping, vividness breadth, vividness depth, temporal consistency, and spatial consistency; (2) cognitive perceptions, which include the variables of flow concentration, flow enjoyment, telepresence, perceived usefulness, and perceived ease of use; and (3) behavioural consequences, which include the variables performance gains, exploratory behaviours, and technology adoption.
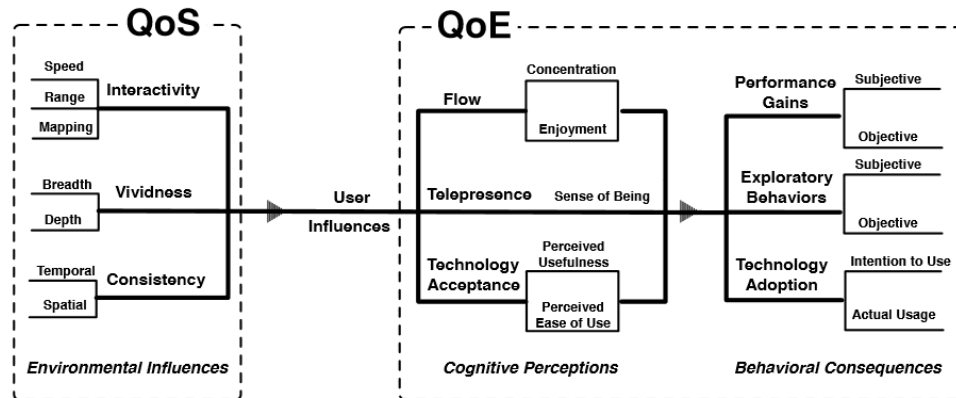

Figure 25. Framework for measuring the quality of experience in distributed interactive multimedia environments

An integrated framework including both the perspectives of QoE and UX is proposed by Geerts et al. (2010), consisting of four main components: user, (IT) product, use process and context. At the user component, a theoretical distinction between the person and the user (person using a specific IT) is made. At the product level, technical characteristics of the product are studied (e.g., application, network, device, and context-sensor related), as well as economic aspects (product and brand strategy, pricing strategy, positioning of the product in the market, and market segmentation and identification of target groups) and product characteristics. Product characteristics can be instrumental, such as utility, efficiency, functionality, effectiveness, usefulness and ease of use, or non-instrumental, encompassing aesthetic qualities, symbolic qualities, and motivational qualities. At the use process framework component, the temporal dimension of interaction is studied, examining macro-temporality, micro-temporality, non-use, and abandoned use. Finally, at the context component, contextual factors are studied and namely socio-cultural context, situational context and interaction context. This framework constitutes a very interesting approach towards the fusion of two relevant yet different theories, these of UX and QoE, however it constitutes a conceptual framework, not providing concrete metrics or methods.

Another framework for the evaluation of Quality of Experience (QoE) in a mobile, testbed-oriented Living Lab setting was proposed by DeMoor et al. (2010). QoE is studied from an interdisciplinary perspective in relation to both QoS and UX. The proposed framework consists of a highly distributed system allowing

measurements on the device (handset-based approach), measurements in the network (network-based approach), and data processing in the back-end. Central component of the framework's architecture is the mobile agent, which is composed of three entities: (i) the QoS monitoring entity, which is in charge of measuring the objective, technical parameters related to the device, the infrastructure, the network, and the application under use; (ii) the contextual monitoring entity, which deals with the determination of the context of the application usage in terms of location, mobility, sensors and other running applications; and (iii) the experience monitoring entity, which interacts with the user by gathering explicit feedback in the form of questionnaires and pictographic feedback.

HAPPYNESS is a QoS assurance framework, exploiting emotional information as a key element in providing personalized context-aware software service, thus enhancing the UX (Condori-Fernandez, 2017). The framework is composed of three main modules: (i) the actionable emotion detection component which aims to detect negative emotions that are actionable (i.e., emotions expressed by a user within the same time interval in which a service is also delivered), (ii) the context-dependent QoS assessment component, which employs socio-technical QoS metrics (e.g., performance, adaptability, usability) to measure the detected services at runtime, and (iii) the emotion aware QoS assurance controller, which is responsible for monitoring services when an actionable emotional level is detected and calibrating QoS levels of service contracts.

In summary, the concept of user experience, although it has been introduced since 2000, is still a challenging topic both in its conception and evaluation. Kashfi, Nilsson, and Feldt (2016) identified eleven challenges related to UX, namely lack of consensus on definition and construct of UX; lack of consensus on the value of UX; low industrial impact of UX models, tools, and methods; more focus placed on objectively measurable aspects; difficulties in engineering UX-related requirements; more focus placed on testing functionalities and usability than UX evaluation which includes emotions, is more holistic, and time-dependent; lack of consensus on UX-related competences and responsibilities; late focus on UX in projects; communication and collaboration gap between UX and non-UX practitioners; customers' resistance to the cost of UX practices; and low user involvement.

Table 8 lists the main attributes involved in the aforementioned frameworks, in total 79, organized under 11 categories.

Table 8. UX attributes proposed by evaluation frameworks

Category	Attributes	References
Usability	Effectiveness, Efficiency, Satisfaction, Functional correctness, Error tolerance, Memorability, Flexibility, Learnability, Accessibility, Reliability, Stability and performance, Visibility of system status, Match to the real world, Consistency, Intuitiveness, Error prevention, Recognition rather than recall, Help and documentation, User control and freedom, User-centricity, Ease of use, Simplicity, Safety	Adikari et al. (2011), Geerts et al. (2010), Harpur and De Villiers (2015), Hellweger and Wang (2015), Irshad and Rambli (2015), Lachner (2016), Miki (2014), Savioja et al. (2014)
Look	Aesthetics, Communicated information structure, Visual branding	Lachner (2016)
Content	Simple and well-organised navigation, Information easily accessible, Relevant, Rich, Lively, Stable, Personalization	Harpur and De Villiers (2015), Irshad and Rambli (2015)
Engagement	Motivation, Symbolic qualities, Emotionally involving, Being in gear, Awareness, Appeal, Voluntariness in use, Entertainment, Mastery	Geerts et al. (2010), Hellweger and Wang (2015), Lachner (2016), Schulze and Krömker (2010)
Emotional state / mood	Surprise, Joy, Fun, Pride, Affective, Excitement, Well-being, Interest	Harpur and De Villiers (2015), Hellweger and Wang (2015)
Ubiquity	Acceptance, Sub-consciousness, Visceral ubiquity	Hellweger and Wang (2015)
Perception	Perceived quality, Sensation, Hedonic, Aesthetic experience, Resources, Perceived value, Suitability for self, Utility, Relatedness, Social value, Influence / Popularity, Pricing strategy, Positioning of the product in the market	Geerts et al. (2010), Hellweger and Wang (2015), Miki (2014), Savioja et al. (2014), Schulze and Krömker (2010)
Context	Socio-cultural context, Situational context, Interaction context	Geerts et al. (2010)
Before use	Users' experience with similar products, User expectations, Expectations for competing products	Thayer and Dugan (2009)
After use	Complaints, Customer loyalty, Purchase motivators, Usage motivators	Miki (2014), Thayer and Dugan (2009)

Cross-platform	Fit for cross-contextual activities, Flow of interactions and content, Perceived service coherence	Wäljas et al. (2010)

From the above it is evident that given the complexity of the concept, the proposed evaluation frameworks have in their majority remained conceptual, and therefore do not bind concepts with concrete methods and metrics that should be used to assess the various facets of UX. Furthermore, it is imperative that frameworks are accompanied by tools facilitating the association of observed performance metrics with metrics related to users' emotions and satisfaction, providing insights and information about the usage of a product or service over the product usage timespan (e.g., single time, long-term). The aforementioned challenges will be addressed by the UXAmI framework and its tools, in the context of systems, products, and applications used in Ambient Intelligence environments.

3.4   EVALUATION OF ADAPTIVE SYSTEMS

An important attribute of Ambient Intelligence environments is their capability to adapt according to the people or objects that reside in them. However, the notion of systems which can adapt according to various requirements and criteria, or even upon request, is not novel. Adaptation can be either triggered by the user (e.g., through customisation) or by the system (self-adaptive systems). Self-adaptation can be further classified to adaptability or adaptivity (Stephanidis, Paramythis, Akoumianakis, & Sfyrakis, 1998): adaptability refers to self-adaptation which is based on knowledge (concerning the user, the environment, the context of use, etc.) available to (or, acquired by) the system prior to the initiation of interaction, and which leads to adaptations that also precede the commencement of interaction; adaptivity on the other hand refers to self-adaptation which is based on knowledge (concerning the user, the environment, the context of use, etc.) that is acquired and / or maintained by the system during interactive sessions (e.g., through monitoring techniques), and which leads to adaptations that take place while the user is interacting with the system. A tool aiming to assist experts in quantifying the adaptivity and adaptability of a system is AnAmeter (Bernard, Marfisi-Schottman, & Habieb-Mammar, 2009). The tool is based on assessing adaptation aspects (presentation, control, abstractions) across a number of adaptation factors (user, platform, environment, activity) and sub-factors. The evaluator has to complete two such grids one for the system's adaptability and one for its adaptivity. However, although the tool can assist in quantifying adaptation, it does not support evaluation of the usability of adaptations. Therefore, one could assess how much adaptation is supported by the system, but not how successful this adaptation is.

The adaptation decisions that are applied before or during the user's interaction with an adaptive system influence the overall user experience. Therefore, an important aspect that should be evaluated in the context of adaptive interfaces is the impact that working in an adaptive interface can have on the user's breadth of experience (i.e., their overall awareness of features in the interface) (Findlater & Gajos, 2009). For instance, an adaptive mechanism designed to improve one aspect of the interaction, often increases effort along another dimension, such as cognitive or perceptual load (Findlater & Gajos, 2009). On the other hand, since the adaptation process often takes time (as the system needs to learn about the user's goals, knowledge or preferences, etc., before adaptation can take place), the observation of any effects of adaptivity may require long-term, or even longitudinal studies, or be based on evaluation designs that explicitly account for that factor (Paramythis, Weibelzahl, & Masthoff, 2010). During the evaluation of adaptive systems, it is therefore important to select the appropriate evaluation method, a task which is more complicate than the evaluation of non-adaptive systems using traditional usability and/or user experience evaluation methods. Important concerns that should be addressed during the evaluation of adaptive systems include, but are not limited to (Mulwa, Lawless, Sharp, & Wade, 2011): the adaptive features of the system should be distinguished from the general usability of the system, it should clarified what constitutes a useful or helpful adaptation, and it is difficult to clarify the origin for a measured effect (i.e., was it the adaptation that caused the effect or another system attribute, such as its usability).

Recognizing the need for an evaluation method that would guide the authors of an adaptive system, Brusilovsky, Karagiannidis, and Sampson (2004) apply the layered evaluation framework proposed by Karagiannidis, and Sampson (2000), according to which the success of adaptation is addressed at two distinct layers (Figure 26): (a) the user modelling, and (b) the adaptation decision making. At the first layer, only the User Modelling (UM) process is being evaluated, aiming to answer questions such as: "are the conclusions drawn by the system concerning the characteristics of the user-computer interaction valid?" or "are the user's characteristics being successfully detected by the system and stored in the user model?". Brusilovsky et al. (2004) propose two potential methods for evaluating the UM layer: (i) through user observation, where experts monitor users as they work with the system and compare their expert opinions with the conclusions stored in the user model and (ii) by users themselves who can evaluate whether the conclusions drawn by the system at any particular instance reflect their real needs. At the second layer, only the adaptation decision making is being evaluated, with the aim to answer the question "are the adaptation decisions valid and meaningful, for the given state of the user model?". This phase can be evaluated through user testing based on specific scenarios, according to the adaptation evaluated (e.g., to evaluate a knowledge-based adaptation the user knowledge can be assessed by direct testing). In addition, this layer can be evaluated through the cooperation of experts with users who will assess

whether specific adaptations contribute to the quality of interaction. It is important to note that the layered evaluation approach did not intend to replace current evaluation practices, since the separate evaluation layers can make use of existing evaluation techniques (Karagiannidis & Sampson, 2000).
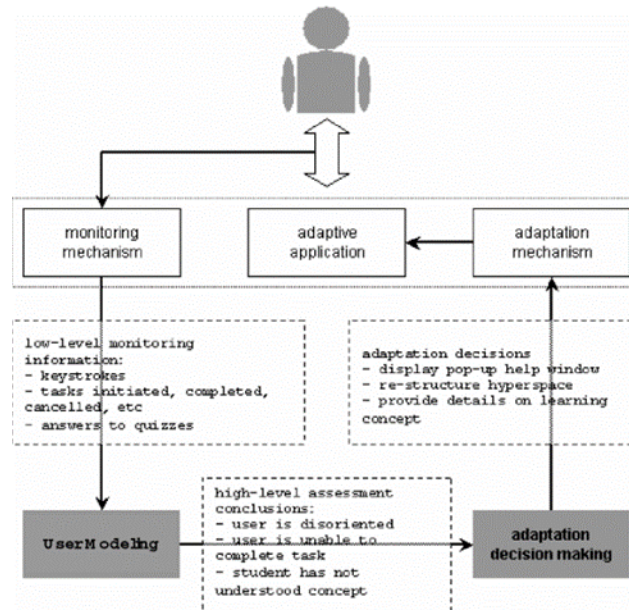

Figure 26. Adaptation decomposition (Brusilovsky et al., 2004)

Although the two-level layered evaluation model is perhaps the simplest layered evaluation model, its contribution to the field is important, as it makes clear the benefits of layered evaluation in comparison to traditional evaluation approaches:

- Traditional evaluation approaches can be used to report a success, however they cannot adequately guide an adaptive system's development process in case of failure. Since adaptive behaviour is evaluated as a whole, the reasons behind unsatisfactory adaptive behaviour are not evident, and the ways to improve the system are not clear.

- Evaluating a system as a whole requires building the whole system before it can ever be evaluated. Layered evaluation however, supports iterative design approaches and can assist in fixing problems earlier in the development lifecycle.

- Traditional evaluation provides no feedback about performance of different system components, thus successful design practices cannot be easily re-used across different applications and services.

Building on the notion of layered evaluation as defined by Karagiannidis and Sampson (2000), Herder (2003) proposed the utility-based evaluation of adaptive systems. According to this method, the evaluation can be seen as a utility function U that maps a system, given some user context, to a quantitative representation of user satisfaction or performance. In more details, the model assumes there is a utility function U_1 that maps the interaction assessment and the resulting user model to a real number that represents its correctness and a utility function U_2 that maps a system, given some user model, to a real number that represents user satisfaction or performance. Then, the whole utility function can be defined as $U = U_1 U_2$.

Weibelzahl and Lauer (2001) proposed a layered evaluation framework for case-based retrieval (CBR) systems, consisting of the following layers: (i) correctness of input data acquisition, (ii) correctness of inference, (iii) appropriateness of adaptation decisions, (iv) change of system behaviour when the system adapts, (v) change in user behaviour when the system adapts, (vi) change and quality of total interaction. An updated version of this model addressing adaptive systems in general (Weibelzahl, 2002) proposed four layers (Figure 27), namely: (i) evaluation of input data, which refers to the evaluation of the reliability and external validity of input data; (ii) evaluation of inference, which assesses the validity of inferences regarding the user properties; (iii) evaluation of adaptation decision, aiming to figure out whether the chosen adaptation decision is the optimal one, given that the user properties have been inferred correctly; (iv) evaluation of total interaction, which assesses the whole system in a summative evaluation, in terms of system behaviour (frequency of adaptation, frequency of adaptation types), user behaviour (task success, performance), usability and behavioural complexity.
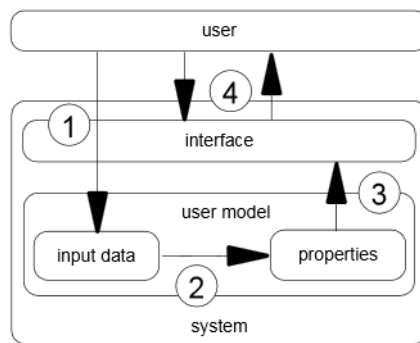

Figure 27. Four-layered evaluation model

Weibelzahl (2002) besides defining the evaluation model, proposes evaluation criteria and methods per layer (summarized in Table 9) and illustrates how the evaluation model was used for the evaluation of an adaptive web-based learning course.

Table 9. Evaluation criteria and methods for each of the four layers of the layered evaluation model (Weibelzahl, 2002)

	Criteria	Methods
Evaluation of input data	• objectivity of data assessment • retest-reliability • split-half reliability	• early exploratory studies: objectivity, reliability and stability user assessment method • controlled evaluation with users: occurrence of adaptation trigger under experimental conditions (e.g., high workload)
Evaluation of inference	• accuracy of system predictions • congruency of assumed user properties and external test (e.g., $\chi2$-test) • correct categorization of users • stability of user model	• controlled evaluation with users: comparison of assumed user properties with external test, expert rating, or self- assessment; comparison of system predictions with actual user behaviour • controlled evaluation with hypothetical users: observation of user model in dependence of different hypothetical behaviours of users • experience with real world use: observation of change in user model; comparison of assumed user properties with external test, expert rating, or self- assessment; comparison of system predictions with actual user behaviour; registration of frequency of adaptation
Evaluation of adaptation decision	• accuracy, precision and recall • amount of required help • amount of requested material • behavioural complexity • budget spent • computation time • difficulty of learning • duration of interaction • fixation times • frequency of adaptation • number of communications • number of errors • number of navigation steps • overall impression • rating of solution • quality	• controlled evaluation with users and experience with real world use: comparison of different adaptation decisions in terms of efficiency, effectiveness, and usability

	• similarity of expert rating and system decision • subjective rating of effect • system preference • task success • usability questionnaire • user satisfaction	
Evaluation of total interaction	• accuracy, precision and recall • amount of required help • amount of requested material • behavioural complexity • budget spent • computation time • difficulty of learning • duration of interaction • fixation times • number of communications • frequency of adaptation • number of errors • number of navigation steps • overall impression • rating of solution quality • similarity of expert rating and system decision • structural information measures • subjective rating of effect • system preference • task success • usability questionnaire • user satisfaction	• controlled evaluation with users: observation of system and user behaviour under different conditions; • experience with real world use: observation of system and user behaviour for different user groups in real world settings in terms of absolute efficiency, effectiveness, and usability

Based on the initial model for the evaluation of adaptive Case-Based Reasoning systems Magoulas, Chen and Papanikolaou (2003) proposed a model integrating heuristic evaluation criteria into the layered evaluation model. The criteria proposed are specific for learning activities, with the aim to assess the impact of usability on user behaviours and consequently their impact on adaptation, at an early design stage of the adaptive system.

An alternative model has been proposed by Paramythis, Totter, and Stephanidis (2001), focusing on the different components involved in the adaptation process. In more detail, the following components of adaptation are defined (Figure 28-left): (i) interaction monitoring, which refers to the facilities that are intended to capture the exchanges between the user and the UI; (ii) interpretation / inferences, which refers to the parts of the Adaptive User Interface (AUI) responsible for interpreting the information that is made available through interaction monitoring, in order to update the models maintained by the system (iii) explicitly provided knowledge, which deals with the information about the users' characteristics, plans, tasks, context, etc. which is explicitly provided to the system, typically by users themselves; (iv) modelling, which refers to explicit or implicit representations of the users, their plans regarding a particular interaction session, the tasks that can be performed with the system, etc.; (v) adaptation decision making, which deals with the part(s) of the AUI responsible for deciding upon the necessity of the adaptations and their required type, given a particular interaction state; (vi) applying adaptations, which refers to the actual introduction of adaptations in the user-system interaction; (vii) transparent models and adaptation rationale, which refers to the particular case of AUIs that enable users to review the models maintained by the AUI, or the rationale that underlies the adaptation decisions made by the system; and (vii) automatic adaptation assessment, referring to the run-time assessment of the success of the applied adaptations. Based on these, the framework proposes evaluation modules, comprising one or more of the aforementioned adaptation components, which can be evaluated individually and in combinations (Figure 28-right). Table 10 summarises these modules, their evaluation goals, the evaluation criteria and the evaluation methods that can be applied for each module.
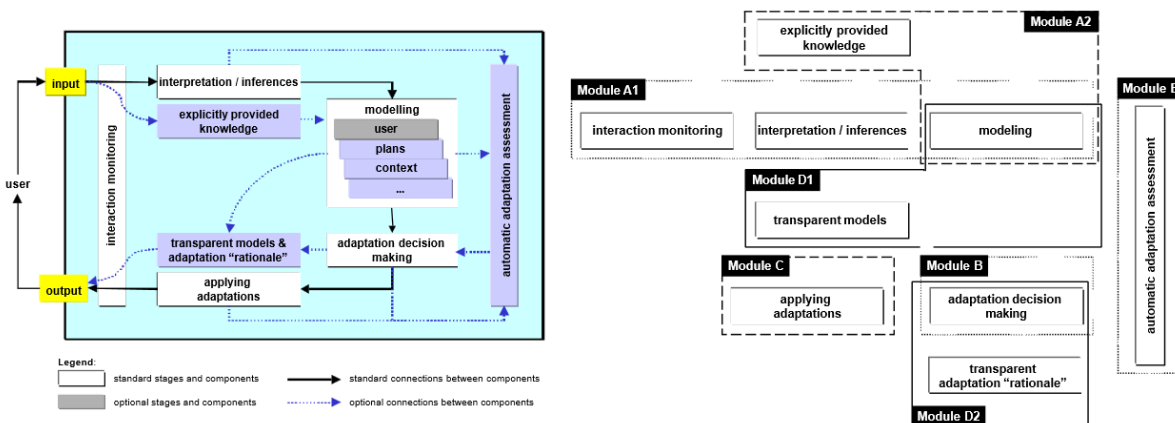

Figure 28. (Left) High level model of adaptation in Adaptive User Interfaces (AUIs); (Right) the correspondence between evaluation modules and AUI model components (Paramythis et al., 2001)

Table 10. Evaluation goals, criteria and methods for the modules of the Paramythis et al. (2001) layered evaluation model

Module A1 [Components: interaction monitoring, interpretations / inferences, modelling]	
Goal	To ensure that the models derived by the system through dynamic interaction assessment are "optimal"
Criteria	• correctness of the interpretations / inferences • comprehensiveness of the model • redundancy of the model • precision of the model • sensitivity of the modelling process
Methods	• self-reporting methods for models that directly or indirectly involve users • expert-based evaluation for correctness of inferencing / interpretations, and comprehensiveness and redundancy of the model
Module A2 [Components: explicitly provided knowledge, modelling]	
Goal	Similar to the preceding one
Criteria	• comprehensiveness of the model • redundancy of the model • precision of the model sensitivity of the modelling process • transparency of the process • overhead on the main interaction
Methods	• self-reporting methods for models that directly or indirectly involve users
Module B [Components: adaptation decision making]	
Goal	To ensure that the adaptation decisions made by the respective component are "correct"
Criteria	• necessity of adaptation • appropriateness of adaptation • acceptance of adaptation
Methods	• formative evaluation methods to assess the necessity and appropriateness of adaptations, based on past empirical findings • user-based evaluation to assess the overall acceptance of an adaptation decision, which are experienced in "real time" • expert-based evaluations or the necessity and appropriateness of adaptations
Module C [Components: applying adaptations]	
Goal	Complementary to the goal of module B (ensure that the adaptation decisions made by the respective component are "correct")

Criteria	• timeliness of adaptation • obtrusiveness of the adaptation • user control over adaptation
Methods	• User-based evaluations, where users are "immersed" in realistic tasks or interaction situations
Module D1 [Components: modelling, transparent models]	
Goal	To ensure that the users' perception of the maintained models matches the actual state of the models
Criteria	• completeness of the presentation • coherence of the presentation • rationality of the presentation
Methods	• End users and experts involved in the evaluation of completeness and coherence • User-based evaluation for rationality, where feedback is requested during the interaction
Module D2 [Components: adaptation decision making, transparent adaptation "rationale"]	
Goal	Similar to D1, with the difference that the user is not presented with a model, rather with the rationale underlying each adaptation
Criteria	• coherence of the adaptation • causality of the rationale
Methods	• End users and experts involved in the evaluation of completeness and coherence • User-based evaluation for rationality, where feedback is requested during the interaction and a (almost) fully functional prototype is available
Module E [Components: automatic adaptation assessment]	
Goal	To ensure that the system shares the same views as the users with regards to the "success", or "failure" of adaptations
Criteria	• optimality of modifications
Methods	• user based evaluation regarding specific adaptations and their effects on interaction, compared to the system's view of such adaptations

An alternative model for Interactive Adaptive Systems (IAS), taking into account all the previous layered evaluation models was proposed by Paramythis, Weibelzahl, and Masthoff (2010), according to which the main levels of adaptation are: (i) collection of input data, which refers to the assembly of user interaction data, along with any other data (available, e.g., through non-interactive sensors) relating to the interaction context; (ii) interpretation of the collected data, where the raw input data previously collected acquire

meaning for the system; (iii) modelling of the current state of the "world", which refers to derivation of new knowledge about the user, the interaction context, etc., as well as the subsequent introduction of that knowledge in the "dynamic" models of the IAS; (iv) deciding upon adaptation, in which the IAS decides upon the necessity of, as well as the required type of, adaptations, given a particular state of the "world", as expressed in the various models maintained by the system; (v) applying (or instantiating) adaptation, which refers to the actual introduction of adaptations in the user–system interaction, on the basis of the related decisions. Table 11 summarises the evaluation goals, criteria and methods that can be applied for each layer of this model.
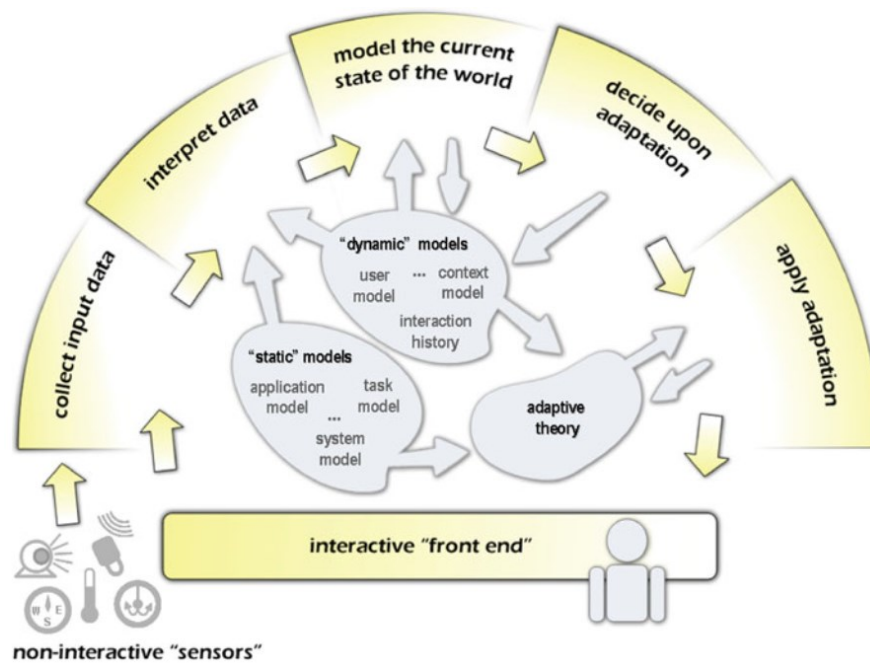
Figure 29. Adaptation decomposition for the layered evaluation model of Paramythis et al. (2010)

Table 11. Evaluation goals, criteria and methods for the layers of the Paramythis et al. (2010) layered evaluation model

Collection of input data	
Goal	Check quality of raw input data
Criteria	• Accuracy • Latency • Sampling rate • Etc.

Methods	- Data mining - Play with layer - Simulated users - Cross-validation
Interpretation of the collected data	
Goal	Check that input data is interpreted correctly
Criteria	- Validity of interpretations - Predictability (of system's interpretations), - Scrutability (of system's interpretations)
Methods	- Data mining - Heuristic evaluation - Play with layer - Simulated users - Cross validation
Modelling of the current state of the "world",	
Goal	Check that constructed models represent real world
Criteria	- Validity of interpretations or inferences - Predictability (of system's modelling behaviour) - Scrutability (of user model) - Comprehensiveness, conciseness, precision, sensitivity (secondary evaluation criteria)
Methods	- Focus group - User-as-wizard - Data mining - Heuristic evaluation - Play with layer - Simulated users - Cross-validation
Deciding upon adaptation	
Goal	Determine whether the adaptation decisions made are the optimal ones
Criteria	- Necessity of adaptation - Appropriateness of adaptation - Subjective acceptance of adaptation - Predictability (of system's adaptive behaviour) - Scrutability (of system's behaviour) - Breadth of experience

Methods	- Focus group - User-as-wizard - Heuristic evaluation - Cognitive walkthrough - Simulated users - Play with layer - User test
Applying adaptation	
Goal	Determine whether the implementation of the adaptation decisions made is the optimal one
Criteria	- Usability criteria - Timeliness - Unobtrusiveness - User control - Acceptance by user - Predictability (of system's adaptive behaviour) - Breadth of experience
Methods	- Focus group - User-as-wizard - Heuristic evaluation - Cognitive walkthrough - User test - Play with layer

Studying the concept of evaluation of adaptation, several efforts have attempted to highlighted which attributes are usually evaluated and how. Such a literature review of approaches towards the evaluation of adaptive and adaptable systems (Van Velsen, Van Der Geest, Klaassen, & Steehouder, 2008) identified that the most common attributes evaluated are usability, perceived usefulness, appropriateness of adaptation, intention to use, and user behaviour. Moreover, the methods most usually employed are questionnaires, interviews, and data log analysis. Based on the literature review findings, a model was proposed for the evaluation of personalized systems, which is based on the user-centred approach, studying the various attributes in four phases: when no system is available with the aim to support design decisions, when a low-fidelity and a high-fidelity prototype is available so as to detect problems, and once the full system is implemented to verify quality. The attributes evaluated, as well as the methods suggested for each phase, are illustrated in Table 12.

Table 12. User-Centred evaluation of personalized systems: methods and attributes to evaluate

	Phase 1	Phase 2	Phase 3	Phase 4
Questionnaires	✔		✔	✔
Interviews	✔	✔	✔	✔
Focus Groups	✔	✔		
Think-aloud			✔	✔
Observations			✔	✔
	Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
User characteristics	✔			
User needs	✔			
Appreciation		✔	✔	✔
Future system adoption		✔		
Perceived usefulness		✔		
Trust and privacy		✔	✔	✔
Appropriateness of adaptation			✔	✔
Comprehensibility			✔	
Usability			✔	✔
User behaviour			✔	✔
User performance			✔	
User experience				✔
User satisfaction				✔

A more recent review (Dhouib, Trabelsi, Kolski, & Neji, 2016) identifies the following usability factors for the evaluation of interactive adaptive systems: predictability, privacy, controllability, breadth of experience, unobtrusiveness, timeliness, appropriateness, transparency, comprehensibility, scrutability, effectiveness, efficiency, and precision. The methods typically employed in any of the possible adaptation layers are identified to be cognitive walkthrough, heuristic evaluation, focus groups, user-as-wizard, task-based experiments, and simulated users (Table 13).

Table 13. Methods employed to study usability factors of interactive adaptive systems per adaptation layer

	CID	ID	MW	DA	AA	Whole
Predictability	--	HE SU	HE FG / UW SU	CW HE FG / UW TE SU	CW HE FG / UW TE	--
Privacy	HE SU	HE	HE FG	CW HE / FG TE	CW HE / FG TE	CW HE / TE
Controllability	HE SU	HE	HE FG	CW HE / FG TE	CW HE FG / UW TE	CW HE / TE
Breadth of experience	--	--	--	CW HE FG / UW TE SU	CW HE FG / UW TE	--
Unobtrusiveness	--	--	--	--	CW HE FG / UW TE	--
Timeliness	--	--	--	--	CW HE FG / UW TE	--
Appropriateness	--	--	--	CW HE FG / UW TE SU	--	--
Transparency	HE SU	HE	HE FG	CW HE / FG TE	CW HE / FG TE	CW HE / TE
Comprehensibility	--	--	HE FG / UW SU	--	--	--
Scrutability	--	HE SU	HE FG / UW SU	CW HE FG / UW TE SU	--	--
Effectiveness	--	--	--	--	--	CW HE / TE

Efficiency	--	--	--	--	--	CW HE TE
Precision	--	--	HE FG UW SU	--	--	--

CID= Collect Input Data, **ID**= Interpret Data, **MW**=Model the current state of the world, **DA**=Decide upon adaptation, **AA**=Apply Adaptation; **CW**=Cognitive Walkthrough, **HE**=Heuristic Evaluation, **FG**=Focus Groups, **UW**=User-as-Wizard, **TE**=Task-based Experiments, **SU**=Simulated Users

In summary, the evaluation of adaptive interactive systems is a challenging and intricate task, not only in terms of deciding whether an adaptation was effective, efficient, and satisfactory to the user, but mainly towards identifying the reasons why the adaptation has failed. To this end, several layered evaluation frameworks have been proposed, involving assessments at the various layers involved in the adaptation process, such as data collection, data interpretation, modelling, decision-making, and application of the adaptation itself. The UxAmI framework takes into account all the potential adaptation evaluation criteria suggested in literature and also, through its Observer tool, it facilitates evaluators in identifying adaptations that are not accepted by users and therefore need to be further investigated.

3.5   EVALUATION OF UBIQUITOUS COMPUTING SYSTEMS

The origins of ubiquitous computing can be attributed to Mark Weiser (1991), who described his vision for the 21st century computing stating that: "The most pro-found technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it". Ubiquitous computing is the term given to the third era of modern computing, which is characterized by the explosion of small networked portable computer products in the form of smart phones, personal digital assistants, and embedded computers built into many of the devices we own—resulting in a world in which each person owns and uses many computers (Want 2010). In Ubiquitous Computing (UbiComp) environments, the interaction goes beyond the one-to-one model prevalent for PCs, to a many-to-many model where the same person uses multiple devices, and several persons may use the same device, while it may be implicit, invisible, or through sensing natural interactions such as speech, gesture, or presence (Bardram & Friday, 2010). In summary, the core properties of a UbiComp environment are (Poslad, 2011): (i) computers need to be networked, distributed, and transparently accessible; (ii) human-computer interaction needs to be hidden more; (iii) computers need to be context-aware in order to optimise their operation in their environment; (iv) computers can operate autonomously, without human intervention,

and be self-governed; and (v) computers can handle a multiplicity of dynamic actions and interactions governed by intelligent decision-making and intelligent organisational interaction.

Ubiquitous computing has constituted an important paradigm shift, but as we are heading in the fourth era of modern computing, its ideas have already pervaded much of computing research and practice (Abowd, 2012). In the next generation of computing, the human–computer experience will be more conjoined than ever before (Abowd, 2012). Such a vision may be fulfilled by Ambient Intelligence, which incorporates the features of UbiComp environments, but focuses on the human inhabitants of the environment, aiming to elevate the overall user experience. This section will study evaluation approaches in UbiComp environments and evaluation frameworks and models, as AmI and UbiComp share many common features and therefore challenges faced in the evaluation of UbiComp applications and services are valid for AmI as well.

Evaluation of ubiquitous computing systems is a challenging research area (Neely, Stevenson, Kray, Mulder, Connelly, & Siek, 2008). It is often the case that sub-parts of a UbiComp system are evaluated with well-known methods from well researched fields, rather than the whole system (Schmidt, 2003). Traditional usability evaluation techniques can certainly be used for the evaluation of ubiquitous applications. However, in the UbiComp paradigm, usability engineering faced new challenges, including the following (Kim, Kim, & Park, 2003):

- the traditional controlled laboratory testing is no longer appropriate and should be replaced by real use in the authentic environment
- the UbiComp applications and services do not have the full attention of the user
- the task-centric approach of traditional usability is not suitable for everyday use of computing in everyday life
- a wider range of factors should be considered in the evaluation process.

Additional problems that have been identified for the evaluation of UbiComp systems are the multi-causality and the evaluation goal (Schmidt, 2003). Multi-causality refers to the difficulty in assessing each individual system component (e.g., deployed devices, context awareness, interaction metaphors) to find the contribution of a particular design decision for the success or failure of the overall system. Furthermore, in UbiComp systems it is not always clear for what the overall system is evaluated (e.g., to demonstrate the feasibility of a concept, show the ease of use, evaluate enhanced user experience, proof the efficiency or stability of an implementation, and estimation of administration effort), whereas different evaluation goals necessitate different evaluation approaches.

Research efforts in the field can be classified in three major categories: new evaluation methods or usability evaluation methods revisited in the context of UbiComp environments; evaluation tools, aiming to provide a means to facilitate and automate evaluation; and evaluation models and frameworks, intending to guide evaluators in selecting a method or applying specific evaluation criteria.

In the first category, four user study techniques, applied in the context of UbiComp applications have been explored by Consolvo, Arnstein and Franza (2002): contextual field research, which is a technique for gathering qualitative data by observing and interacting with users in their everyday environments rather than the laboratory, while users perform their normal activities rather than contrived tasks; intensive interviewing, which involves gathering qualitative data by asking users open-ended questions about their work, background and ideas, spending several hours for each user; usability testing, in the lab or in the field, for gathering empirical data by observing users as they perform tasks with the application that is being evaluated; and lag sequential analysis, which is a technique for gathering quantitative data by observing users as they perform their normal activities and studying the behaviour of person-to-person interaction by measuring the number of times certain behaviours precede or follow a selected behaviour. The authors discuss the techniques applied for the evaluation of a UbiComp application and conclude that traditional usability testing is not an appropriate evaluation method, as it is difficult to apply task-centric evaluation techniques to everyday situations and the lab environment cannot lead to deep, empirical evaluation results.

Trying to address the need for evaluating ubiquitous applications outside the laboratory environment and before fully implementing them in the context of a ubiquitous environment, the Wizard of Oz technique has been used for the evaluation of a doorman application using spoken language and speech recognition (Mäkelä, Salonen, Turunen, Hakulinen, & Raisamo, 2001). The evaluation aimed at studying the use of speech synthesis and spoken language, as well as how to combine synthesized speech and pointing gestures. The experiment resulted in identifying several issues that should be addressed and providing guidelines for the design process. Consequently, a conclusion from the process that was applied was that more tests should be carried out during the development of ubiquitous computing systems, in order to make the process iterative and gain valuable information on how to improve the system. The same technique was used to mimic sensors deployed in home settings for supporting eldercare (Consolvo, Harrison, Smith, Chen, Everitt, Froehlich, & Landay, 2007). This study also confirmed that important design and system requirements can be discovered before much development effort has been put into building the underlying system, while applying the Wizard of Oz technique contributed even better requirements, as participants could get a realistic feel for what it would be like to actually use the technology as part of their everyday lives.

An approach that has received interest from the UbiComp community is that of "living laboratories", which are naturalistic environments equipped with sensing and observational technologies used for experimental evaluation, assisting researchers to bridge from laboratory testing to larger studies in real UbiComp environments (Intille et al., 2006). PlaceLab is such a live-in, apartment-scale research facility, which has been used in the context of three 10-day pilot studies (Intille, Larson, Beaudin, Nawyn, Tapia, & Kaushik, 2005), each with one participant who moved into PlaceLab and was directed to conduct his or her life as normally as possible for the study period. It is noteworthy that each stay yielded 200-250GB of data, which needed to be carefully reviewed by the researchers to identify and mark behaviours of interest. A custom visualization tool facilitating the process of reviewing the data features a floor plan of the facility, permitting the researcher to click on any sensor in the environment and immediately be taken to an audio-visual record of what was happening at the time of the sensor activation.

On the other hand, in-situ evaluation supports exploring how a system is actually used in its real environment, avoiding artificial situations imposed in laboratory and field testing (Fields, Amaldi, Wong, & Gill, 2007), being perhaps messy yet realistic (Kjeldskov & Skov, 2014) and allowing researchers to understand how the environment itself impacts the user experience (Rogers et al., 2017). Comparison of in situ and laboratory evaluation, carrying out evaluations in the exact same way, indicated that more usability problems were identified in situ, while it was only this type of evaluation that revealed problems related to cognitive load and interaction style (Nielsen, Overgaard, Pedersen, Stage, & Stenild, 2006).

A method that has been widely employed as a formative technique is experience sampling, a procedure that consists of asking individuals to provide systematic self-reports at random occasions during the waking hours of a normal week (Larson & Csikszentmihalyi, 1983). The method has been used to reveal where UbiComp solutions might be appropriate, as well as to define requirements for UbiComp solutions, and in more details to study information needs and what output devices were available to users throughout the day (Consolvo & Walker, 2003). Based on their experience, the researchers identify the major considerations regarding the application of the method, including the type of alerts (random, scheduled, or event-based), scheduling requirements, delivery mechanism, and data delivery options.

The experience sampling method has been further extended to include context awareness, and to acquire feedback from users in particular situations that are detected by sensors connected to a mobile computing device context (Intille, Rondoni, Kukla, Ancona, & Bao, 2003). Context-aware experience sampling has been applied for studying travel behaviours and place visit activities of users who had been at a specific place for at least ten minutes, as well as to study how technology can help encourage people to use a mobile phone and pedometer to provide personal awareness of physical activity (Consolvo et al., 2007).

To this end, the Context Aware Toolbox is a toolkit which provides a means to rapidly equip environment with sensors and actuators and to enable the monitoring of activities that can trigger an experience sampling survey or that that the actuators and sensors themselves can be used to collect data (Keyson, 2010).

One more variation of the experience sampling method introduced in the context of UbiComp environments is adaptive experience sampling (Vastenburg & Herrera, 2010), which aims to enhance existing experience sampling methods by providing instruments that enable researchers to easily inspect the preliminary findings in relation to context and product usage data, and to easily adapt timing and content of experience sampling on the fly. Key challenges identified towards achieving adaptive experience sampling are to detect unexpected behaviours or that anticipated behaviours did not occur, and also to be able to view the experience sampling findings in relation to time.

A combination of activity logs with experience sampling in the context of in-situ evaluation of mobile computing activities is employed in the MyExperience system (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007). MyExperience features a three-tiered architecture of sensors, triggers and actions. The triggers combine streams of sensor data with conditional logic (e.g., every time the mobile phone connects to a new cell ID) to invoke actions. The system can be used in two ways: as a stand-alone application configured via XML and scripting or as a library within another application.

In summary, the majority of user experience / usability studies in the field employ well-known methods as applied in the desktop GUI paradigm. A literature review regarding user experience evaluation in UbiComp (Väänänen-Vainio-Mattila, Olsson & Häkkilä, 2015) highlighted that out of the 75 papers reported, many of them (47 – 62.66%) involved field studies in real contexts of use. The data gathering methods employed were questionnaires (60%), interviews (53.33%), system logging (29.33%), observations (16%), diaries and probes (6.66%), as well as experience sampling (4%).

On the other hand, efforts towards evaluation tools aim to facilitate testing of prototypes and include Momento, a remote evaluation tool, and ActivityStudio, an open source suite of tools for prototyping and in situ testing of low-cost UbiComp applications. Momento (Carter, Mankoff, & Heer, 2007) provides integrated support for situated evaluation of ubiquitous computing applications and in more details it supports remote testing, helps with participant adoption and retention by minimizing the need for new hardware, and supports mid-to-long term studies to address infrequently occurring data. It can also gather log data, experience sampling, diary, and other qualitative data. ActivityStudio (Li & Landay, 2008) supports high-end target devices (e.g., tablet PC) through a virtual machine (VM) that can schedule and run multiple prototypes and periodically retrieve users' activity data from ActivityServer, and low-end

devices not able to run a VM, through a web browser or an extra-thin client that periodically retrieves interface screens from the ActivityServer. Additionally, ActivityStudio employs an architecture that can incorporate numerous activity-sensing components, which makes testing possible with multiple participants at the same time. Designers can monitor these test sessions, including users' activities and the state of the interface screens on the target devices.

In the context of UbiComp, but with a clear focus on the design and evaluation of mobile phone applications, MoPeDT is prototyping tool facilitating user testing as well (Leichtenstern, & André, 2009). The evaluation component supports recording of all user interactions during a user study synchronized with the audio-visual recording of the user and their environment, a live stream of the user's mobile phone screen displayed on the evaluator's computer, automatic capturing of screen shots, live annotations, and recording of the environmental context (i.e., information from the environment sensors). The analysis component of the tool provides the time-line based visualisation of the recorded data as well as the possibility for exporting annotated data in different formats supported by statistical analysis tools (e.g., SPSS).

A platform logging and analysing all user operations on a smartphone used in a real-time context over long periods of time is described by Wu, Liao, Chen, Hsu, and Li (2014). The platform includes three main modules: a log charting service, a log query service and the real-time monitor. The log charting service provides an overview of the collected data, such as application information, time stamp, usage frequency, etc. Example visualizations supported include daily and overall user activity, types of applications run by a user, number of times that an application runs at each time of the day on a daily and weekly representation, as well as an application's usage distribution for all users. The log query service can be used to make comparisons of the number of log files per application through a 12-month period. Finally, the real-time monitor can be used to detect patterns of behaviour based on the recorded events (e.g., a user might dial a phone number either by manually inputting the number, or by searching their contacts list).

Another framework is BaranC, a service-oriented framework that monitors all user interactions with a digital device and collects all available context data, in order to build a full model of user application behaviour (Hashemi & Herbert, 2016). The framework can be used in the context of User Centred Design (UCD) activities to assist in the analysis of users' interactions. The authors illustrate how the framework has been used by a service to monitor a user working with an Android smartphone, and to learn their patterns of application use at various levels of detail, in the context of a two-month user study. The service produced information summaries and patterns, such as heat map showing the most frequent days and

times of a day that users interact with their devices, time spent by a user on different application categories, or patterns of use of individual applications.

Focusing on handheld devices, the HUIA testing framework (Au, Baker, Warren, & Dobbie, 2008), carries out specific tests, including a comparison of the expected actions with the actual actions carried out by a user, assertion analysis evaluating specific usability metrics based on upper and lower threshold values, and hotspot analysis. To this end, the developer has to create the Expected Actions Script, which is an XML document describing the intended usage of the interface. The tool primarily focusses on analysing the usability of forms.

Finally, another line of research in the evaluation of UbiComp environments concerns simulation tools, which allow cheap and quick testing of applications and systems. UbiREAL (Nishikawa et al., 2006) is a simulator which reproduces behaviour of application software on virtual devices in a virtual 3D space. UbiREAL provides functions to facilitate deployment of virtual devices in a 3D space, simulates communication among the devices, and reproduces the change of physical environment characteristics caused by networked appliances (e.g., the room becomes brighter due to turning on a lamp). UBIWISE (Barton & Vijayaraghavan, 2003) is a simulator concentrating on computation and communication devices situated in physical environments. It presents two views, a 3D world and a close-up view of the devices and objects that the user may manipulate. TATUS is a ubiquitous computing simulator based on a 3D games engine (O'Neill, Klepal, Lewis, O'Donnell, O'Sullivan, & Pesch, 2005). An important benefit of the simulator is that it is independent of the system-under-test, which is a separate module connected to the simulator that makes decisions to change its behaviour in reaction to user movements, behaviour and environmental factors, such as network conditions, ambient noise, or social setting. Following a slightly different approach, the hybrid simulation method has been used to carry out user studies in order to evaluate pervasive interactions (Leichtenstern, André, & Rehm, 2010). More specifically, a virtual representation of a pervasive environment has been created, as well as a virtual mobile phone to assist in extending users' real world activities with their mobile phone in the virtual world. The hybrid study was compared to a real study in a pervasive environment, and yielded similar results.

Realising the need for a systematic approach for the evaluation of UbiComp systems, several research approaches have focused on creating evaluation frameworks and models. Building on the technology acceptance model (Davis, 1985) and its extensions, and further advancing them to address pervasive and ubiquitous environments, Connelly (2007) introduced PTAM, the pervasive technology acceptance model. As an alternative to evaluating a pervasive computing application, in situ or in the laboratory, the model aims to predict user acceptance and long-term usage, after minimal exposure to a prototype. PTAM (see

Figure 30), adds the following constructs to existing approaches: (i) trust, which in pervasive environments is very important due to the nature of data that are collected by the environment; (ii) integration, aiming to assess whether the technology is well-integrated into the environment and does not distract users or interferes with their other activities. Furthermore, it defines usage motivation, and socio-economic status as motivators, along with other user attributes, such as gender, age, and experience. Given the large corpus of research related to technology acceptance models, most of the parameters that PTAM introduces have already been addressed in other models. Nevertheless, the construct of integration is very important for ubiquitous environments. Yet, the framework does not include specific suggestions on how to measure the integration construct and has not been validated.
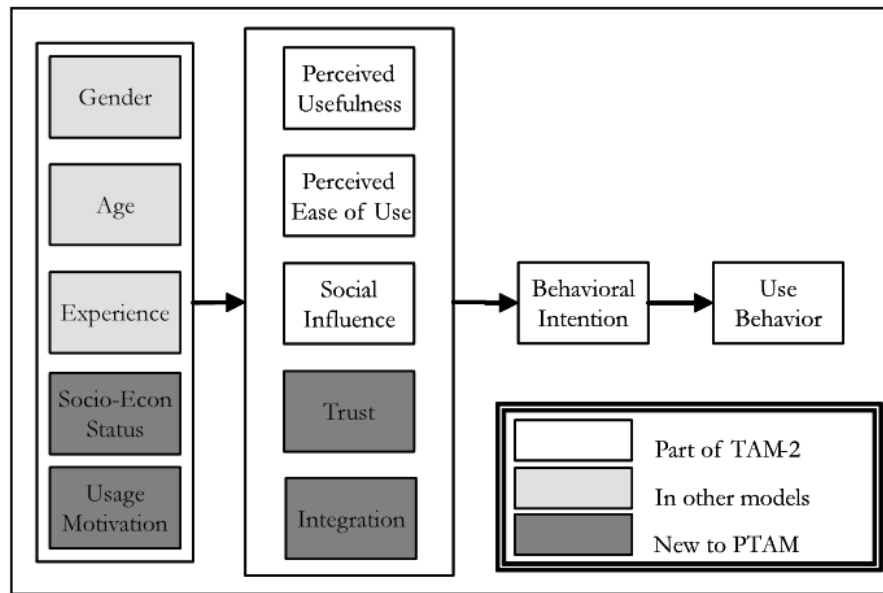
Figure 30. Pervasive Technology Acceptance Model

Scholtz and Consolvo (2004) developed a framework for ubiquitous computing evaluation, defining a set of evaluation areas, sample metrics and measures. In more details, the following nine evaluation areas and their related metrics are foreseen by the framework (Figure 31): (i) attention, with metrics focus and overhead; (ii) adoption, which can be measured by rate, value, cost, availability, and flexibility; (iii) trust, with privacy, awareness and control metrics; (iv) conceptual models, measured with the help of predictability of application behaviour and awareness of application capabilities; (v) interaction, measured by effectiveness, efficiency, user satisfaction, distraction, interaction transparency, scalability, collaborative interaction; (vi) invisibility, with metrics intelligibility, control, accuracy and customization; (vii) impact and side effects, measured through utility, behaviour changes, social acceptance, and

environment change; (viii) appeal, with metrics fun, aesthetics, and status; and (ix) application robustness, with metrics robustness, performance speed and volatility.

Attention
- Focus
- Overhead

Adoption
- Rate
- Value
- Cost
- Availability
- Flexibility

Trust
- Privacy
- Awareness
- Control

Conceptual models
- Predictability of application behavior
- Awareness of application capabilities
- Vocabulary awareness

Interaction
- Effectiveness
- Efficiency
- User satisfaction
- Distraction
- Interaction transparency
- Scalability
- Collaborative interaction

Invisibility
- Intelligibility
- Control
- Accuracy
- Customization

Impact and side effects
- Utility
- Behavior changes
- Social acceptance
- Environment change

Appeal
- Fun
- Aesthetics
- Status

Application Robustness
- Robustness
- Performance speed
- Volatility

Figure 31. Framework of Ubiquitous Computing Evaluation Areas (Scholtz & Consolvo, 2004)

An approach towards measuring the calmness of a ubiquitous computing environment is proposed by Carvalho, Andrade and Oliveira (2015). The evaluation is based on two main questions, namely if the application is capable of interacting with users at the right time, and if it effectively uses the periphery and the centre of user's attention. The first question is proposed to be assessed by measuring the adaptation degree, adaptation correctness degree, indicator of transparent mobility, availability degree and context-awareness timing degree. Transparent mobility is defined as the capability of the application to move from one device to another, keeping the past interactions and adapting resources to the new device, so that the user can continue their tasks seamlessly, and is measured through observations. Availability and context-awareness timing degree are measured through user forms. The second question

can be answered by measuring the number of irrelevant focus changes, proactivity of the application, number of failures, relevancy degree, and courtesy degree. Proactivity of the application aims to identify what the degree of proactivity is by counting – through a form filled-in by the developer – how many user actions the application is able to replace. Relevance and courtesy degree are values provided by users through forms.

Yang, Chen, Abdulrazak, and Helal (2010) identify the following attributes that should be considered to assess the performance of a pervasive computing system: invisibility, compatibility, deployment, safety, usability, resource usage, speed and efficiency, programmability, and sentience. Parameters to study the aforementioned issues are classified into system-centric, user-centric, as well as user and system centric, and further subdivided to quantifiable and non-quantifiable parameters (Table 14). Although system-centric parameters may affect the overall user experience, they are out of the scope of the current thesis, which aims to provide a framework and tools to assist researchers in measuring user experience itself for a given system (with specific performance, data storage, programming efficiency etc.). It should be noted that despite the fact that parameters are characterized by the framework as quantifiable or not, specific measures or methods are not described.

Table 14. User-centric, as well as system-and-user-centric parameters of the performance evaluation framework for pervasive systems (Yang et al., 2010)

	User-centric		System- and user- centric	
	Quantifiable	Non-quantifiable	Quantifiable	Non-quantifiable
User performance	✔			
Learnability	✔			
User effort	✔			
Modality	✔			
Acceptance		✔		
Satisfaction		✔		
Usefulness		✔		
Effectiveness			✔	
Adaptability & self-organization			✔	
Error			✔	
Explicitness			✔	
Adaptability characteristics				✔
Economic considerations				✔

Functionality
- Correctness
- Robustness
- Reliability

Usability
- Ease of use
- Accessibility
- Non-obtrusiveness

Security & Privacy
- Violation reports
- Configurable privacy and access control
- Encryption strength

Architecture
- Modularity
- Interoperability

Cost
- Installation cost
- Maintenance cost
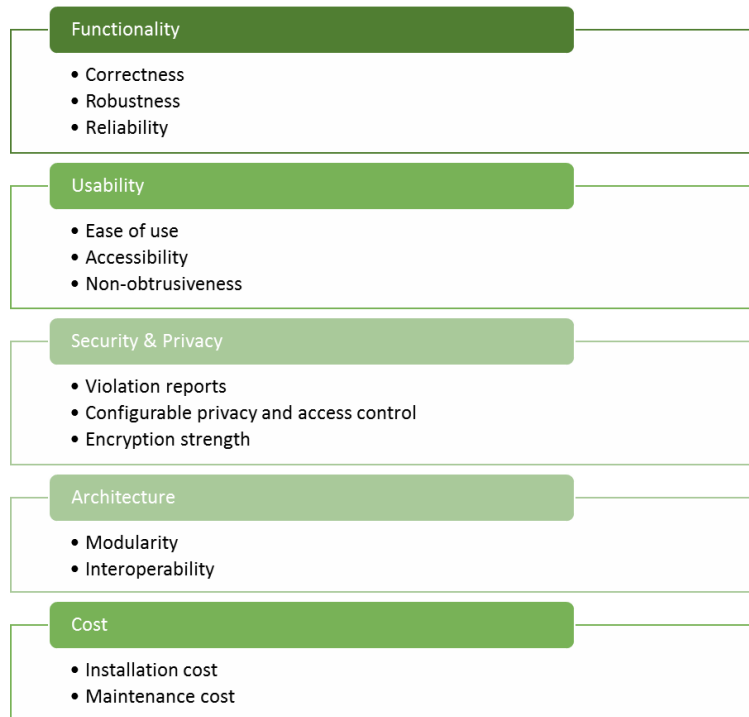
Figure 32. Evaluation framework to assess the quality of assistive environments (Metsis, Le, Lei, & Makedon, 2008)

In the context of pervasive and ubiquitous environment, but focusing on assistive environment, Metsis, Le, Lei, and Makedon (2008) proposed an evaluation framework to assess the quality of assistive environments (Figure 32). The framework identifies a set of attributes that are considered critical to user adoption, which is a requirement for the environment's success. The framework further proposes sample metrics and possible approaches to measure them. In more details, the attributes are organized under five categories, namely: (i) functionality, featuring the attributes of correctness, robustness, and reliability; (ii) usability, measured by ease of use, accessibility, and non-obtrusiveness; (iii) security and privacy, with the attributes of violation reports, configurable privacy and access control, and encryption strength; (iv) architecture, featuring the attributes of modularity, and interoperability; and (v) cost, attributed with installation cost and maintenance cost.

Carvalho, de Castro Andrade, de Oliveira, de Sousa Santos, and Bezerra (2017) propose a set of 27 quality characteristics that should be considered for the evaluation of ubiquitous computing systems, namely acceptability, attention, availability, calmness, context-awareness, device capability, ease of use, effectiveness, efficiency, familiarity, interconnectivity, mobility, network capability, predictability, privacy, reliability, reversibility, robustness, safety, scalability, security, simplicity, transparency, trust, usability, user satisfaction, and utility. Also, a detailed list of 218 software measures to achieve the aforementioned

evaluation of quality characteristics is proposed, with an indication of how well they are defined in the referenced sources. It is notable that out of the 218 measures only 36 are well defined, and the remaining 182 are either defined but without measurement function or not defined at all.

The TRUU Quality Model (Santos, de Oliveira, Andrade, Santos, & Lima, 2013) aims to guide quality evaluation of UbiComp systems by proposing four main characteristics that should be evaluated and by defining their sub-characteristics, as well as specific measurements for each one of them. More specifically, TRUU suggests that the system's trustability, resource-limitedness, usability, and ubiquity should be measured. Trustability is further analysed by security, privacy, control, and awareness. Resource-limitedness is decomposed in device capability and network capability. Usability is suggested to be measured through satisfaction, ease of use, efficiency, effectiveness, and familiarity, while ubiquity by context-awareness, transparency, availability, focus, and calmness. Moreover, specific measures for context awareness are described, and namely adaptation correctness, context correctness, context frequency, and adaptation time.

In summary, the evaluation frameworks proposed in the domain, encompass the notion of usability as well as various other features, sometimes overlapping. Table 15 lists the main UbiComp attributes classified in categories, as included in the aforementioned frameworks. In total, 70 attributes (organized under 17 categories) should be taken into account for the evaluation of UbiComp environments.

Table 15. UbiComp attributes proposed by evaluation frameworks

Category	Attributes	References
Attention	Focus, Overhead, Unnoticed events, Distractions, Frustration, Performance degradation	Carvalho et al. (2017), Santos et al. (2013), Scholtz and Consolvo (2004)
Conceptual models	Predictability of application behaviour, Awareness of application capabilities, Vocabulary awareness	Carvalho et al. (2017), Scholtz and Consolvo (2004)
Invisibility	Intelligibility, Control, Accuracy, Customization	Scholtz and Consolvo (2004)
Ubiquity	Context-awareness, Transparency, Availability	Carvalho et al. (2017), Santos et al. (2013)
Functionality	Correctness, Reliability	Carvalho et al. (2017), Metsis et al. (2008)

Calmness	Capability to interact with users in the right moment, Effective usage of the periphery and the centre of user's attention	Carvalho et al. (2015), Carvalho et al. (2017)
Usability	Effectiveness, Efficiency, Safety, User satisfaction, Learnability, Simplicity, User effort, Ease of use, Accessibility, Non-obtrusiveness, Familiarity	Carvalho et al. (2017), Metsis et al. (2008), Santos et al. (2013), Scholtz and Consolvo (2004), Yang et al. (2010)
Interaction	Interaction transparency, Scalability, Collaborative interaction	Metsis et al. (2008), Scholtz and Consolvo (2004), Yang et al. (2010)
Appeal	Fun, Aesthetics, Status	Scholtz and Consolvo (2004)
Trust / Trustability	Privacy, Awareness, Control / Configurable privacy and access control, Security	Carvalho et al. (2017), Metsis et al. (2008), Scholtz and Consolvo (2004)
Impact and side effects	Utility, Usefulness, Behaviour changes, Social acceptance, Environment change	Carvalho et al. (2017), Scholtz and Consolvo (2004), Yang et al. (2010)
Security and Privacy	Violation reports, Encryption strength, Expressiveness of the security policy, Unobtrusiveness of security mechanisms	Carvalho et al. (2017), Metsis et al. (2008)
Architecture	Modularity, Interoperability	Metsis et al. (2008)
Resource limitedness	Device capability, Network capability	Carvalho et al. (2017), Santos et al. (2013)
Application Robustness	Robustness, Performance speed, Volatility	Carvalho et al. (2017), Metsis et al. (2008), Scholtz and Consolvo (2004)
System-centric (other)	Adaptability and self-organization, Error, Explicitness, Mobility, Interconnectivity, Reversibility	Carvalho et al. (2017), Yang et al. (2010)
Adoption	Rate, Value, (Installation & Maintenance) Cost, Economic considerations, Availability, Flexibility, Acceptance	Carvalho et al. (2017), Metsis et al. (2008), Scholtz and Consolvo (2004), Yang et al. (2010)

Although the above frameworks introduce concepts relevant to ubiquitous environments and provide a classification of measures and metrics, in their majority they do not systematically assist evaluators in deciding which evaluation method to choose, or which exact metrics, according to the specific evaluation context (e.g., the context of use of the system evaluated, the development stage of the system, the users or experts that will be involved in the evaluation). On the other hand, given the high complexity of ubiquitous and pervasive computing environments, frameworks often end up to an unmanageable number of parameters, attributes and constructs that should be evaluated. To this end, the need for a systematic approach that will act as a guide to evaluators of UbiComp systems still remains to be addressed. Given the complexity of UbiComp environments, such a framework is not expected to cover all the potential systems, users and contexts of use; instead an extensible approach taking into account the various parameters of interaction in UbiComp environments could constitute the ground for further research in this direction. Furthermore, it is evident that such a framework should support a variety of methods, and be accompanied by appropriate tools to facilitate evaluation and reduce as much as possible the parameters that evaluators should assess on their own.

3.6   EVALUATION IN AMBIENT INTELLIGENCE ENVIRONMENTS

Evaluation in Ambient Intelligence is a challenging objective and a field which has not yet been extensively explored, due to the inherent difficulties it imposes. Stephanidis (2012) highlights that the evaluation of AmI technologies and environments needs to go beyond traditional usability evaluation in a number of dimensions, concerning both the qualities of the environment to be assessed and the assessment methods. A major concern is that evaluation should go beyond performance-based approaches to evaluation of the overall user experience (Stephanidis, 2012; Gaggioli, 2005), which should be further articulated in the context of AmI environment. Moreover, evaluation should take place in real world contexts (Stephanidis, 2012; Gaggioli, 2005), which is a challenging task by itself. Traditional evaluation practice has also been pointed out as insufficient for new HCI systems that feature new sensing possibilities, shift in initiative, diversification of physical interfaces, and a shift in application purpose (Poppe, Rienks, & van Dijk, 2007). Challenges include the interpretation of signals from multiple communication channels in the natural interaction context, context awareness, the unsuitability of task-specific measures in systems which are often task-less, as well as the need for longitudinal studies to assess the learning process of users.

User experience in Ambient Intelligence environments goes far beyond UX in the desktop paradigm, and should also be differentiated by UX in the UbiComp paradigm. Stephanidis (2012) identifies eight factors

determining user experience in AmI environments, namely natural interaction, accessibility, cognitive demands, emotions, health, safety and privacy, social aspects, cultural issues, and aesthetics. A guideline regarding the usability evaluation of smart home environments, issued by the ITG (Information Technology Group) technical society of VDE, a European technical and scientific association[7], identified that the following usability aspects should be considered: consistency, transparency, obtrusiveness, personalization, absence of barriers, adequacy to multiple users, trust and security, and robustness (Moeller, Engelbrecht, Hillmann, & Ehrenbrink, 2014). The guideline also takes into account the various services that are offered in a smart home environment, be them technological services (e.g., intelligent services, adaptive services, persuasive services, sensor and actuator services, input and output services, speech-based services) or domain-specific services (e.g., communication, safety, energy management, household control, entertainment, health and fitness). Finally, a taxonomy of performance and quality factors that should be taken into account for the evaluation of multimodal interfaces for AmI environments is proposed by Möller, Engelbrecht, Kühnel, Wechsung, and Weiss (2009). The taxonomy can be used to guide the evaluators, while authors suggest that because current systems cover a wide range of applications and domains, it is anticipated that an open framework will be needed to enable meaningful evaluation for specific contexts. The taxonomy consists of three layers:

- Quality factors layer, studying user, context and system factors. User factors are further subdivided to static and dynamic, context is defined by environmental and service factors, while system factors may be agent or functional.
- Interaction performance aspects including factors that pertain to the user and to the system. User factors are the perceptual effort, cognitive workload and response effort. System parameters include input performance, input modality appropriateness, interpretation performance, dialogue management performance, contextual appropriateness, output modality appropriateness, form appropriateness.
- Quality aspects including hedonic, acceptability, usefulness, and pragmatic aspects, namely: appeal as influenced by aesthetics and system personality; interaction quality affected by output quality, cooperativity, and input quality; usability as joy of use and ease of use; intuitiveness studied through effectiveness, efficiency, and learnability; and utility.

Ozkul and El Zarka (2013) describe the process of assessing the smartness of a UI as a four-step procedure, measuring: (1) how many steps it takes to achieve a goal, (2) how easy it is for the user to decipher the screen and find the correct button to activate the operation, (3) how long it takes for the device to respond to the action requested by the operator, and (4) how easy it is for the human operator to convey their

---

[7] https://www.vde.com/en/about-us

actions to the UI. A model is proposed based on the aforementioned four parameters, which calculates an interface quality score, based on values acquired through testing a system with 10 users.

Gaggioli (2005) discusses the concept of optimal experience, which is a state during which an activity is perceived as enjoyable and intrinsically worth doing for its own sake. Other features of this experience are: the perception of high challenges in the task at hand, personal skills adequately high to face those challenges, high concentration, involvement and enjoyment, absorption in the task, unselfconsciousness, control, and clear feedback on the course of the activity. Under this perspective, Gaggioli (2005) introduces the experience sampling method (see section 3.3.6) and the context-aware experience sampling (see section 3.5), which can be applied in AmI environments to address the inherent evaluation challenges.

Very few efforts have focused on providing a framework for evaluation in AmI environments. Such a framework, albeit quite generic and focusing on the processes and not the metrics, is the Experience Research theory, that supports user-centred design in Ambient Intelligence environments (De Ruyter & Aarts, 2010). The Experience Research theory involves studies in: (i) context, which focuses on collecting initial user requirements without introducing any new technology applications; (ii) the laboratory, with the aim to evaluate the new propositions in a controlled setting; and (iii) the field, which allows long-term testing in real life settings. Therefore, three dimensions can be identified in the process of generating experiences for AmI environments: Experience@Context, which involves trend studies, insight generation and validation; Experience@Lab, which may encompass concept definition, experience prototyping and user-centred design and engineering; and Experience@Field, involving involves field tests, longitudinal studies and trials (Aarts & de Ruyter, 2009). Implementing the Experience Research theory, an experience lab has been set up, which consists of a HomeLab, a ShopLab, and a CareLab, and aims to study user experiences of test participants during their stay in the ExperienceLab (De Ruyter, Van Loenen, & Teeven, 2007). In more details, for studying user experiences when setting up an experiment in the ExperienceLab, the researchers design a coding scheme for the observation session, listing all prototypical behaviours that are expected to be observable during the session. During the experiment, the observers mark the occurrence of these behaviours. Recapitulating their experience from testing in the ExperienceLab, the authors identify lessons learned, including the conclusion that although user experience is by nature subjective, there is a need to capture and analyse user experiences by means of objective methods.

Living labs (simulated AmI spaces) have been a very popular approach for the evaluation of AmI environments. Such a living lab approach was introduced by van Helvert and Wagner (2016), and involves two phases: (i) a preliminary experiment consisting of two phases and (ii) a final evaluation study. In the

first phase of the preliminary experiment, users are instructed to interact with the AmI environment based on specific scenarios. Execution of scenarios is followed by interview and filling-in a questionnaire. The second phase involves a focus group where participants from the first phase are presented with a futuristic scenario based on potential capabilities of the AmI space and engage in a discussion of their own views and visions. In the final evaluation study, users are allowed to interact with the AmI environment unconstrained (free-play), and freely explore the full range of functionality of the environment according to their own instinctive patterns of thought and action. The user interaction is video-recorded. The researcher and the user replay the video and provide a commentary, while gaps in the interaction – identified according to the Sense-Making method (Dervin, 1998) – are further elaborated. The proposed method has not yet been validated or employed in the context of evaluation in AmI environments.

Ambient Assisted Living (AAL) is a domain directly relevant to Ambient Intelligence, exhibiting considerable progress and results. Pereira, Teixeira and e Silva (2014) identify that most AAL evaluations rely on standard practices like enquiries, however they neglect contextual information or user-related data and propose an enquiry-based evaluation platform for use within a Living Lab. Using the proposed platform, the evaluator defines the questions that will be asked to the user and constructs a workflow for the instantiation of the question engine. During the evaluation, users' answers to the questions are provided to the online platform, therefore the evaluator can have direct access to the enquiry results and analyse the resulting information. A living lab approach has been used for the usability evaluation of two AAL applications (Dias et al., 2015), employing direct observation and questionnaires. The evaluation took place in a living lab simulating a regular house living room, while participants were asked to carry out specific tasks using the two AAL applications. Direct observation was employed to collect metrics such as task execution time, task completion rate, assistance during task completion, and the participant's visible emotional state. Users' satisfaction was measured by a custom questionnaire. Direct observation, questionnaires, reviewing the recorded sessions and log usage was employed for a scenario-based evaluation of multimodal interfaces for the smart home, in a laboratory setting simulating a living room (Fernández, Peláez, López, Carus, & Lobato, 2012), whereas the authors conclude that it would especially useful to develop methods to evaluate multimodal interfaces in multiuser environments. In the context of a Living Lab, a user evaluation has been carried out to assess the acceptance and fear of the smart home technology by the elderly (Portet, Vacher, Golanski, Roux, & Meillon, 2013). The experiment involved co-discovery of the smart home alternating between interview and wizard of Oz periods followed by a debriefing.

A framework oriented towards Wizard of Oz experiments in AmI environments, combining contextual rapid prototyping and the Wizard of Oz method was developed and evaluated through its application for

the assessment of a workflow support system in a semi-conductor factory (Zachhuber, Grill, Polacek, & Tscheligi, 2012). The Contextual Interaction Framework (CIF) allows the setup and handling of different contextual situations during user studies and consists of components to support modular programming, scenario configuration functionalities, context simulation possibilities, as well as a WoZ module to allow controlling the configured setup. A tool cooperating with the CIF framework is ConWIZ, which is used to carry out the contextual Wizard of Oz experiment and allows to control a specific prototype and the parameters included in the contextual settings (Grill, Polacek, & Tscheligi, 2012). For instance, to simulate the smart home context, the wizard is prepared to control different home appliances (e.g., turn light on/off, set TV volume, close/open curtains, etc.), while in the car context it is possible to simulate handling phone calls by voice, where the speech recognition is replaced by the wizard.

Another approach for evaluating AAL and AmI systems and services is based on simulation. A simulation-based approach to predict user errors is proposed by Halbrügge, Quade, Engelbrecht, Möller, and Albayrak (2016), combining model-based UI development with cognitive models. More specifically, the authors describe an integrated system targeted at predicting erroneous omissions of task steps depending on UI element characteristics. A validation experiment comparing data from usage of a system with real users against data predicted by the system highlighted that the model predicted the same errors, however with a different pattern. SISARL is a simulation environment to support the design and development of smart devices and systems for the elderly (Chen, Chen, Shih, & Liu, 2008). The developer needs to provide the simulator with the operational view specification of the device and models of the users. The operational view specification of a device comprises a package of workflows, i.e., definitions of activities and workflow graphs, together with resource components that simulate or implement activities and rules and policies that govern allocation of resources to workflows, device operations and device-user interactions. In terms of human models, the SISARL simulation environment only supports models defined by time parameters that specify the durations of user activities. An alternative method for simulating the context where a system is expected to be used is introduced by Singh et al. (2006). The authors propose capturing imagery and sound at the site of the intended deployment of a location-based service. Then, for the evaluation of a mobile or ambient application, the intended environment can be simulated by projecting the recorded images and videos (e.g., through a CAVE [8]) and simulating the sensor infrastructure, recreating thus the user's experience in a laboratory setting.

A hybrid approach is that of employing real users in virtual AmI environments, avoiding thus problems that have been reported in the use of model-based simulations (Fuchkina, Fischer, Tien, von der Heide, & Hornecker, 2016) and at the same time alleviating cost- and feasibility- related issues of living labs. Such an approach to the user experience evaluation of a context-aware smart home through hybrid reality-

---

[8] A cave automatic virtual environment is an immersive virtual reality environment where projectors are directed to between three and six of the walls of a room-sized cube (https://en.wikipedia.org/wiki/Cave_automatic_virtual_environment)

based interactive visualization is proposed by Seo, Kim, Kim, and Lee (2016), combining virtual and augmented reality (VR/AR) with physical reality. The proposed framework consists of three main layers: (1) reality-virtuality interface layer to support communication among physical and virtual objects, (2) context management layer based on a middleware for ubiquitous devices' context awareness, and (3) hybrid reality-based user experience layer, which is linked to the virtual-physical components through the context management layer, 3D augmentation, and synchronization with physical objects. In the VR-based interactive visualization environment, the user – wearing an HMD – views and navigates the immersive smart home space. At the same time, the user can interact with virtual objects through a Leap Motion[9] sensor that is attached to the HMD.

Malý, Curín, Kleindienst, and Slavík (2008) propose the transformation of participant interaction within an AmI environment into data represented in a virtual environment to address the challenges of data complexity, unification of different data formats and ethical issues. To support data transformation, the USEd tool has been developed which features a 3D visualization of an AmI space, options to facilitate an expert in creating the virtual counterpart of a live interaction (e.g., transformation of a video to the virtual environment tracing users' paths). The tool can be used to visualize users' interactions and also to support experts' evaluation through scenarios.

Well-established methods from the desktop paradigm have been extensively used in the context of AmI evaluations. A multi-method approach was used for the evaluation of an AAL service with the aim to test the feasibility of employing multiple methods for assessing the usability of AAL products or services (Martins, Queirós, Silva, & Rocha, 2016). The methods employed were: (i) self-perceived usability, (ii) usability evaluation based on the opinion of the evaluator on the users' performance, registration of quantitative performance data and namely task success or failure, task time, and number of errors, and (iii) registration of critical incidents. The benefits of a multi-method approach have also been pointed out in the evaluation of a home simulated environment by a group of elderly people (Casacuberta, Sainz, & Madrid, 2012). The study gathered the following quantitative and qualitative data: objective system performance variables (time and error measures) and subjective performance (questionnaire), users' spontaneous comments and verbalizations, observations and recording of participants' behaviour by experts, as well as assessment of system and control devices ease of use through questionnaires at different stages of the test.

A combination of implicit and explicit methods for the evaluation of an ambient persuasive display that provides information to operators in a semiconductor factory was adopted by Strasser, Weiss, Grill, Osswald, and Tscheligi (2012). The implicit evaluation method used was AMP (Payne, Cheng, Govorun, & Stewart, 2005), according to the following procedure: each participant is shown a sequence of three

---

[9] https://www.leapmotion.com/

images (i) an image of the system being evaluated, (ii) an abstract neutral image such as a Chinese character and (iii) a masked (e.g., with white noise) version of the abstract image. Participants are asked to indicate if they feel that the abstract image is pleasant or unpleasant, however their rating actually indicates how they feel about the preceding image. The explicit evaluation method employed was the Persuasion Questionnaire (PeQ), which consists of three parts, each aiming to assess: (i) if participants noticed that there was a display and what alerting method exactly made them aware of the display, given that ambient displays do not necessarily catch user's attention; (ii) the persuasive effect of the evaluated interface; and (iii) the areas of the display the participants would like to customize. The results of the evaluation carried out with the combination of these two methods indicated that both measurement methods show good incremental validity to each other.

Realizing the importance of ambient displays in an AmI environment, Mankoff et al. (2003) have proposed guidelines for the heuristic evaluation of ambient displays, with the aim to provide a low-cost evaluation approach. A comparison of the new guidelines for the evaluation of ambient displays to the original heuristic evaluation guidelines (Nielsen, 1994a) indicated that more severe problems were found, and also that 40-60% of known usability issues were identified. The final set of heuristic evaluation guidelines for ambient displays (including some guidelines as originally introduced by Nielsen) are: (i) sufficient information design, (ii) consistent and intuitive mapping, (iii) match between system and real world, (iv) visibility of state, (v) aesthetic and pleasing design, (vi) useful and relevant information, (vii) visibility of system status, (viii) user control and freedom, (ix) easy transition to more in-depth information, (x) "peripherality" of display, (xi) error prevention, and (xii) flexibility and efficiency in use.

In the context of ambient displays, a field study aiming to assess user experience of an Ambient Intelligence system in a retail store, involved user observation in the field, in situ interviews and use of video material recorded through a Spectacles Camera, that is a video camera built-in a pair of ordinary sports glasses (Reitberger, Obermair, Ploderer, Meschtscherjakov, & Tscheligi, 2007). The combination of immediate consumer perspectives through the Spectacles Camera, interview statements and direct observations over four days enabled the researchers to gain valuable insights into general shopping behaviour, as well as into the ambient display system that was evaluated.

An alternative effort towards combining objective data and enhancing them with information related to the user experience is experience tagging, a mechanism to annotate sensor data using subjective tags, enabling thus users to add a subjective view to the sensor data, and can be linked to the activity traces of the user in the smart environment (Vastenburg & Herrera, 2011). To implement the experience tagging concept, a touch-screen display was used as an interactive awareness display in the homes of seniors and

family caregivers. The display showed two activity traces – based on sensor data – one for the family caregiver, and one for the senior in need of care. In the context of the deployed prototype, passive infrared sensors were used to detect physical activity in the kitchen, living room and bedroom doorway. Users could select a mood from nine predefined mood tags and they could also add text to further comment on a sensor data or explain their tag. The evaluation of the system indicated that participants in general appreciated the system, however they exhibited a variance in how often and in what way they used the experience tags.

A framework oriented towards recognizing the user social attitude in multimodal interaction in smart environments is proposed by De Carolis, Ferilli and Novielli (2012). According to the proposed framework, signals of social attitude in multimodal interaction can be decomposed into signals in language, speech, and gestures. As a result, the user modelling procedure of the framework integrates (i) language analysis for linguistic cues extraction, (ii) prosodic analysis and (iii) gesture recognition into a Dynamic Belief Network. At the beginning of interaction, the model is initialized, while at every dialog step, knowledge about the evidences produced by the multimodal analysis are entered and propagated in the network, while the model revises the probabilities of the social attitude node. The new probabilities of the signs of social attitude can be used for planning how the environment will behave. In the context of the aforementioned framework the authors have classified gestures as (i) open attitude gestures, including arm(s) open, knees apart, elbows away from body, hands not touching, and legs uncrossed, (ii) closure attitude gestures, featuring crossed arms, gripping own upper arms, crossed legs and (iii) negative attitude gestures including, adjusting cuff, watchstrap, tie, etc., using an arm across the body, touching or scratching shoulder using arm across body, picking nose, pinching bridge of nose, and neck scratching.

Although studying the issue of evaluation from different perspectives, all the aforementioned approaches have recognised the importance of moving beyond the performance-based evaluation in the laboratory towards the evaluation of the entire user experience in real-world or realistic settings. Some of the state-of-the-art approaches described in this section constitute single evaluation experiments, while others are aimed towards establishing a more generic evaluation approach for AmI environments. In all cases, however, although the approaches are interesting and constitute a step beyond traditional usability and UX assessment, the evaluation scope is rather narrow, focusing on either specific evaluation topics or methods. Additionally, although several frameworks have been proposed in the UbiComp context, none have been explored for Ambient Intelligence environments. AmI as a concept is the direct extension of the concept of UbiComp, but it is much more than this, as AmI systems should be adaptive and responsive to the user's needs and behaviour (Bibri, 2015), therefore it is doubtful whether UbiComp models can be adequate for AmI. As a result, there is a need for a framework that will cater for evaluation of AmI

environments taking into account the different attributes of such an environment, the characteristics of its users and the various contexts of use, thus providing a useful tool for evaluators of AmI technologies and environments.

3.7   TOWARDS A USER EXPERIENCE EVALUATION FRAMEWORK FOR AMI ENVIRONMENTS

In view of the not distant realization of Ambient Intelligence environments, the main characteristic of which is that they are oriented towards anticipating and satisfying the needs of their inhabitants, there is an emerging need for understanding and scoping how evaluation should be carried out and what should be evaluated. To this end, this thesis has carried out a systematic review of 42 evaluation frameworks in the fields of usability, user experience, adaptive systems, ubiquitous computing systems, smart and AmI environments.

Table 16 presents a classification of these frameworks along five dimensions: (i) *type*, which may be conceptual if the constructs/sub-constructs to be evaluated are studied, methodological if the methodologies that can be applied for the evaluation are analysed, or both; (ii) if *metrics* are included (Y: Yes, N: No, Q: Questionnaire only, P: Partially, when metrics are given as an example, M: Mixed when both conceptual and detailed metrics are included); (iii) the evaluation *field* (e.g., usability, UX, adaptivity, UbiComp quality); (iv) *context*, which may be generic or a specific application domain (e.g., mLearning, eHealth); (v) if the framework has been *applied* for the evaluation of a specific system.

Table 16. Classification of reviewed evaluation frameworks

Framework	Type	Metrics?	Field	Context	Applied?
Adikari et al. (2011)	C	Q	UX	Generic	Y
Al-Azzawi (2013)	C	N	Usability	Generic	N
Carvalho et al. (2015)	C, M	Y	Calmness	UbiComp	Y
Carvalho et al. (2017)	C	Y	Quality	UbiComp	N
Connelly (2007)	C	N	Acceptance	UbiComp	N
Cota et al. (2014)	C	N	Usability	mLearning	N
Daniels et al. (2007)	M	P	Usability	eHealth	N
De Carolis et al. (2012)	C	P	Smart Env.	Social attitude	Y
Dhouib et al. (2016)	C, M	N	Adaptivity	Generic	N
Ferre et al. (2005)	M	N	Usability	Generic	N

Geerts et al. (2010)	C	N	UX, QoE	Generic	N
Harpur & De Villiers (2015)	C	M	UX	mLearning	Y
Hellweger & Wang (2015)	C	N	UX	Generic	Y
Heo et al. (2009)	C	P	Usability	Mobile	Y
Hussain & Kutar (2009)	C	Y	Usability	Mobile	N
Irshad & Rambli (2015)	C	N	UX	Mobile AR	N
Jin et al. (2009)	C	Q	Usability	Dishwashers	Y
Karagiannidis & Sampson (2000)	C	N	Adaptivity	Generic	N
Khan et al. (2011)	C	M	Usability	Haptic systems	N
Kurosu (2015)	C	N	Usability	Generic	N
Lachner (2016)	C	Q	UX	Generic	Y
Magoulas et al. (2003)	C	Y	Adaptivity	eLearning	N
McCarthy et al. (2005)	C	N	UX	Generic	Y
Metsis et al. (2008)	C	P	Quality	AAL	N
Miki (2014)	C	N	UX	Generic	N
Möller et al. (2009)	C	P	AmI	Multimodality	Y
Mourouzis et al. (2006)	C	N	Usability	Generic	N
Paramythis et al. (2010)	C, M	M	Adaptivity	Generic	N
Pu et al. (2011)	C	Q	Usability	Recommender systems	N
Santos et al. (2013)	C	M	Context-awareness	UbiComp	Y
Savioja (2014)	C	Q	UX	Complex systems	Y
Scholtz & Consolvo (2004)	C	M	Generic	UbiComp	N
Schulze & Krömker (2010)	C	N	UX	Online products	Y
Seffah et al. (2006)	C	P	Usability	Generic	N
Thayer & Dugan (2009)	M	N	UX	Generic	N
Van Velsen et al. (2008)	C, M	N	Adaptivity	Generic	N
Vavoula & Sharples (2009)	M	N	Usability	mLearning	Y

Wäljas et al. (2010)	C	N	UX	Cross-platform systems	N
Weibelzahl (2002)	C, M	Y	Adaptivity	Generic	Y
Weibelzahl & Lauer (2001)	C, M	N	Adaptivity	CBR	N
Yang (2010)	C	N	Performance	UbiComp	N
Zhang & Walji (2011)	C	Y	Usability	eHealth	Y

In summary, 32 of the frameworks (76.19%) are conceptual, 4 (9.52%) are methodological, and 6 (14.29%) are both conceptual and methodological (Figure 33 left). In terms of metrics included, only 6 (14.29%) frameworks include concrete metrics, 20 (47.62%) do not include any metrics at all, 6 (14.29%) partially include metrics as examples, 5 (11.90%) include mixed metrics using both conceptual (abstract) metrics and some concrete actually measurable parameters, and 5 (11.90%) use only questionnaires as metrics (Figure 33 right). Furthermore, only 16 (38.09%) of the studied frameworks have been used for the evaluation of a system. Regarding their scope, 22 of these frameworks are generic, while the remaining 20 target a specific subject (e.g., calmness) or application domain (e.g., mLearning, mobile AR, etc.). It is noteworthy that only 1 of the generic frameworks proposes concrete metrics, and only 5 have been actually applied in the context of an evaluation. Finally, only 2 of the total 42 frameworks move beyond conceptualization to specific metrics and methods, however they are both targeted to a specific evaluation factor, namely calmness and adaptivity respectively.
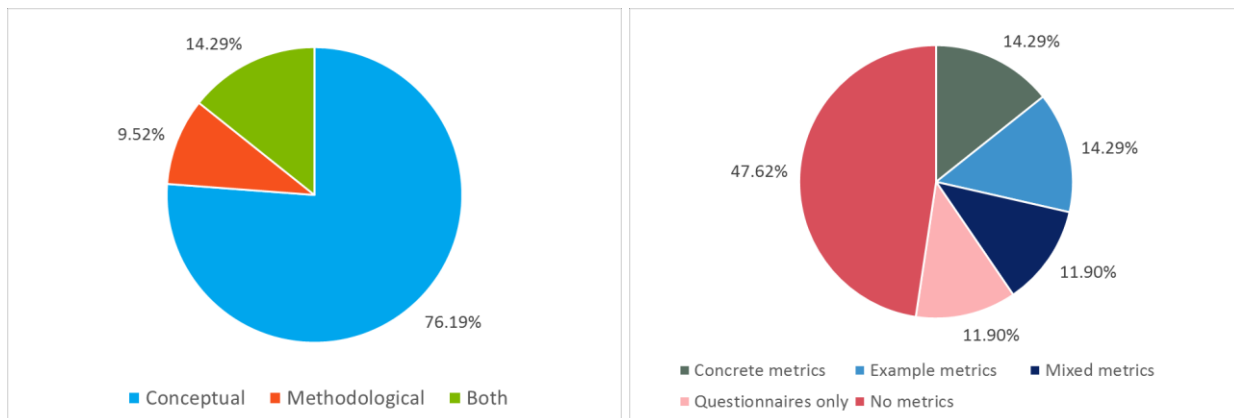
Figure 33. Analysis of reviewed frameworks: (Left) per type; (Right) per metrics included

Stemming from this review and analysis of existing approaches, a number of challenges need to be met towards defining a framework for the evaluation of AmI environments. An important concern refers to the scope of the framework, which should not be merely constrained at a conceptual level, and define

high-level concepts to evaluate. Although such a conceptual approach is an important first step in guiding evaluators, it is not adequate to assist the actual evaluation of AmI environments. In order to not create a purely theoretical framework, but a valuable tool for evaluators, such a framework should proceed to associating concepts with specific metrics and concrete methods.

Inevitably, as the complexity of the environment increases, the number of parameters that should be studied grows as well. The intersection of two already complex concepts, UX and AmI, may lead to the conception of frameworks with an extremely large number of parameters and metrics to be studied. An extensive framework would require 97 attributes for usability (Table 7), 79 for UX (Table 8), and 70 for UbiComp (Table 15), leaving out any AmI specific metrics. Even if duplicates are removed from the above indicative calculation (as for instance the UX attributes include some of the usability attributes), the number still remains very large. Taking into account that each attribute should be measured by at least a couple of metrics, the practical applicability of such a complex framework seems doubtful.

Therefore, a challenge that lies ahead is to create a clean-cut framework to evaluate AmI environments from a UX perspective, taking into account its various facets and temporal attributes, providing not only concepts, but also concrete metrics and methods to measure them. Although the high complexity of AmI environments may increase the evaluation complexity, their architecture and sensors' infrastructure may assist the evaluation procedure. AmI environments are equipped with sensors that provide information about the environment and users themselves, and with powerful reasoning mechanisms, which can be exploited by evaluation frameworks and tools to alleviate evaluators from the burden of recording dozens of parameters and synchronizing them with video recordings. The proposed framework aims to address these challenges, and along with the tools that have been developed, assist evaluators by presenting ready-to-use insights and statistics produced by the environment itself, allowing them to easily identify issues that hinder user interaction or environment characteristics and behaviours that should be improved in order to ensure high quality interaction and elevated user experience.

4 THE UXAMI FRAMEWORK FOR UX EVALUATION IN AMBIENT INTELLIGENCE ENVIRONMENTS

This section introduces the UX evaluation framework for AmI environments (UXAmI), a conceptual and methodological framework, by discussing its scope and main objectives (section 4.1), the characteristics and attributes of an AmI environment that the framework aims to assess, presenting the framework from a conceptual perspective (section 4.2), as well as the methodological perspective of the framework, by analysing the evaluation approaches according to which each attribute can be measured (section 4.3). Following, the UXAmI framework is presented in details (section 4.4), along with the evaluation that has been carried out and its results (section 4.5).

4.1 SCOPE AND OBJECTIVES

The proposed framework aims to provide a useful tool for UX evaluation experts towards designing and evaluating Ambient Intelligence environments. Taking into account best practices in the literature, and more specifically approaches for the evaluation of adaptive systems, UbiComp systems, as well as for UX and usability evaluation and technology acceptance, the proposed framework introduces a holistic approach that can be applied in any context of use. Given the complexity of AmI environments, and the wide range of potential contexts and target users, the framework does not constitute a panacea for any potential system; instead, it is an extensible approach taking into account the various attributes of AmI environments and parameters of interaction. It aims to provide a solid and clean-cut basis for the UX evaluation in any AmI environment, which can be further augmented with context-specific metrics if needed (e.g., metrics related to enhanced visitor flow in an AmI museum, support of medical practices in an AmI hospital, etc.).

The UxAmI framework constitutes both a conceptual and a methodological tool, describing not only attributes that should be measured, but concrete metrics as well, along with suggestions on the methods to be used towards acquiring the specified metrics. A challenge towards the development of the framework was the immense number of parameters that should be studied, given the complexity and multidimensionality of AmI environments, as well as the different temporal dimensions of UX and its multiple facets. As a result, an important concern that has guided the development of the framework was the trade-off between a huge list of metrics that would probably cover every possible aspect of an AmI system and the practical applicability of the framework in real contexts. To this end, the framework foresees the evaluation of an AmI system/environment through different phases and supports both

formative and summative evaluations. The UX practitioner is therefore provided with a consolidated, easy to manage list of metrics for each evaluation approach/phase.

Taking advantage of the infrastructure of smart and AmI environments (Cook & Das, 2004; Cook, Augusto, & Jakkula, 2009), UxAmI identifies a number of metrics and parameters that can be automatically calculated during a user testing session, alleviating the need for observers to explicitly record them. At the same time, this inherent support by the AmI environment provides an alternative to the common practice of asking users about almost everything, ending up with very lengthy questionnaires, requiring much time to answer and administer. Besides facilitating evaluators and users, the approach of automatically calculating metrics constitutes the missing link in mismatches and gaps often noticed in observers' recordings and users' questionnaire responses.

Another important concern for the development of the framework was to encompass best practices for the evaluation of AmI environments and to support both short-term and long-term evaluations with real users in simulation spaces (Living Labs), and facilitate practitioners in employing the appropriate metrics for each experiment type. For instance, in the case of short-term task-based evaluations, it is straightforward and meaningful to calculate task success, whereas this is almost impossible in situations where users are instructed to use the environment at their own discretion without a specific scenario. Towards this direction, UxAmI not only indicates the method to be applied (i.e., user testing) but also specifies the experiment type for which a metric is better suited. By the same token of guiding evaluators to apply the framework, a clear distinction of the attributes that should be measured along the different temporal dimensions of UX (i.e., before use, during use, shortly after use, long-term after use) is made in the case of UX experiments.

In addition, two significant research directions that are recognized and embraced by the UxAmI framework are technology acceptance theories and models, as well as the layered evaluation approach. With regard to the first, as pointed out in section 3.1.8, common practice so far has been to assess every aspect of the user's attitude through questionnaires, in order to calculate and predict the acceptance of a given technology. UxAmI provides a new means for substituting user-provided metrics related to one's experience with the system with observed and automatically calculated metrics. At the same time, it includes metrics stemming from users themselves, reflecting thereby their opinions, with a clear indication on when they should be measured according to the temporal UX dimensions. Concerning the layered evaluation approach, UxAmI has adopted the suggestion that an appropriate adaptation is a result of correct input data, valid inferences, and suitable instantiation of the adaptation itself, and guides evaluators towards assessing each of the above separately.

Finally, although several approaches in literature have attempted to quantify UX or usability by calculating specific indexes (e.g., degree of adaptation, aesthetics score, understandability index) (Carvalho et al., 2015; Wu, Chen, Li, & Hu, 2011; Seffah et al, 2001), UxAmI has not followed this path, as its purpose was not to provide a means to measure the performance of an AmI environment towards benchmarking. Instead, espousing the notion that user experience is unique for each individual and much more than simple adherence to guidelines, the UxAmI framework aims at constituting a tool for evaluators and designers to identify potential UX problems and eliminate them, by adopting a multi-method evaluation approach. However, apart from constituting a guide, the framework fosters the adoption of the proposed metrics and approaches through its accompanying tools, as described in section 5.

4.2 ATTRIBUTES OF AMI ENVIRONMENTS – CONCEPTUAL OVERVIEW

The UXAmI framework foresees the evaluation of seven fundamental attributes of Ambient Intelligence environments, namely intuitiveness, unobtrusiveness, adaptability and adaptivity, usability, appeal and emotions, safety and privacy, as well as technology acceptance and adoption (Figure 34). This section describes the UXAmI framework from a conceptual point of view, discussing the importance of each of the seven attributes in the context of the UX evaluation, and presenting the main high-level characteristics that determine each attribute.

Intuitiveness and unobtrusiveness are two important characteristics that AmI environments should exhibit. Intuitiveness is desirable for any system, underpins good design, and in general it means that the system employs pre-existing action-perception (motor) routines and socially (and culturally / historically) acquired "know-how", thus allowing users to focus on achieving a target goal through a system rather than on interacting with it (Turner, 2008). In the context of AmI, where novel means of interaction are inherently supported, applications may be pervasive, devices interconnected, and the system proactively anticipates and in some cases acts on behalf of the user, intuitiveness becomes a major need and challenge. The proposed framework suggests two main characteristics that should be assessed in this direction, namely that users are aware of the application/system capabilities and of the interaction vocabulary. Unobtrusiveness suggests that the system should not obstruct the users' main tasks (Paramythis et al., 2010) or generally place demands to the user's attention that reduce the user's ability to concentrate on their primary tasks (Ryu, Hong, & James, 2006). As a result, systems comprising the AmI environment should be appropriately embedded in the physical environment, and support user interactions without inducing distractions.
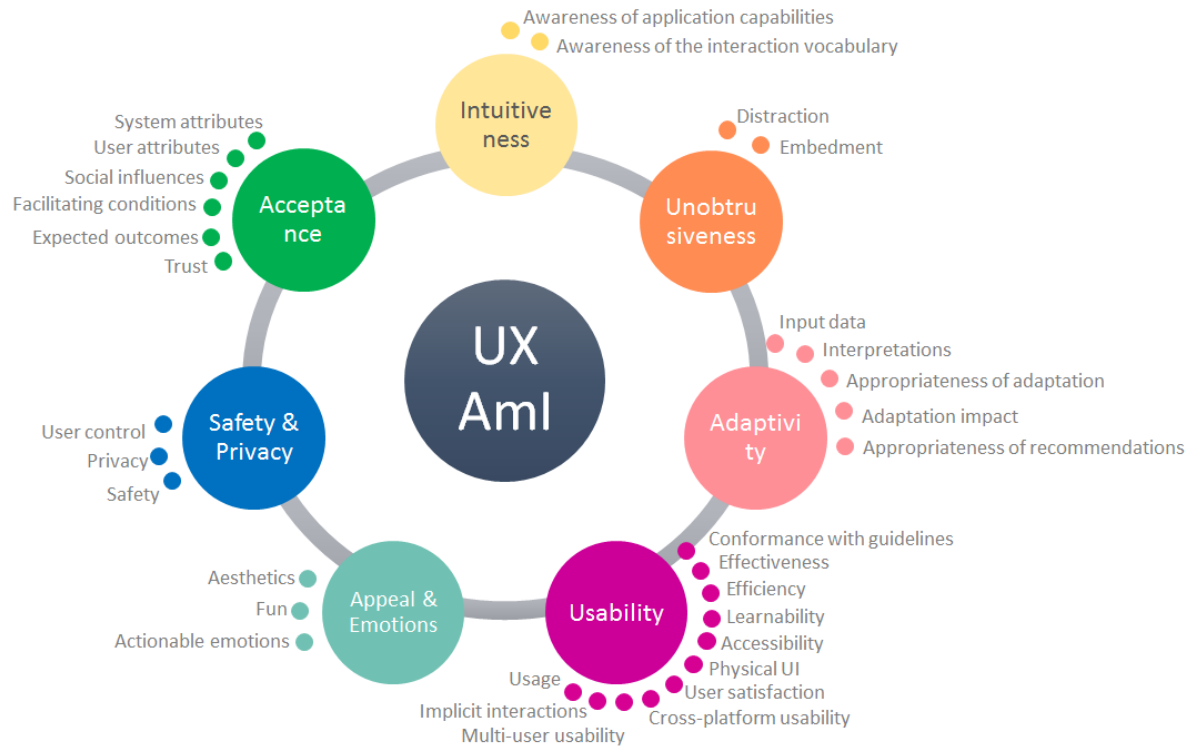
Figure 34. AmI attributes and characteristics evaluated in the UXAmI framework

Adaptability and adaptivity are core attributes that deal with the static and dynamic adaptations of the AmI environment according to each different user or user group and context of use. Context of use refers to the devices, the environment characteristics (e.g., light and sound levels) and the domain under which the system is being used (e.g., work, education, leisure, entertainment). Following the layered evaluation approach, adaptations are proposed to be studied in different layers, namely regarding the accuracy of data acquired through the environment's sensors, validity of interpretations, and appropriateness of an adaptation studied along three dimensions: interaction modalities supported, output provided and content delivered. The impact of an adaptation should also be explored, referring to how users react once an adaptation has been applied (e.g., if errors are increased). Last, as recommendations are also based on the same layers as adaptations, requiring valid input data, and appropriate inferences based on user and context models, the appropriateness of recommendations is another system characteristic assessed in the context of adaptability and adaptivity.

The cornerstone of the overall user experience is usability, referring to usability issues of each AmI system and to usability of the entire AmI environment, studying cross-platform usability, multi-user usability and implicit interactions, issues that are imperative to be evaluated given the confluence of platforms and

systems and the pervasiveness of applications, as well as the multiple users who may interact with the environment explicitly or implicitly, posing sometimes conflicting demands and requirements. The individual systems' usability refers to the qualities of each system that comprises the AmI environment, qualities which allow users to interact with it in an effective, efficient and satisfactory manner, also including learnability, accessibility, and conformance to relevant guidelines. Furthermore, the physical UI design of the individual systems should be assessed, as interaction in AmI environments goes beyond the typical desktop paradigm to using and interacting with novel objects. As a final point, the actual usage of the individual systems and applications of the AmI environment should be considered, with the aim to identify any usage patterns or preferences, and also detect systems and applications that are not used often or that are used for short periods of time.

Taking into account that user experience goes beyond usability assessment into looking users' emotions, perceptions, as well as physical and psychological responses, the framework includes the attribute of appeal and emotions. To this end, it deals with the aesthetics of the AmI environment and the systems that compose it, assesses how fun the users perceive the AmI environment and/or its systems to be, and how they actually feel. The latter is explored through users' reporting their affective reactions, as well as through detecting potential emotional strain through physiological measurements.

Safety and privacy are important parameters of the overall user experience and user acceptance of any technology. Especially for Ambient Intelligence environments and given their inherent capability to collect data on people's everyday interactions and to search large databases of that collected data, the issue of privacy becomes critical. Under this perspective, the framework studies the control that a user will have over the data that are collected by the environment and the information dissemination (i.e., if and what data will be communicated to other systems), as well as identity security issues. In addition, the level of control that the AmI environment has over the individual should be assessed. Finally, issues related to safety should be taken into account, including commercial damage, operator health and safety, public health and safety, as well as environmental harm.

Last, taking into account the holistic approach of user experience, studying the user's perceptions before, during and after the use of a specific product, the framework caters for studying the overall technology acceptance and adoption of an AmI environment. This can be further analysed by studying system features as perceived by the user, user attributes, and social influences to use a specific system, facilitating conditions, expected outcomes, and trust.

4.3   EVALUATION APPROACH – METHODOLOGICAL OVERVIEW

A fundamental constraint in existing approaches is that several of the user experience qualities that the evaluation aims to assess are measured through questionnaires, by recording the user's subjective opinion on a matter. As a result, if one would like to study a plethora of issues, the evaluation questionnaire would end up being too large to be answered. To this end, the proposed framework aims to assess as many issues as possible through other methods. However, user testing is the most fundamental evaluation method (Nielsen, 1994b) and cannot be replaced by any other method, therefore it constitutes a core evaluation approach of the framework.

Following an iterative approach, the framework proposes a combination of formative and summative evaluation methods, namely expert-based reviews and user testing (Figure 35). These two methods are the most popular and the ones actually employed during evaluations, as indicated by the literature review that has been carried out (Martins et al., 2014). During the design and prototyping phases of an AmI environment application/system the framework proposes evaluation through expert-based reviews. As the centre of the iterative design approach is the recurrence of evaluation and the improvement of designs and prototypes based on the evaluation results, expert-based evaluations can be planned by the evaluator when appropriate. Once a fully functional prototype is available, or when the evaluator deems proper, user-based UX evaluation can take place. It should be noted that the framework describes what should be measured and how, and simply provides suggestions as to when. Evaluators can employ the proposed methods according to their own experience and needs during the lifecycle of the development of an application or system that will be deployed in an AmI environment. Moreover, the scope of the evaluation may vary from a specific application running in one system, to a system including many applications, a pervasive application running in multiple systems, or an entire AmI space.

In a nutshell, UXAmI proposes combining formative and summative methods for better results. This combination is common practice in evaluations (Martins et al., 2014), since through formative evaluations several major problems can be eliminated without the need for involving actual users, or running resource demanding long-term experiments. On the other hand, it has been shown that the different assessment approaches and more specifically expert-based reviews and user-based evaluations find fairly distinct sets of usability problems, therefore they complement each other (Nielsen, 1994b).

Expert-based reviews may be used to assess various aspects of the individual systems in the AmI environment, such as embedment, validity of interpretations, appropriateness of recommendations, compliance with general and domain-specific guidelines, accessibility, physical UI, aesthetics, user control over the data collected and the behaviour of the AmI environment, as well as privacy and safety. User

testing constitutes a vital approach for the evaluation of user experience in Ambient Intelligence environments. It should be noted that all the user-testing protocols (e.g., thinking-aloud, retrospective testing, coaching, co-discovery learning, co-operative evaluation, etc.) can be applied, while user testing is used as a term for any type of test that employs users and namely (task based) tests in simulation spaces (Living Labs), in situ evaluations, or real long-term usage in Ambient Intelligence environments. An important contribution of the framework is that it will enhance the evaluation process with automated measurements provided through the AmI environment. Moreover, for the majority of metrics pursued to be recorded through observation, the ones for which automation support through tools is feasible are clearly marked. A combination of the automated measurements, metrics with automation support, user observation, questionnaires and interviews is expected to allow evaluators to gain insight into the composite issue of user experience. In order to effectively combine all the aforementioned information deriving from different sources, an important concern that should be addressed is that of synchronizing automated measurements, evaluator observations, and video recordings, in order to further assist the evaluator in comprehending interaction difficulties and deducing useful conclusions. This challenge is effectively addressed in the current thesis through the UXAmI Observer tool (Section 5.1).
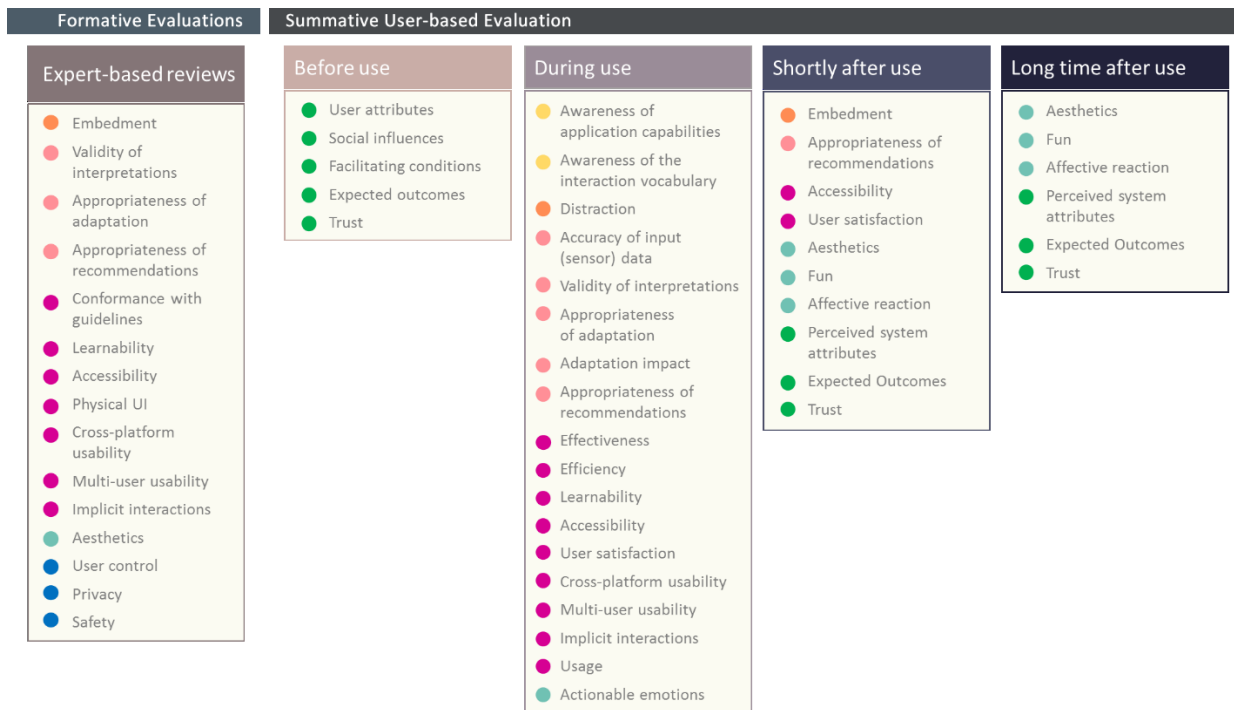
Figure 35. Evaluation approaches employed in the context of the UXAmI framework

4.4   THE UXAMI EVALUATION FRAMEWORK

Having studied the AmI environment attributes that the framework aims to assess, as well as the evaluation methods that will be employed to this end, this section presents the proposed framework, including metrics and measurement approaches for each attribute.

In the context of intuitiveness, the awareness of application capabilities can be measured by identifying the functionalities that have been used for each system, as well as the undiscovered functionalities. These metrics can be provided automatically by the AmI environment itself, given that some kind of instrumentation exists. More specifically, two preconditions need to be met: (i) declaration of the entire set of functionality supported by an application and (ii) communication of the application with the AmI environment infrastructure to identify when a specific functionality is used. Awareness of the interaction vocabulary is based on exploring input commands provided by the users, and more specifically: (i) calculating percentages of input modalities used, that is which exact modalities are used by the user in their interaction with the system and how often, highlighting thus users' preferences regarding the supported input modalities, (ii) identifying erroneous user inputs per input modality (e.g., gesture, speech, etc.), specifically user input commands that have not been recognized by the system, and (iii) percentage of erroneous user inputs per input modality, providing a general pointer as to how easy it is for a user to employ the specific modality. The aforementioned measurements can also be automatically acquired.

With regard to unobtrusiveness, distraction is measured through the number of times that the user has deviated from the primary task, as well as the time elapsed from a task deviation until a user returns to the primary task. Both metrics mainly apply to task-based evaluations or free exploration through thinking aloud, as in free exploration and usage it is not possible to know or to always correctly infer the user's goal, unless explicitly stated by users themselves. Evaluators can be assisted towards calculating these metrics, by having only to mark (e.g., through pressing a specific key) when a task deviation starts and when it ends. The characteristic of embedment, and more specifically whether the system and its components are appropriately embedded in the surrounding architecture, are suggested to be evaluated by experts, as well through questionnaire and interviews with the users after their interaction with the system.

Adaptability and adaptivity are proposed to be evaluated through assessing five main characteristics, following the paradigm of layered evaluation. First, the accuracy of input data perceived by the system should be assessed (e.g., accuracy of the data received by the sensors). This can be carried out through user testing. Automation support can be provided, by displaying to the evaluator all the input data acquired through the environment sensors not in a raw format, but elaborated in a semantically

meaningful form. The next assessment level refers to the validity of interpretations, a metric which can be calculated through expert-based review of the adaptation logic, and user testing with automation support. Automation support in this case refers to displaying, in a meaningful manner, the specific inferences of the reasoning mechanism, prior to applying an adaptation. At the next level, the appropriateness of an adaptation is evaluated, by means of exploring whether the interaction modalities, the system output, and the content are appropriately adapted according to the user profile and context of use, through user testing with automation support. The metric of adaptations that have been manually overridden by the user, indicates whether an adaptation is not only appropriate but acceptable as well, and can be acquired through automation supported user testing. Automation refers to the potential of the environment to detect when a user interaction possibly denotes an objection of the adaptation applied, by changing the state of a system that was also modified in the context of an adaptation (e.g., if the environment dims the lights following a suggestion by a reasoning agent, while the user turns them to full bright). The confirmation of whether the adaptation was actually rejected by the user should be provided by the evaluator. Besides being appropriate and acceptable, an adaptation may impose difficulties to a user, therefore its impact should also be assessed. To this end, the automated measurements of the number of erroneous user input commands once an adaptation has been applied and percentage of manually overridden adaptations can be employed. Additionally, the number of erroneous user interactions (e.g., selecting a wrong menu item) can provide an indication on the impact of the adaptation, which can be automatically calculated based on instances of interaction errors marked by the evaluator. Finally, the appropriateness of recommendations can be assessed through the following metrics: if adequate explanations of any recommendations are given by the system (assessed through user testing with automation support), if it is possible for a user to express and revise their preferences (by expert-based review), if recommendations are appropriate for the specific user and context of use (via expert-based review and user testing with automation support), which specific recommendations have not been accepted by the user (user testing with automation support), percentage of accepted system recommendations (automated measurement in user testing), and finally user's satisfaction by the system recommendations assessed through questionnaire and followed up by interviews if needed.

The next attribute, usability of the specific systems and the entire AmI environment, is studied through the evaluation of eleven characteristics analysed in specific metrics. The system's conformance with guidelines should be at first evaluated by expert-based review, taking into account all the guidelines that are relevant for the systems and applications under inspection. Effectiveness can be measured by two fully automated metrics, number of input errors and number of system failures, and two metrics with automation support, namely task success and number of interaction errors, where the environment can

produce calculations based on actual values indicated by the evaluator. Efficiency is proposed to be measured by the automated metric of time on task, and two metrics with automation support, number of help requests and time spent on errors. Learnability can be evaluated via cognitive walkthrough carried out by experts, as well as by studying users' performance (number of interaction errors and number of input errors) and help requests over time, metrics which are calculated automatically. Accessibility can be inspected by experts assisted by semi-automated evaluation tools to assess conformance with accessibility guidelines. Accessibility refers both to electronic and physical accessibility and can be assessed both by experts and by user testing, focusing on observations regarding accessibility problems and retrieving users' opinion through interviews. Electronic accessibility deals with the qualities of the software systems that constitute the AmI environment, which should allow their effective and efficient usage by users with functional limitations due to disability or aging. Physical accessibility, on the other hand, refers to the attributes of the environment that constitute it usable by diverse target user groups (e.g., elderly, disabled, children). The overall physical design should be assessed by experts studying whether the system violates any ergonomic guidelines and checking whether the size and position of the system and its interactive controls is appropriate for manipulation by target user groups. The latter can also be explored through user testing by observing users' interaction with the physical elements of systems in the AmI environment. User satisfaction is typically assessed through questionnaires aiming to elicit users' opinion regarding the system. Besides, during a user testing session the following can be recorded as indicators of user satisfaction: favourable and unfavourable user comments, statements expressing frustration, and declarations of clear joy. Although these need to be manually indicated by the observer, automatic calculation of percentages and total numbers of the above indicators constitute metrics of user satisfaction. The characteristic of cross-platform usability involves metrics studying consistency among the user interfaces of the individual systems, appropriateness of content synchronization and actions, which can be inspected by experts. Additional metrics refer to user interaction and behaviour once the user switches devices (platforms) and in more details: the time spent to continue the task from where it was left, help requests after switching devices and comparisons of cross-platform task success and task times, for task-based evaluations. All these metrics can be acquired and calculated through user tests, either with automation support or fully automated. In all cases, the environment can effectively detect when the user has changed device, requesting evaluator input only for metrics that cannot be fully automated (e.g., task success). Multi-user usability involves measuring, through automated measurements, the number of collisions with activities of others and conflicts resolved by the system. The evaluator can also observe in a user testing and indicate conflicts resolved by users themselves and the correctness of the system's conflict resolution, supported by the environment in calculating total numbers and percentages. Last, experts should carry out inspections of the behaviour

of the AmI environment to verify that the system does not violate social etiquette. Implicit interactions refer to actions performed by the user that is not primarily aimed to interact with a computerised system but which such a system understands as input (Schmidt, 2000) and can be explored by reviewing which implicit interactions occur and of what type (e.g., location-based, emotion-based, etc.), that can be automatically calculated. It is also important to study the appropriateness of system responses to implicit interactions, a task which can be supported by the environment by displaying all system responses after an implicit interaction, allowing evaluators to assess its appropriateness, and by calculating numeric metrics based on evaluators' judgement. Finally, the metrics regarding the actual system and application usage in the AmI environment, which are all acquired through user testing and are automatically provided by the environment are: number of usages per hour on a daily, weekly and monthly basis for the entire AmI environment, for each system and each application; time duration of users' interaction with the entire AmI environment and also with each individual system and application, analysis (percentage) of applications' used per system, as well as analysis (percentage) of systems to which a pervasive application is actually deployed.

Evaluation of appeal and emotions involves examining metrics related to aesthetics, fun, and users' emotions. More precisely, aesthetics are evaluated by experts reviewing if the systems follow principles of aesthetic design and reporting any violations, as well as by asking users their opinion on the matter through questionnaires. Fun and users' affective reaction to the systems are also suggested to be assessed by users' responses to questionnaire. Finally, taking into account that physiometrics can be acquired through sensors of the AmI environment, actionable emotions can be automatically detected and brought to evaluators' attention.

Characteristics and metrics related to safety and privacy are proposed to be evaluated through expert based reviews. In particular, user control can be assessed by verifying that user has control over the data collected and the dissemination of information, and also that they can customise the level of control that the AmI environment has on behalf of the user (e.g., acts on behalf of the person, gives advice, or simply executes user commands). Privacy involves inspecting the availability of the user's information to other users of the system or third parties, the availability of explanations to a user about the potential use of recorded data, as well as the expressiveness of the security and privacy policy. Lastly, safety involves inspecting if the AmI environment is safe for its operators and safe in terms of public health, and it does not cause environmental harm or harm to commercial property, operations or reputation in the intended contexts of use.

Technology acceptance characteristics are pursued through users' responses to questionnaires delivered, before, shortly, and/or long after the user's interaction with the system. System attributes aimed to be assessed are its perceived usefulness and ease of use, its trialability, relative advantage, as well as installation and maintenance cost. Questions regarding cost should not be necessarily addressed to the end-users, as they are not always the ones directly responsible for it (e.g., in an organizational, or public setting). User attributes that should be explored include the user's self-efficacy, computer attitude and personal innovativeness, as well as their age and gender. Metrics regarding social influences include subjective norm and voluntariness, while the ones related to facilitating conditions are end-user support and visibility. Expected outcomes can be explored in terms of perceived benefit, long-term consequences of use, observability, and image. Finally, user's trust towards the system should also be assessed, as it is an important parameter affecting adoption intentions.

All the specific metrics that the UXAmI framework proposes, categorised under characteristics and general AmI attributes to be assessed are listed in Table 17, reporting the appropriate methods for each metric. Metrics acquired through user testing include the following additional indications:

- Whether automation is possible, with the indication *automated measurement* for full automation and *automation support* whenever full automation is not possible, but the evaluator can be assisted in calculations and observation recording. In general, fully automated measurements are based on the architecture of AmI environments and the typical information flow in such environments, whereby interactors (e.g., people) perform their tasks, some of these tasks trigger sensors, and these in turn activate the reasoning system (Augusto, 2007). Therefore, interaction with a system in the AmI environment is not a "black box", instead it goes through sensors and agents residing in the environment, resulting in knowledge of interactions by the environment. A more detailed analysis of how the architecture of AmI environments is used for the implementation of such automated measurements is provided in section 5.2.2.
- If the metric should be acquired before the actual system usage (⏱ *B*), during (⏱ *D*), shortly after (⏱ *sA*), or long after it (⏱ *lA*).
- If the metric pertains to a task-based experiment *(Task-based)*, or if it should be applied only in the context of real systems' usage *(e.g.,* in *in-situ or field studies).*
- If the metric is to be acquired through a specific question in the *questionnaire* that will be filled-in by the user after their interaction with the system, or as a discussion point in the *interview* that will follow up.

Table 17. The UXAmI framework: concepts, attributes, metrics and methods

INTUITIVENESS		
Awareness of application capabilities	Functionalities that have been used for each system	User testing [🕐 D]: Automated measurement
	Undiscovered functionalities of each system	User testing [🕐 D]: Automated measurement
Awareness of the interaction vocabulary	Percentage of input modalities used	User testing [🕐 D]: Automated measurement
	Erroneous user inputs (inputs that have not been recognized by the system) for each supported input modality	User testing [🕐 D]: Automated measurement
	Percentage of erroneous user inputs per input modality	User testing [🕐 D]: Automated measurement
UNOBTRUSIVENESS		
Distraction	Number of times that the user has deviated from the primary task	User testing [🕐 D] [Task-based, or Think Aloud]: Automation support
	Time elapsed from a task deviation until the user returns to the primary task	User testing [🕐 D] [Task-based, or Think Aloud]: Automation support
Embedment	The system and its components are appropriately embedded in the surrounding architecture	Expert-based review User testing [🕐 sA]: Questionnaire, Interview
ADAPTABILITY AND ADAPTIVITY		
Input (sensor) data	Accuracy of input (sensor) data perceived by the system	User testing [🕐 D]: Automation support
Interpretations	Validity of system interpretations	Expert-based review User testing [🕐 D]: Automation support
Appropriateness of adaptation	Interaction modalities are appropriately adapted according to the user profile and context of use	Expert-based review

		User testing [🕐 D]: Automation support
	System output is appropriately adapted according to the user profile and context of use	Expert-based review User testing [🕐 D]: Automation support
	Content is appropriately adapted according to the user profile and context of use	Expert-based review User testing [🕐 D]: Automation support
	Adaptations that have been manually overridden by the user	User testing [🕐 D]: Automation support
Adaptation impact	Number of erroneous user inputs (i.e., incorrect use of input commands) once an adaptation has been applied	User testing [🕐 D]: Automated measurement
	Number of erroneous user interactions once an adaptation has been applied	User testing [🕐 D]: Automation support
	Percentage of adaptations that have been manually overridden by the user	User testing [🕐 D]: Automation support
Appropriateness of recommendations	The system adequately explains any recommendations	Expert-based review User testing [🕐 D]: Automation support
	The system provides an adequate way for users to express and revise their preferences	Expert-based review
	Recommendations are appropriate for the specific user and context of use	Expert-based review User testing [🕐 D]: Automation support
	Recommendations that have not been accepted by the user	User testing [🕐 D]: Automation support
	Percentage of accepted system recommendations	User testing [🕐 D]: Automated measurement
	User satisfaction by system recommendations (appropriateness, helpfulness / accuracy)	User testing [🕐 sA]: Questionnaire, Interview

USABILITY		
Conformance with guidelines	The user interfaces of the systems comprising the AmI environment conform to relevant guidelines	Expert-based review
Effectiveness	Task success	User testing [⏱ D] (Task-based): Automation support
	Number of interaction errors	User testing [⏱ D]: Automation support
	Number of input errors	User testing [⏱ D]: Automated measurement
	Number of system failures	User testing [⏱ D]: Automated measurement
Efficiency	Task time	User testing [⏱ D] (Task-based): Automated measurement
	Number of help requests	User testing [⏱ D]: Automation support
	Time spent on errors	User testing [⏱ D]: Automation support
Learnability	Users can easily understand and use the system	Expert-based review (cognitive walkthrough)
	Number of interaction errors over time	User testing [⏱ D]: Automated measurement
	Number of input errors over time	User testing [⏱ D]: Automated measurement
	Number of help requests over time	User testing [⏱ D]: Automated measurement
Accessibility	The system conforms to accessibility guidelines	Expert-based review Semi-automated accessibility evaluation tools
	The systems of the AmI environment are electronically accessible	Expert review User testing [⏱ D]

	The AmI environment is physically accessible	Expert review User testing [🕐 D] User testing [🕐 sA]: Interview
Physical UI	The system does not violate any ergonomic guidelines	Expert-based review
	The size and position of the system is appropriate for its manipulation by the target user groups	Expert-based review User testing [🕐 D]
User satisfaction	Users believe that the system is pleasant to use	User testing [🕐 sA] [🕐 lA]: Questionnaire
	Percent of favourable user comments / unfavourable user comments	User testing [🕐 D]: Automation support
	Number of times that users express frustration	User testing [🕐 D]: Automation support
	Number of times that users express clear joy	User testing [🕐 D]: Automation support
Cross-platform usability	After switching device: time spent to continue the task from where it was left	User testing [🕐 D]: Automation support
	After switching device: number of interaction errors until task completion	User testing [🕐 D]: Automated measurement
	Consistency among the user interfaces of the individual systems	Expert-based review
	Content is appropriately synchronised for cross-platform tasks	Expert-based review
	Available actions are appropriately synchronised for cross-platform tasks	Expert-based review
	Help requests after switching devices	User testing [🕐 D]: Automated measurement
	Cross-platform task success compared to the task success when the task is carried out in a single device (per device)	User testing [🕐 D] (Task-based): Automation support

	Cross-platform task time compared to the task time when the task is carried out in a single device (per device)	User testing [🕐 D] (Task-based): Automated measurement
Multi-user usability	Number of collisions with activities of others	User testing [🕐 D]: Automated measurement
	Correctness of system's conflict resolution	User testing [🕐 D]: Automation support
	Percentage of conflicts resolved by the system	User testing [🕐 D]: Automated measurement
	Percentage of conflicts resolved by the user(s)	User testing [🕐 D]: Automation support
	Social etiquette is followed by the system	Expert-based review
Implicit interactions	Implicit interactions carried out by the user	User testing [🕐 D]: Automated measurement
	Number of implicit interactions carried out by the user	User testing [🕐 D]: Automated measurement
	Percentages of implicit interactions per implicit interaction type	User testing [🕐 D]: Automated measurement
	Appropriateness of system responses to implicit interactions	Expert-based review User testing [🕐 D]: Automation support
Usage	Global interaction heat map: number of usages per hour on a daily, weekly and monthly basis for the entire AmI environment	User testing [🕐 D]: Automated measurement
	Systems' interaction heat map: number of usages for each system in the AmI environment per hour on a daily, weekly and monthly basis	User testing [🕐 D]: Automated measurement
	Applications' interaction heat map: number of usages for each application in the AmI environment per hour on a daily, weekly and monthly basis	User testing [🕐 D]: Automated measurement
	Time duration of users' interaction with the entire AmI environment	User testing [🕐 D]: Automated measurement

	Time duration of users' interaction with each system of the AmI environment	User testing [⏱ D]: Automated measurement
	Time duration of users' interaction with each application of the AmI environment	User testing [⏱ D]: Automated measurement
	Analysis (percentage) of applications used per system (for systems with more than one applications)	User testing [⏱ D]: Automated measurement
	Percentage of systems to which a pervasive application has been deployed, per application	User testing [⏱ D]: Automated measurement
APPEAL AND EMOTIONS		
Aesthetics	The systems follow principles of aesthetic design	Expert-based review
	The AmI environment and its systems are aesthetically pleasing for the user	User testing [⏱ sA] [⏱ lA]: Questionnaire
Fun	Interacting with the AmI environment is fun	User testing [⏱ sA] [⏱ lA]: Questionnaire
Actionable emotions	Detection of users' emotional strain through physiological measures, such as heart rate, skin resistance, blood volume pressure, gradient of the skin resistance and speed of the aggregated changes in the all variables' incoming data.	User testing [⏱ D]: Automated measurement
	Users' affective reaction to the system	User testing [⏱ sA] [⏱ lA]: Questionnaire
SAFETY AND PRIVACY		
User control	User has control over the data collected	Expert-based review
	User has control over the dissemination of information	Expert-based review
	The user can customise the level of control that the AmI environment has: high (acts on behalf of the person), medium (gives advice), low (executes a person's commands)	Expert-based review
Privacy	Availability of the user's information to other users of the system or third parties	Expert-based review

	Availability of explanations to a user about the potential use of recorded data	Expert-based review
	Expressiveness of the security (privacy) policy	Expert-based review
Safety	The AmI environment is safe for its operators	Expert-based review
	The AmI environment is safe in terms of public health	Expert-based review
	The AmI environment does not cause environmental harm	Expert-based review
	The AmI environment will not cause harm to commercial property, operations or reputation in the intended contexts of use	Expert-based review
TECHNOLOGY ACCEPTANCE AND ADOPTION		
System attributes	Perceived usefulness	User testing [🕐 sA] [🕐 lA]: Questionnaire
	Perceived ease of use	User testing [🕐 sA] [🕐 lA]: Questionnaire
	Trialability	Field study / In situ evaluation [🕐 sA] : Questionnaire
	Relative advantage	User testing [🕐 sA] [🕐 lA]: Questionnaire
	Cost (installation, maintenance)	Field study / In situ evaluation [🕐 sA]: Questionnaire
User attributes	Self-efficacy	User testing [🕐 B]: Questionnaire
	Computer attitude	User testing [🕐 B]: Questionnaire
	Age	User testing [🕐 B]: Questionnaire
	Gender	User testing [🕐 B]: Questionnaire

	Personal innovativeness	User testing [⏱ B] : Questionnaire
Social influences	Subjective norm	User testing [⏱ B]: Questionnaire
	Voluntariness	User testing [⏱ B]: Questionnaire
Facilitating conditions	End-user support	Field study / In situ evaluation [⏱ sA] [⏱ IA] : Questionnaire
	Visibility	Field study / In situ evaluation [⏱ B] : Questionnaire
Expected outcomes	Perceived benefit	User testing [⏱ B] [⏱ sA] [⏱ IA]: Questionnaire
	Long-term consequences of use	User testing [⏱ B] [⏱ sA] [⏱ IA]: Questionnaire
	Observability	User testing [⏱ sA] [⏱ IA]: Questionnaire
	Image	User testing [⏱ sA] [⏱ IA]: Questionnaire
Trust	User trust towards the system	User testing [⏱ B] [⏱ sA] [⏱ IA]: Questionnaire

In summary, the framework includes 103 specific metrics that can be collected through a combination of methods, as shown in Figure 36. More specifically, 20 metrics are assessed through expert-based reviews, 72 metrics through user testing, and 11 by both methods.

Although the number of metrics to be studied through user testing is large, evaluators will not be required to observe and collect data for all the 83 metrics. In particular, as shown in Figure 37, 30 (36.14%) of these metrics are automatically calculated by the AmI environment, 25 (30.12%) feature automation support, 2 (2.40%) need to observed manually, 25 (30.12%) will be obtained through subjective methods, and 1 (1.20%) should be acquired through interviews and manual observations. The 26 subjective metrics are proposed to be retrieved by means of interview (1), questionnaires (23), or both questionnaires and

interviews (2), when additional clarifications will be useful towards identifying potential UX problems or specific user attitudes.
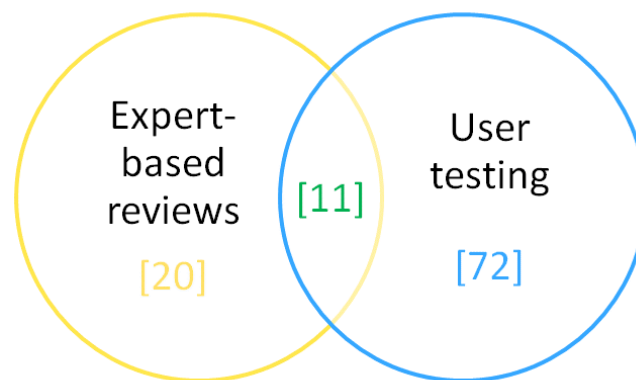
Figure 36. Distribution of metrics to specific methods
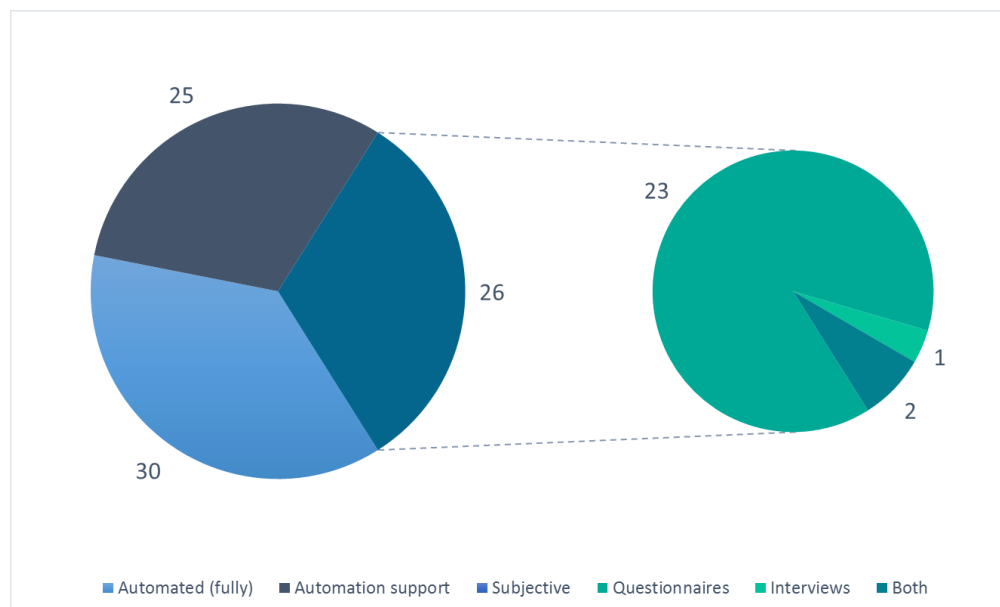
Figure 37. Analysis of metrics explored through user testing

In summary, the UXAmI framework proposes that UX evaluation of an AmI application system or entire environment should be carried out following a combination of methods and aims at minimizing the number of metrics that should be observed by the evaluator during an evaluation experiment with users. However, the role of experts and evaluators in the process is very significant. It is important to note that

human expertise cannot be substituted by any automated evaluation or simulation tool. Instead, these tools aim to provide aggregated metrics, and present them in an appropriate manner in order to facilitate human evaluators in understanding the results and combine them with their own findings and data collected from questionnaires, interviews, or any other usability and UX evaluation methods, so as to effectively comprehend and analyse user experience in an Ambient Intelligence environment.

Appendix D lists all the metrics and concepts, referencing to frameworks or theories from which the metrics were derived.

4.5   EVALUATION OF THE UXAMI FRAMEWORK

The proposed framework has been evaluated with the participation of six UX practitioners, three of whom were experts in the field, and three knowledgeable. All participants were familiar with the concept of Ambient Intelligence, while three of them had actually carried out a few evaluations of systems operating in AmI environments in the past. In particular, three of the participants were experts in AmI systems, having designed and developed systems for more than six years, two were knowledgeable, having less experience as designers of such systems, while one was familiar with such systems, however without any expertise in their design or development. In terms of evaluation of AmI systems, one participant was expert, having planned and carried out evaluations of such systems for more than four years, two were knowledgable with two years of active participation in such evaluations, while three were familiar with evaluations of AmI systems, having participated as observers in a small number (less than five) of such evaluations. Table 18 summarizes the aforementioned data regarding the evaluation participants.

Table 18. Evaluation participants' data

Age		Usability / UX expertise		AmI expertise		Evaluation in AmI	
20 – 30	2	Expert	3	Expert	3	Expert	1
30 – 40	2	Knowledgeable	3	Knowledgeable	2	Knowledgeable	2
40 – 50	2			Familiar	1	Familiar	3

The goal of the evaluation was twofold: (i) assessing if evaluators would plan and carry out a more detailed and inclusive evaluation with the UXAmI framework and (ii) evaluating the comprehensibility and usability of the framework and retrieving feedback from the evaluators. To this end, the following hypothesis were tested:

H1. Evaluators will plan a multi-method evaluation with the UXAmI framework.

H2. The number of metrics that evaluators will examine with the UXAmI framework will be larger (compared to the number of metrics that evaluators would plan to measure without the framework).

H3. The UXAmI framework is usable for evaluators.

Involving participants who are simply familiar with the concepts of usability and UX, having no practice in actually planning and running evaluations, was considered inappropriate for the context of the current evaluation. Participants should be at least knowledgeable in the field in order to be able to criticise and provide feedback on the framework constructs. Nevertheless, beginner UX practitioners can be involved in future evaluations, where they will be able to use the tools of the UXAmI framework in order to acquire concrete results.

4.5.1 PROCEDURE

A major goal of the evaluation of frameworks would be to assess how usable they are for the intended target audience (Heo et al., 2009), and retrieve qualitative feedback regarding their readability, understandability, learnability, applicability, and usefulness (Sommerville & Dewsbury, 2007). The evaluation of the proposed framework mainly targeted at retrieving qualitative feedback from evaluators regarding its usability, however a cognitive exercise was also included in order to retrieve some quantitative metrics as well. More specifically, the evaluation involved two phases: (a) planning an evaluation without the UXAmI framework and (b) planning the same evaluation with the framework. In order to place them in context, an introduction phase preceded, where participants were introduced to their role, being the lead UX expert in the design team of an AmI home, whose task is to plan, organize, and carry out evaluations of the systems being developed. In addition, participants were given a specific evaluation target, namely the TV system located in the living room of the AmI home and three short scenarios exemplifying its usage by the home residents. The scenario (given to participants as follows) exemplified not only the possible interaction and functionality of the television, but also addressed the topics of implicit interactions, system adaptation, multi-user usage, and system recommendations.

Living room TV (Interaction: gestures, speech, and remote control)

Jenny enters home after a long day at work. On her way home, she heard on the radio about an earthquake in her home island. Worried, she turns on the TV through the remote control. She switches to her favourite news channel through the remote control and turns up the volume by carrying out a gesture, raising up her palm that faces the ceiling. The news channel is currently showing statements of the Prime Minister

for a hot political topic. While listening to the news, she does some home chores and prepares dinner. She is cooking, when she listens that a report about the earthquake is presented and returns to the TV area. It turns out that the earthquake was small after all and no damages have been reported.

Jenny is on her way home, when she listens to the radio about an earthquake in her home island. She arrives at home, and in a hurry types her code in the home lock. As soon as she enters, the lights are turned on. Worried, she turns on the TV. She switches to her favourite news channel and turns up the volume. The news channel is currently showing statements of the Prime Minister for a hot political topic. Anxious to find out what happened she navigates to the news application of the TV to browse through the news, and see if she can find anything relevant. In the meantime, the environment has detected that she is stressed and starts playing her favourite jazz song, and at the same time dims the lights. Irritated she shouts "Music", "Stop". Browsing through the news, Jenny locates an article related to the earthquake. She selects it, reads it and eventually finds out that the earthquake was small and that no damages have been reported. Relieved she switches back to the news channel, and heads towards the kitchen to prepare dinner. Having detected her new location, the AmI Home transfers the TV sound to the kitchen speaker.

Peter has returned home from work and is currently reading the news through the living room TV. While reading, he receives a message from Jenny that she is on her way home and that he should start the dishwasher. Peter heads towards the kitchen (lights are turned on), selects a dishwasher program to start and returns to the living room (while kitchen lights are automatically turned off). After some time, Jenny arrives at home and unlocks the front door. As Jenny's preferred lighting mode is full bright, while Peter has dimmed the lights, a message is displayed on the active home display, the living room TV, asking whether light status should change to full bright. Peter authorizes the environment to change the lighting mode, welcomes Jenny and they both sit on the couch to read the news. Peter tells Jenny about an interesting article regarding an automobile company and the recent emissions scandal, and opens the article for her to read. Having read the article, Jenny recalls something interesting that she read at work about a new car model of the specific company and how it uses IT to detect drivers' fatigue. She returns to the news categories, selects the IT news category and they both look for the specific article. Peter reads it and they continue selecting collaboratively interesting news articles. After some time, and since they have to wait for Arthur – their 15 year old son – to come back from the cinema, they decide to watch a movie. The system recommends movies based on their common interests and preferences. Peter selects the movie, Jenny raises the volume, while the environment dims the lights to the pre-set mode for watching TV. Quite some time later, and while the movie is close to ending, Arthur comes home. As soon as he unlocks the door and enters, the lights are turned to full bright and the movie stops, since the movie is rated as inappropriate for persons younger than 16 years old. Jenny and Peter welcome their son, and then resume

the movie, as they think that it is not inappropriate for Arthur anyway, plus it is about to end. The movie ends and Jenny heads to the kitchen to serve dinner. Arthur and Peter browse through their favourite radio stations, and select one to listen to. The dinner is served, the family is gathered in the kitchen, and the music follows along, as it is automatically transferred to the kitchen speaker.

Having the scenarios, phase A of the evaluation was initiated, where participants were asked to think and organize the evaluation of the television, noting which methods they would use and what they would measure. They were given one day to think and plan their evaluation. After that, the evaluation method and metrics proposed by each participant were recorded. Following, the UXAmI framework was introduced by describing its main purpose, the multi-method approach advocated, the main AmI environment attributes assessed, as well as the full or partial automation support proposed in the context of user testing. Moving to phase B of the evaluation, participants were given printouts of the UXAmI framework and were asked to read it carefully and think again how they would plan this time the evaluation and also comment on metrics that were not understandable. They were given three days to prepare and plan their evaluation, taking into account that they had to read all the metrics and have the chance to comprehend how the framework works. It should be noted that they were not given a description of what each parameter means, or how important it is in the context of an AmI environment, and why it had been included in the framework. After completing phase B, evaluators' preferred metrics and comments were recorded. Finally, they were interviewed following a semi-structure interview approach featuring the following questions:

1. What is your overall impression of UXAmI?
2. Would you consider using it? Why?
3. Was the language clear and understandable?
4. What was omitted that should have been included?
5. What could be improved?
6. Would it be helpful in the context of carrying out evaluations in AmI environments in comparison to existing approaches you are aware of?

4.5.2 RESULTS

Analysis of the results of the evaluation revolves around the three hypothesis and explores if and how they are supported.

H1. Evaluators will plan a multi-method evaluation with the UXAmI framework

In Phase A (prior to using the framework), the following methods were employed for the evaluations that were planned:

- User testing: suggested by all six participants (100%)
- Expert-based reviews: suggested by only one participant (16.66%)

Specifically, in terms of user testing, the following methods were suggested (Figure 38 left): observation (6 participants: 100%), questionnaires (5 participants: 83.33%), interview (3 participants: 50%), Experience Recollection Method (1 participant: 16.66%), and UXGraph (1 participant: 16.66%).

In addition, two participants suggested that logs could be used, without however being able to explain how to use them or associate any specific metrics with this method.

In Phase B, as illustrated in (Figure 38 right), all the evaluators selected the expert-based review and the user testing method employing automated measurements, observation through automation support, as well as questionnaires. Interview was selected by 5 participants, while the methods of Experience Recollection and UXGraph were suggested to be used by the same participant who also employed them in phase A.
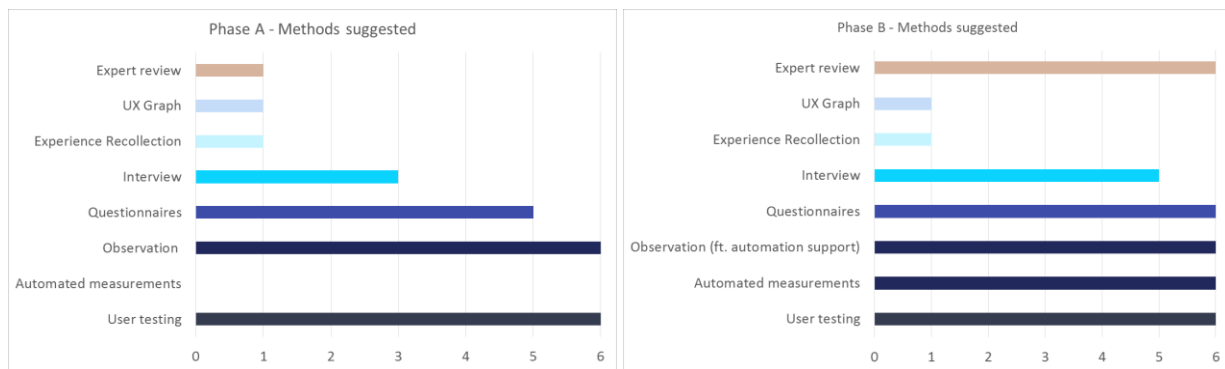
Figure 38. UXAmI Framework Evaluation: Methods employed in Phase A (left) and in Phase B (right

By comparing the results acquired in the two phases regarding the methodologies used, the following conclusions hold:

- Although in phase A only one participant selected expert-based reviews as a method to be employed, in phase B six participants selected it, embracing the multi-method approach advocated by the framework.

- Interviews were selected by two more participants in phase B.
- Automated measurements were selected by all the participants in phase B.
- Observations through automation support were selected by all the participants in phase B.

Based on the above, it is evident that hypothesis *H1* is supported, as a multi-method approach was selected by all the participants who used the UXAmI framework, although without it they had not catered for such a perspective and in their majority had focused on user testing only. Further looking at the metrics selected for each approach, it holds that in phase A only one participant employed expert-based review for a single metric. In Phase B however, not only the number of participants suggesting expert-based reviews increased, but also the number of metrics that would be assessed with the use of experts was much higher, leading thus to a more well-balanced iterative approach.

Table 19 provides the number of expert-based review metrics employed by each participant and in average in phases A and B, as well as the percentage of adoption in phase B of the UXAmI proposed expert-based review metrics, calculated as per Equation 1, where *p* is the number of metrics proposed by the participant and *30* is the total number of expert-based metrics proposed by UXAmI.

$$(p) = \frac{p}{30} \hspace{4cm} (1)$$

Table 19. Number of expert-based review metrics per phase

	P1	P2	P3	P4	P5	P6	AVG.
Phase A	0	0	0	0	0	1	0.16
Phase B	3	22	25	16	30	29	20.83
UXAMI adoption	10%	73.33%	83.33%	53.33%	100%	96.66%	69.44

H2. The number of metrics that evaluators will examine with the UXAmI framework will be larger (compared to the number of metrics that evaluators would plan to measure without the framework)

In Phase A, a total of 46 metrics was proposed by the participants towards measuring UX of the envisioned AmI system, some of which overlapped. The final list of metrics proposed was:

A. Observation
 1. Time to complete a task
 2. Number of times that an interaction modality is used

3. Interaction modality changes for a given task

4. Number of errors

5. Input errors

6. Interaction errors

7. Time spent recovering from errors

8. Number of help requests

9. Number of times that the user "undoes" automatic changes

10. Interaction modality accuracy

11. Interaction modality selected first

12. Task success

13. Number of tries to achieve a task

14. Unexpected actions or movements

15. User confidence with interaction modalities

B. Think aloud user statements

16. Input modalities that the user wanted to use but did not remember how to

17. Number of times the user expresses frustration

18. Number of times the user expresses joy

19. If the user understands the changes happening in the environment

C. Questionnaires

20. Age

21. Gender

22. Computer attitude

23. Preferable interaction technique

24. User satisfaction (questionnaire)

25. How well did the system manage multiple users?

26. Correctness of system adaptations

27. Level of fatigue

28. Users' experience of the intelligence

29. How intrusive did they find the environment

30. Effectiveness (questionnaire)

31. Efficiency (questionnaire)

32. User feelings

33. Learnability

34. System innovativeness

35. System responsiveness

36. System predictability

37. Comfortability with gestures

38. Promptness of system adaptations to user emotions

39. Comfortability with tracking and monitoring of activities

D. Interview

40. User feedback for each modality

41. Likes

42. Dislikes

43. Additional functionality desired

E. Experience Recollection Method (ERM)

44. User experience

F. UXGraph

45. User satisfaction from the overall user experience

G. Expert-based review

46. Functionality provided for setting preferences

Figure 39 illustrates the number of parameters suggested per participant during phase A. The distribution of the proposed metrics per method is illustrated in Figure 40, whereby it is evident that 22 metrics (36.66%) pertain to observed user behaviours, 40 metrics (63.49%) are user-reported (i.e., derived through statements vocalised in a think-aloud protocol, questionnaires, or interviews), and 1 metric (1.66%) is based on expert-based reviews. The exact number of proposed metrics per method and per participant is provided in Table 20. In general, it was observed that participants suggested metrics that were reasonable and important in the context of AmI environments (e.g., preferable interaction technique), however they typically resorted in measurements through users' self-reporting, with the exception of well-established usability metrics that were suggested to be measured, such as task success, time to complete a task, etc.
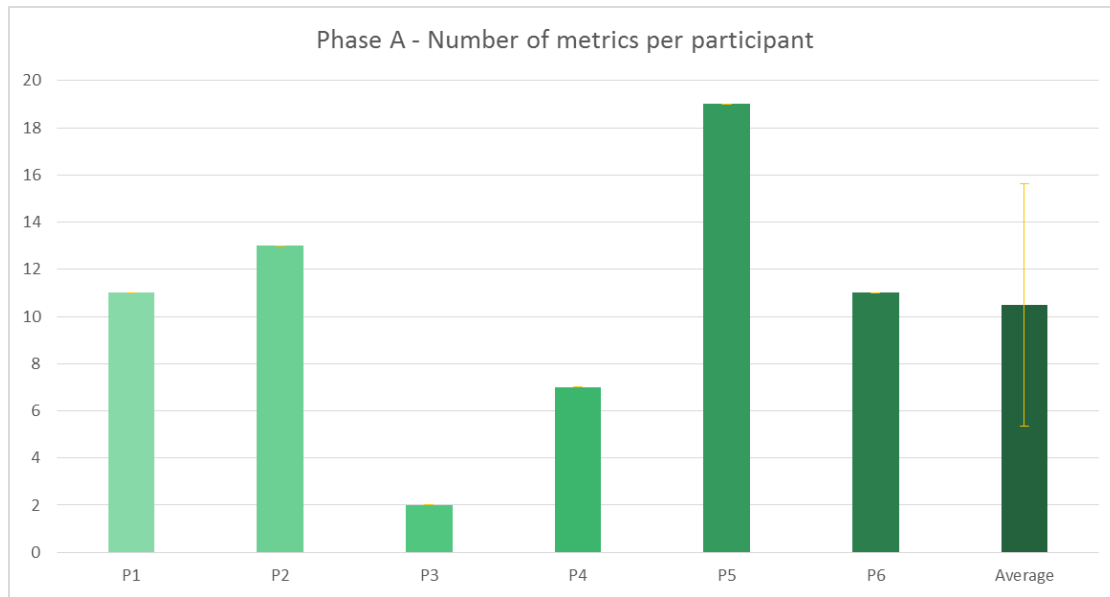
Figure 39. UXAmI Framework Evaluation: Number of metrics suggested per participant in phase A
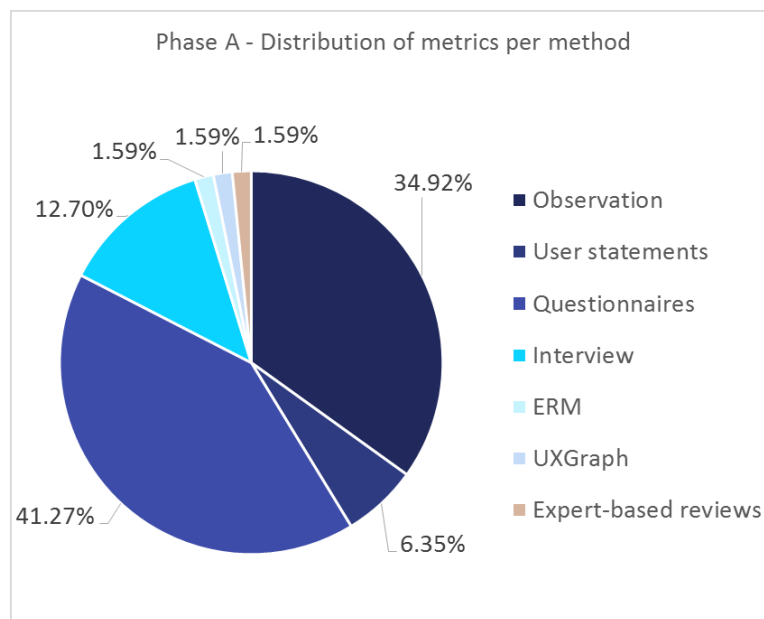
Figure 40. UXAmI Framework evaluation: Distribution of metrics per participant in phase A

Table 20. UXAmI Framework evaluation: Number of metrics per participant per method in phase A

Participant	Observation	User statements	Quest.	Interview	ERM	UXGraph	Expert	Total
P1	7	4	0	0	0	0	0	11
P2	3	0	8	0	1	1	0	13
P3	1	0	0	1	0	0	0	2
P4	4	0	1	2	0	0	0	7
P5	5	0	10	4	0	0	0	19
P6	2	0	7	1	0	0	1	11
Total	22	4	26	8	1	1	1	63

Several of the metrics proposed by the participants have not been included in the framework. Some of these will be considered in future versions, and some should not be included at all, for the reasons explained in Table 21 below.

Table 21. UXAmI Framework evaluation: Metrics proposed by participants that have not been included in the framework

OBSERVATION METRICS	
Interaction modality changes for a given task	Although a user may change the modality used during a task they are performing, this change does not necessarily imply that there is a problem. Therefore, as long as the user does not perform an input error with the specific interaction modality, this change cannot be associated with a UX problem or point of caution.
Interaction modality selected first	Even though the user may select a specific modality first, this does not mean that they will use it effectively and efficiently, or that they will develop a preference over it.
Number of tries to achieve a task	While a high number of tries is an indicator of a task that cannot be carried out efficiently, yet this metric will not add any new knowledge, given that other efficiency metrics are included in the framework, such as task success, time to complete a task, as well as monitoring of interaction and input errors. Furthermore, it is not clear what constitutes a "try" to achieve a task, and when a previous try ends, therefore it is a construct difficult to accurately measure.

Unexpected actions or movements	This is another ambiguous metric, which cannot be measured accurately and thoroughly. What constitutes unexpected action or movement for an evaluation observer might not be considered unexpected by another observer. Also, if seen out of context the metric can cause confusion and lead to wrong experiments, as for instance a user might be whistling in front of his living room television with the intention to interact with his coffee maker. To this end, the UXAmI framework constructs of input errors addresses the aforementioned concern.
User confidence with interaction modalities	This metric relies on the interpretation of user' behaviour from the evaluation observer, which is subjective. The actual usage of interaction modalities is addressed by the UXAmI framework through concrete observable metrics, therefore there is no need of including this construct as well.
Input modalities that the user wanted to use but did not remember how to	This metric is entirely based on user vocalizations of their intentions during a think-aloud experiment, therefore it is highly subjective and error-prone. The UXAmI framework has foreseen specific straightforward metrics towards assessing the learnability and intuitiveness of an AmI environment, system or application, therefore this indicator is not necessary to be included in the framework.
Whether the user understands the changes happening in the environment	An ambiguous metric, as an AmI environment should be intuitive and unobtrusive. The metric is proposed to be assessed through statements during a think aloud process. However, it is not clear whether noticing such changes is good or bad, while this interpretation would be a highly subjective, depending on the tone of the user statements. More objective metrics towards assessing intuitiveness and unobtrusiveness have been included in the UXAmI framework, therefore there is no need for including this metric as well.
USER REPORTED METRICS	
Preferable interaction technique	A user-reported metric aiming to identify which interaction techniques each user prefers. This has been anticipated in the framework through the observable metric of percentage of input modalities used.
How well did the system manage multiple users?	Although a valid question to ask as an evaluator, pursuing answers from users themselves is not an ideal solution. UXAmI includes a list of automated and semi-automated measurements providing concrete indications of how the system manages issues raising from multi-user usage.
Users' experience of the intelligence	An ambiguous metric, as on the one hand users may not perceive all the reasoning processes and decision-making on behalf of the AmI environment, and on the other hand the question itself is not straightforward and easy to answer. The

	concept behind this metric is of paramount importance in AmI environments, however it cannot be pursued and accurately measured through a single question addressed to users themselves.
How intrusive did they find the environment	Intrusiveness is one of the main AmI attributes that the UXAmI framework addresses, avoiding however to explicitly ask users about it. Rather, the framework anticipates measuring intrusiveness, through the unobtrusiveness metrics, as well as metrics pertaining to the appropriateness of adaptations and recommendations, and metrics regarding the impact of adaptations.
Effectiveness (questionnaire)	Effectiveness is a well-established metric of usability and has been foreseen by the framework through concrete observable measurements.
Efficiency (questionnaire)	Efficiency is a well-established metric of usability and has been foreseen by the framework through concrete observable measurements.
System innovativeness	Research on Technology Acceptance Models and theories has indicated that it is not the system innovativeness that impacts its adoption and acceptance. Instead, it is the user's personal innovativeness that has an effect, a metric already included in UXAmI.
System predictability	Even though the predictability of a system is advocated as a usability characteristic, it should not be measured through questioning users about it, who might even be unware of what it means and whether it is something good or bad. An unpredictable system will lead to increased user errors, depending on the domain that has not been designed to be predictable (e.g., input errors for unpredictable input vocabulary, interaction errors for unpredictable UI behaviour, adaptation rejections for unpredictable decisions and automations, etc.). The specific causes of unpredictability are already explored through the UXAmI framework.
Comfortability with gestures	A user-reported metric that is not expected to add insights to information already acquired through the automated measurements foreseen in UXAmI, namely usage of each interaction modality and errors per modality.
Promptness of system adaptations to user emotions	A very specific metric, inspired by the scenarios given to users, which is an instance of system responsiveness. UXAmI also includes several metrics regarding implicit interactions, to which emotions belong.
User feedback for each modality	Although it is important to receive user feedback, UXAmI has included specific metrics that provide insights regarding the interaction modalities. Evaluators of AmI environments and systems can additionally include questions in interviews,

however it is not considered necessary, as they will have enough insights and evidences through the UXAmI metrics. Also, the practicality of asking users about everything is questionable and with uncertain results.

Besides the aforementioned metrics, some suggestions made by the participants are quite important and should be included in UX evaluations. The parameters that should be considered are:

- Level of fatigue: an important consideration for the evaluation of systems supporting gestures. Although context-specific, as gestures are expected to be a fundamental interaction modality in AmI environments this metric will be included in future versions of the UXAmI framework, along with other metrics examining the most fundamental interaction modalities. The current version of the framework has focused on metrics pertaining to any AmI environment, no matter what its context and supported interaction modalities are. In addition, such specific concerns are expected to be studied by evaluators either as part of the "Conformance with guidelines" metric, where they will apply all the guidelines that may be of relevance to the specific system, or as part of the guideline "The AmI environment is safe for its operators".

- System responsiveness: a system characteristic which obviously impacts the overall user experience that should always be examined during software testing. Future versions of the framework will consider adding this variable to the expert-based measurements and simulations, however not as a user reported metric.

- Comfortability with tracking and monitoring of activities: a fundamental concern in AmI environments is whether users accept the fact that the environment collects information based on their activities. UXAmI has included attributes regarding safety, privacy, and user's control over the behaviour of the AmI environment. In addition, the trust metric in the acceptance and adoption category aims to retrieve users' attitude on how much they trust the AmI environment. Future versions of the framework will explore if a specific question for activity monitoring should be included as well.

- User likes and dislikes: two metrics pursued through interviews or questionnaires that may be considered by future versions of the UXAmI framework, taking into account the potential of their contribution against the induced complexity and load for the users, as the number of questions to be asked increases.

- Additional functionality required: it could be added as an interview question to be asked in the context of iterative design processes, as a means for eliciting user requirements.

- User experience: this metric could be included as an indication that additional methods estimating user experience as perceived by the users can be employed (e.g., how satisfied they are from the system during the various phases of using it).

In summary, as illustrated in Figure 41, out of the 43 metrics proposed, 17 (39.53%) are already included in the UXAmI framework, 19 (44.19%) are not and should not be included in the framework, based on the preceding analysis, while 7 (16.28%) will be considered in future framework versions, taking into account their actual impact against the complexity they may induce. Figure 42 illustrates the analysis of the proposed metrics against the ones included in the UXAmI framework, for observation and user-reported metrics separately. Regarding observation metrics, a total of 15 measurements were proposed, 11 (73.33%) of which are already included in the framework, and 4 (26.66%) not included currently or in future versions. The proposed user-reported metrics were 27, with 5 (18.51%) already included in the framework, 7 (25.92%) that could potentially be included in future versions, and 15 (48.14%) not (to be) included.
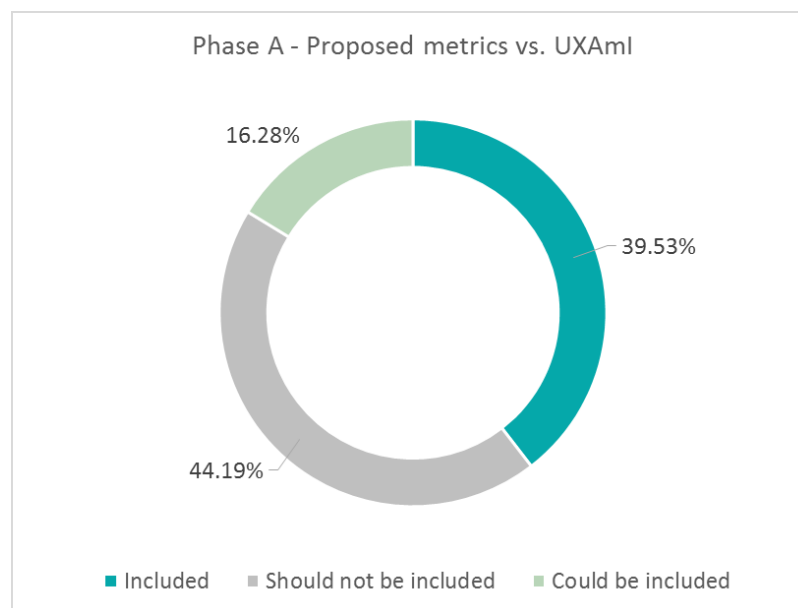
Figure 41. UXAmI Framework Evaluation: Proposed metrics against the UXAmI metrics
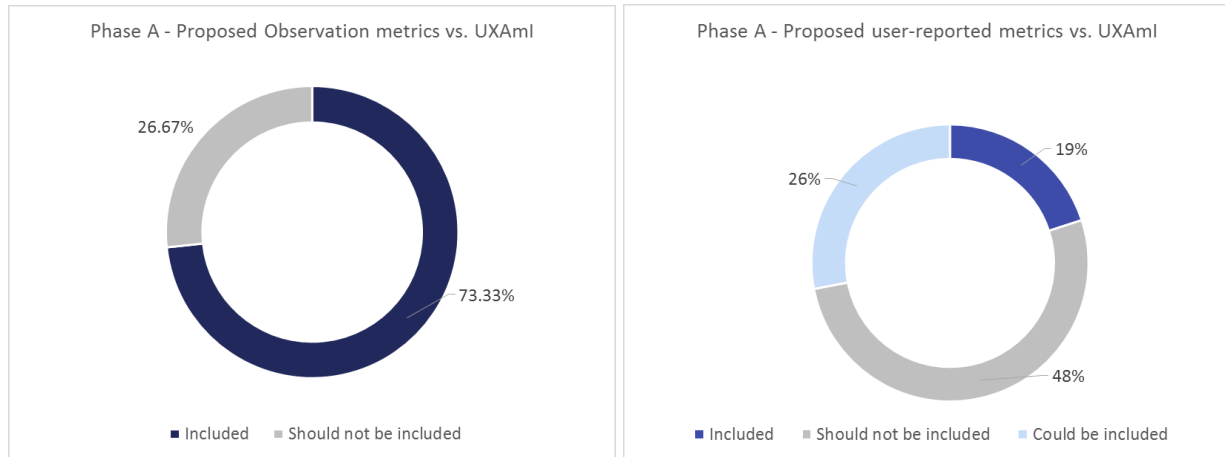
Figure 42. UXAmI Framework Evaluation: (Left) proposed observation metrics against UXAmI metrics; (Right) proposed self-rated metrics against UXAmI metrics

Examining the proposed metrics from the perspective of the AmI attributes and characteristics, excluding only parameters that were too ambiguous to be measured or asked to the users, therefore comparing only the quantity of metrics, aggregated according to the suggestions from all the evaluation participants, it turns out that a small proportion of attributes that should be examined in an AmI context was suggested to be included (Figure 43). In particular, the suggested metrics address the issue of User Experience in Ambient Intelligence environments in a rather low percentage (31.73%). It is noteworthy that certain attributes – although fundamental – are inadequately met, such as privacy and safety (10%), or adaptivity (12.50%) and adoption (26.31%). Moreover, the majority of attributes are only partially explored, e.g. unobtrusiveness (33.33%), usability (39.13%), as well as appeal and emotions (40%).
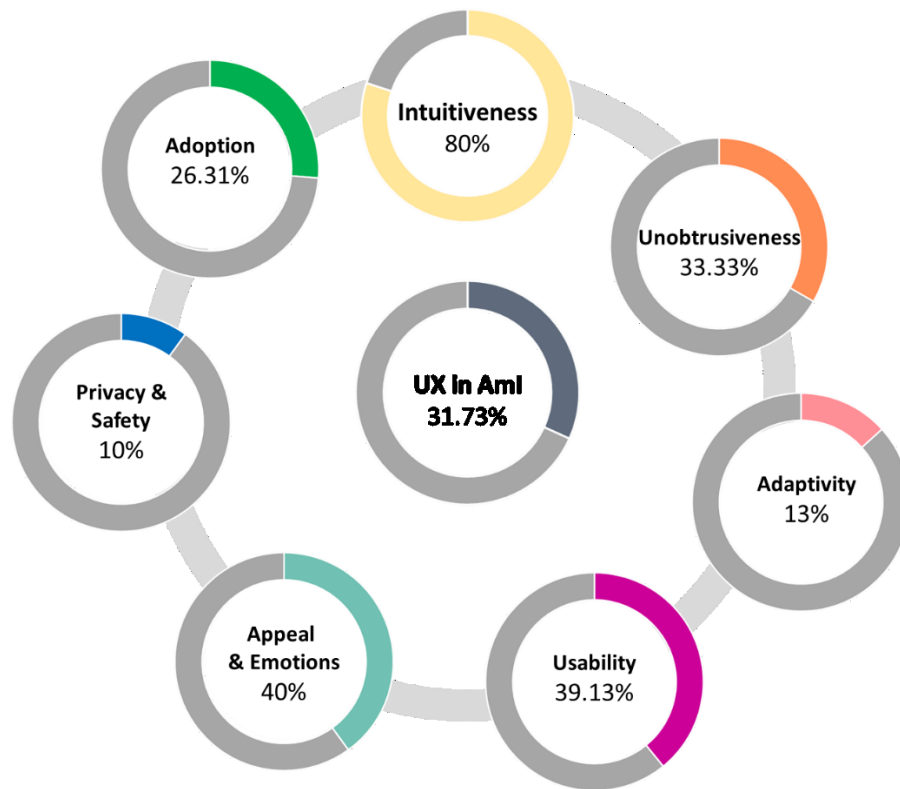
Figure 43. UXAmI Framework Evaluation: Proportion of AmI attributes evaluated by the aggregated metrics suggested by the evaluation participants

In Phase B, the number of parameters suggested by the participants was considerably larger. Overall, the aggregated number of suggested metrics is 103 (i.e., all the UXAmI framework metrics) plus one, namely perceived user experience. In particular, the evaluator who initially suggested using the ERM and UXGraph methods suggested to employ them again, however as these constitute generic user experience questionnaires they both correspond to one metric, namely perceived user experience, which will be included in future versions of the UXAmI framework. In Phase A, the aggregated number of metrics was 46, it is therefore directly evident that hypothesis H2 is supported, as the number of proposed metrics substantially increased.

Further, besides the aggregated number of metrics the individual number of metrics per participant also increased considerably, as illustrated in Figure 44. It is notable that the minimum number of metrics was 30 in phase B (P1), while the maximum number suggested in phase A was 19 (P5). This increase is clearly demonstrated by the increase in the average number of metrics proposed, which was 10.5 in phase A, against 74 in phase B.
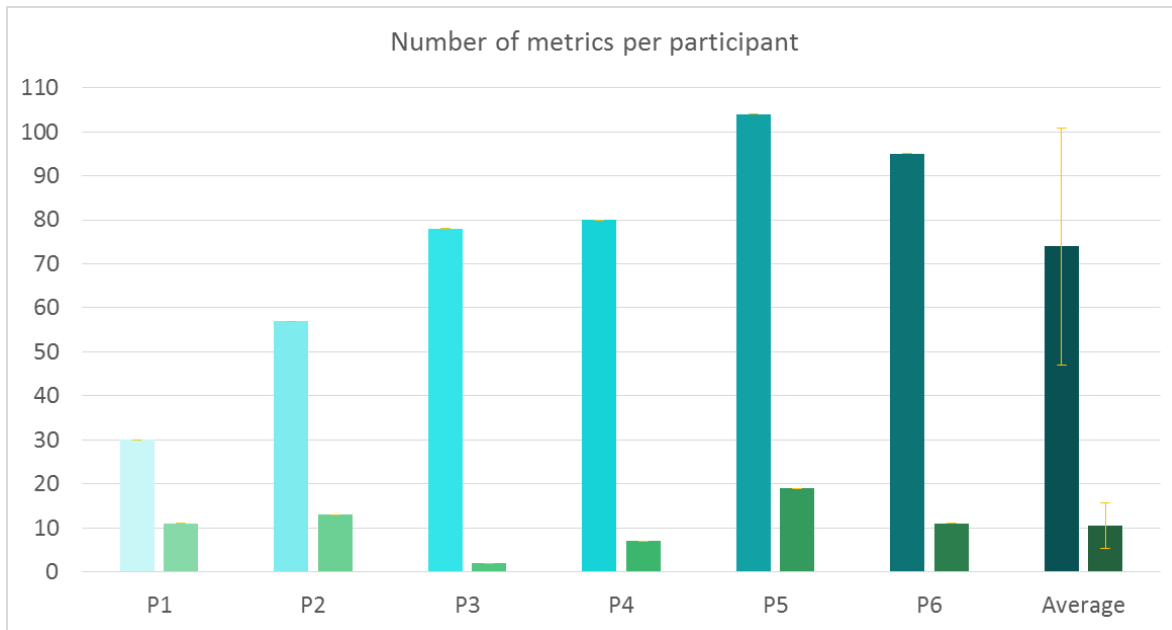
Figure 44. Number of metrics per participant in Phase A (green) and Phase B (teal)

The distribution of metrics to methods has also apparently changed when using the UXAmI framework, employing metrics assessed by expert-based reviews, and embracing all the automated metrics. As the entire set of UXAmI metrics has been involved in total by all evaluators, the metrics distribution per method is the one advocated by the framework: 31 metrics to be evaluated with expert-based reviews, 30 metrics to be automatically calculated by the AmI environment, 25 to be observed with automation support, 2 to be observed manually, 25 to be obtained through subjective methods, and 1 to be acquired through interviews and manual observations (note that 11 metrics are to be evaluated both by expert-based reviews and user testing methods). Finally, with UXAmI all the AmI attributes would eventually be assessed in their entirety by the six evaluation participants, in contrast to the extremely partial assessment of phase A (Intuitiveness: 80%, Unobtrusiveness: 33.33%, Adaptivity: 13%, Usability: 39.13%, Appeal and Emotions: 40%, Safety and Privacy: 10%, Adoption: 26.31%, Overall UX: 31.73%).

In conclusion, hypothesis H2 is supported as the number of metrics that were employed in phase B was greater for each participant individually, in average, and in total, having aggregated all the individual parameters suggested by each participant into one set.

H3. The UXAmI framework is usable for evaluators

To explore this hypothesis the participants' answers provided in the semi-structured interview that followed phase B are discussed.

Regarding their overall impression of UXAmI, participants indicated that it is complete, structured, thorough and in general very good. All six participants provided positive answers, an example being the following statement: *"A thorough and exhaustive collection of the most important evaluation metrics and heuristics, which is by itself very useful for the evaluator"*. In terms of understandability, all the participants agreed to the fact that all the metrics were clear and understandable, with the exception of certain specific metrics pertaining to Technology Acceptance. However, as one of them clarified, it only required a few minutes to refresh their memory of what these metrics mean by looking into the related theories. It should be mentioned that with the goal to assess how intuitive the metrics are, evaluators were not given any explanation or accompanying material regarding the metrics, although an extensive analysis is available in this thesis. To resolve this issue, a short list with terms and definitions has been prepared (Appendix B), which will accompany the UXAmI framework.

Evaluators were also asked what was omitted from the framework. Regarding omissions, all the evaluators agreed that they could not find any metrics or aspects of AmI environments missing. Some evaluators suggested employing expert-based reviews beyond user-based testing for four specific metrics. Their suggestions have been adopted and have already been included in the framework. One evaluator highlighted the need for been directed towards which questionnaires to employ, with an emphasis on standardized ones. Although the initial intention of the UXAmI framework was to allow evaluators to employ any specific user testing method and protocol, as well as any questionnaires they prefer, this suggestion will be adopted in future versions of the framework, which will include potentially useful questionnaires that could be used, without however forcing evaluators to adopt them.

Regarding improvements, the majority of evaluators suggested that the framework could be accompanied by tools to facilitate automated measurements and inspections, a concern that has already been addressed in the context of this thesis (Section 5). Furthermore, half of the evaluators suggested that they would have liked to have distinct tables for each method. The approach of one unified table was initially preferred, as on the one hand it provides an overview of all the metrics that fall under a specific AmI environment attribute, and on the other hand it makes clear that some metrics can and should be evaluated in a multi-method approach. Appendix C provides six tables offering this classification of metrics per method and per UX evaluation phase, and includes: metrics that should be assessed through expert-based reviews (Table 30), questionnaire-based metrics for user-based experiments to be acquired before the experiment (Table 31), observation metrics that can be automatically acquired with the help of the AmI environment during the experiment (Table 32), observation metrics regarding the experiment that need to be marked by the evaluator and receive automation support for calculations through tools (Table 33), metrics that should be pursued through questionnaires (Table 34) or interviews (Table 35) shortly

after the system usage in a user-based experiment, as well as metrics that should be acquired long time after the system usage through questionnaires (Table 36). Urged by the same need of easily retrieving metrics per method, two evaluators suggested that an electronic version of the framework, offering filters and step-by-step guides would also be useful, an observation that will be certainly followed up in future work.

Finally, evaluators were asked if they would consider actually using the framework and how helpful they think it would be in the context of AmI evaluations. All responses were unanimous, highlighting that they would definitely use the framework in any evaluation (not only AmI oriented), as it is thorough, systematic, well-structured, *"a real problem solver".* In addition, it was stressed that using the framework will reduce the time required for preparing and running an evaluation, and that one of its major benefits is that it minimizes the need for long questionnaires and lengthy interviews and substitutes them with actually measurable behaviours. Especially with regard to AmI environments, evaluators pointed out that it is the first framework that they know of regarding AmI environments, therefore it outweighs existing approaches. Further, the automated measurements it suggests are highly valuable and make it possible to collect data otherwise impossible to retrieve.

Based on the above analysis of evaluators' responses in the interview, it can be concluded that H3 is supported and that the UXAmI framework is not only usable, but actually useful and valuable for evaluators.

4.5.3 CONCLUSIONS

According to the analysis of hypotheses H1 and H2, it turns out that evaluators employed a more balanced approach in Phase B, where by using the UXAmI framework they were able to avoid estimations based entirely on user-reported perceptions and moved towards metrics objectively assessed through the environment itself, or through observed behaviours analysed systematically with the potential assistance of tools (automation support). It is also evident that they all realized the importance of expert-based reviews and decided to adopt an iterative evaluation approach, gaining all the benefits it promises. Moreover, with the help of the framework evaluators were able to plan a more thorough evaluation of user experience, based on metrics beyond the typical ones employed in usability evaluations (e.g., errors or task success), and to incorporate attributes of AmI environments that would have been otherwise neglected. Moreover, using the UXAmI framework, the evaluation catered for all the temporal facets of UX, namely before, during, shortly after, and long after using a product.

Evaluators' interview responses were analysed towards exploring hypothesis H3, affirming that the UXAmI framework is understandable, detailed, complete, and well-structured. All the evaluators acknowledged its usefulness towards any evaluation and highlighted its innovativeness in terms of evaluation in AmI environments. The usage of automated measurements was emphatically appraised, along with other benefits of the framework, such as that it provides a complete guide, facilitating evaluators in planning and carrying out thorough evaluations in a more "standardized" manner with minimum time required for preparation. These benefits will become even more concrete and substantial with the use of the UXAmI framework tools, allowing evaluators to retrieve guidelines for expert-based reviews and facilitating analysis of short- and long- term user-based experiments.

4.5.4 LIMITATIONS

When carrying out phase B, several participants assumed that they had to reproduce the same experiment as the one they had initially planned, leaving out several parameters that employed other methods beyond observation in user-testing. Furthermore, the scenario that was given as a guide towards understanding the functionalities of the system that would be evaluated, was considered by some evaluators as the scenario that would be given to users to execute, leading them to assume that they had to follow a task-based user testing approach. For those participants who reported these misconceptions, clarifications were given and therefore this limitation was addressed.

4.6 SUMMARY AND FUTURE WORK

This section has described the UXAmI conceptual and methodological framework, featuring concrete metrics with the aim to guide evaluators in planning and carrying out evaluations in AmI environments. The framework includes 103 metrics, organized in subcategories, which are in turn classified under seven fundamental attributes of AmI environments, namely intuitiveness, unobtrusiveness, adaptivity, usability, appeal and emotions, safety and privacy, and acceptance. The proposed AmI attributes, characteristics, and metrics are based on the extended literature review that has been carried out (Section 3), combining fundamental usability and UX concepts, as well as attributes of adaptive systems, UbiComp and AmI environments. Appendix D lists all the metrics and concepts, referencing to frameworks or theories from which the metrics were derived.

Besides combining knowledge from the aforementioned domains towards specifying which metrics should be assessed for the evaluation of AmI environments (or systems and applications therein), UXAmI proposes 39 novel metrics, mainly motivated by the need to complement expert-based reviews, user-reported metrics, and observers' remarks with objective measurable behaviours. To this end, novel

metrics have been proposed predominantly along four directions. First, taking advantage of the infrastructure of AmI environments it is possible to distinguish input errors from interaction errors, leading to the possibility to assess separately the interaction vocabulary and the user interface of applications. Then, these metrics - along with a few others indicating users' performance (e.g. help requests) - are studied in a variety of situations that may occur in an AmI environment or under different perspectives, for instance after an adaptation has been applied to assess its impact, after switching device to explore cross-platform usage behaviour, or over time to study the learnability of the system. Second, besides assessing the appropriateness of an adaptation or recommendation (through experts or users' self-reporting), the framework proposes to assess users' acceptance of the system's initiatives and recommendations, by exploring which and how many have been rejected. Third, an important activity that takes place in AmI environments, and is anticipated by the proposed framework, is that of implicit user interactions, which trigger the reasoning mechanisms of the environment and in specific circumstances lead to system actions. Finally, inspired by work on web analytics, a number of novel UXAmI metrics is related to the actual usage of the environment, its systems and applications, with the aim to provide insights to evaluators for the long-term usage of an AmI environment by its inhabitants.

An important contribution of the proposed framework is that it minimizes effort required by the observer in recording detailed notes for users' actions and behaviours, while it increases the objectivity of the metrics through the automated measurements it suggests. This contribution has been acknowledged in the evaluation of the UXAmI framework, along with several other contributions, such that it constitutes a means for carrying out consistent evaluations with the minimum possible effort in preparing and running them, as well as that it is systematic and thorough.

Future endeavours regarding the framework include exploring the addition of specific metrics for the most common interaction modalities employed in AmI environments and creating framework add-ons, that is, metrics to be accounted for in specific contexts (contexts of use, interaction modalities employed, or specific user categories, such as children or the elderly). Creating an electronic version of the framework constitutes a high priority future development, accompanied by specific suggestions of ready to use standardized questionnaires for user-based evaluations, and sets of guidelines for various contexts to be reviewed by experts.

5 FRAMEWORK TOOLS

UXAmI framework is accompanied by two main tools, one for each usability evaluation approach employed, as well as by a social networking platform for UX engineers that interoperates with the tools. These tools as well as the overall architecture to support them (Section 5.1) are presented in this chapter.

UXAmI Observer (Section 5.2) relates to experiments carried out with users in the laboratory or in the field and aims to combine data acquired from different sources during an experiment. Such data include video taken during the experiment, notes from the experiment observer, or data from automated measurements (e.g., psychophysiological metrics, system response metrics, user input commands, etc.). As a result, the evaluator acquires an aggregated view of the experiment along a timeline and is able to better comprehend and interpret user reactions through the various statistics available in the tool and therefore to assess the overall user experience.

To address the challenge of acquiring guidelines that should be considered during expert-based reviews of an AmI system or application, given the multi-platform and multi-modal interaction in AmI environments, UXAmI Inspector (Section 5.3) supports evaluators in retrieving guidelines relevant to the AmI environment project they are working with, guidelines that can be used both during design and evaluation. Furthermore, the tool supports the inspection process itself, facilitating error reporting and rating. UXAmI Inspector interoperates with the UXAmI Online Community to retrieve guidelines following a tag-based classification approach.

The UXAmI Online Community (Section 5.4) aims to become a knowledge resource and personal repository for UX engineers. Mechanisms to assist content contributions and address the challenges associated with tag-based crowdsourcing content classification have been embedded. Moreover, the community features an innovative adaptive reward scheme to ensure high quality and adequate quantity of user contributions towards reaching the critical mass required for succeeding and thriving.

The chapter ends by providing a use case (Section 5.5) of an AmI residence, to illustrate how the proposed framework and its tools can all be employed to carry out the evaluation following an iterative approach.

5.1 UXAMI ARCHITECTURE

The UXAmI architecture is based on the Service Oriented Architecture (SOA) model, decoupling the implementation of the UXAmI tools from the basic UXAmI software core. Such an approach entails several benefits in comparison to monolithic approaches, in terms of scalability, extensibility, high availability and data persistence (Bieberstein, 2006). Figure 45 illustrates the functional stack of the UXAmI architecture,

built on layered components, which are responsible for substantiating the functionality of the UXAmI core. Each component of a layer provides the necessary functionality and input to the components of the upper layers.
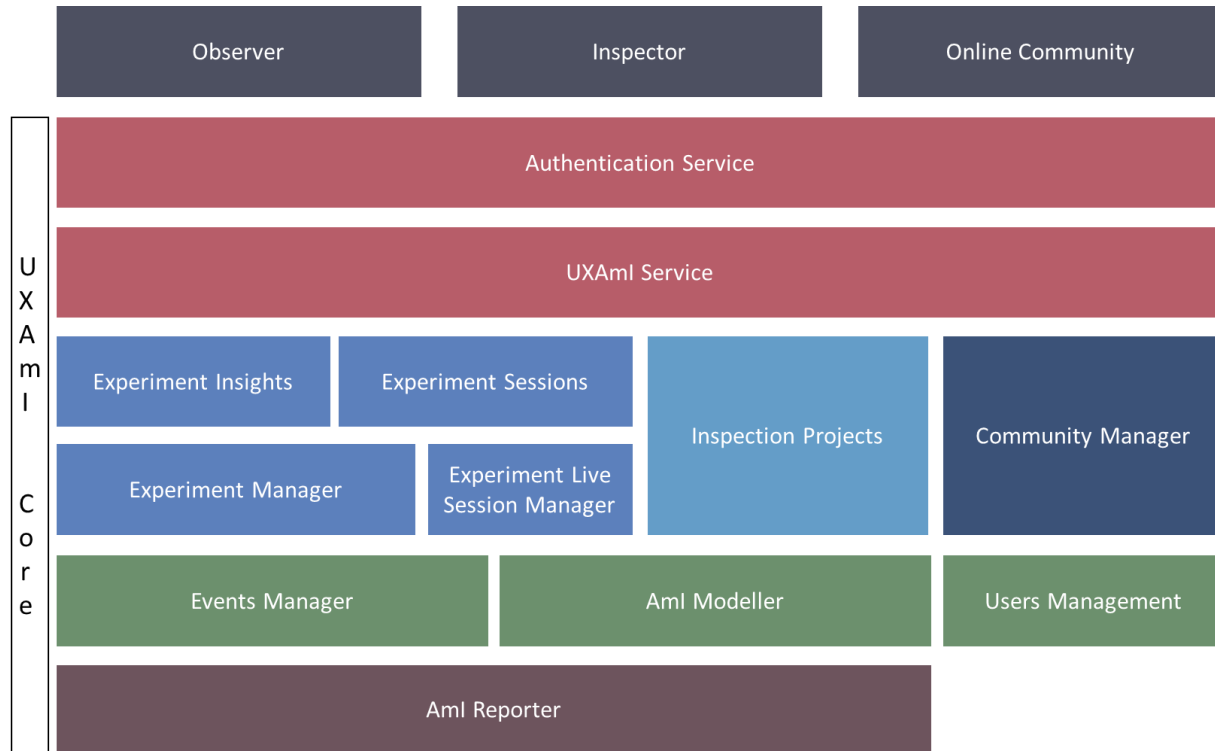
Figure 45. UXAmI Architecture

Specifically, *Authentication Service* and *UXAmI Service* constitute the UXAmI's core endpoints, providing RESTful APIs[10] so that applications, and therefore the users using the core, are able to authenticate and acquire the consolidated information regarding user-based experiments, expert-based inspections, as well as to access the UXAmI Online Community. Registration and authentication of an UXAmI user is provided through the *Authentication Service* endpoint. In more details, as illustrated in Figure 46, users can register themselves either through the UXAmI web front-end or directly with the UXAmI core (through a REST API call). After a user has registered, they must login to the system at least once in order to acquire a unique hash string (API token) that should be carried along within any other request to the UXAmI service. The lifetime of an API token is defined in the UXAmI core configuration.

---

[10] https://restfulapi.net/
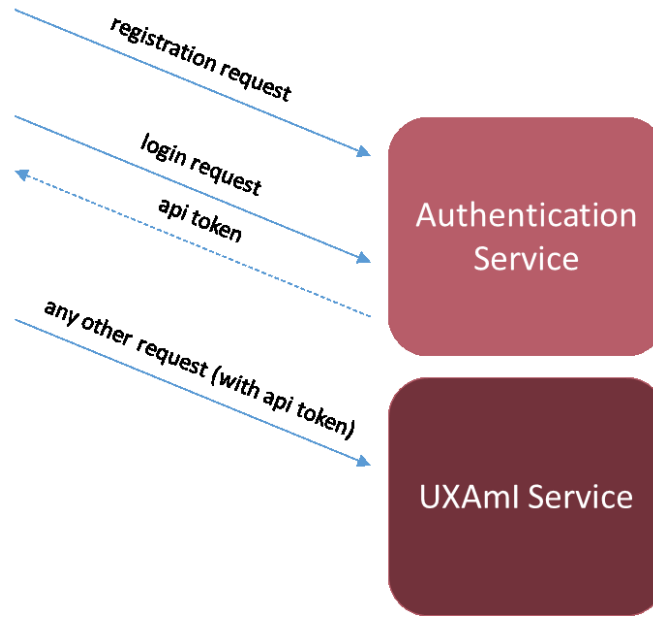
Figure 46. User authentication process

The *Experiment Sessions* component delivers information regarding the execution of a specific session of an experiment, acquiring data from the *Experiment Live Session Manager* (e.g., annotated POIs during the live execution of the experiment, session videos), the *Experiment Manager* regarding the basic parameters of the experiment (e.g., participants, tasks) and the *Events Manager* regarding the events that have occurred during this session (e.g., user input commands, system responses). The *Experiment Sessions* component is then responsible for calculating all the automated measurements, such as input error POIs or adaptation rejections. The *Experiment Insights* component provides consolidated statistical information, which has been acquired and analysed during the execution of an evaluation experiment or at a post-experiment processing phase, as described in section 5.2.3. The information is provided aggregated for all the participants of an experiment and all the systems and applications involved, as well as in sub-clusters pertaining to specific systems or applications. The *Experiment Insights* component acquires information from the *Experiment Sessions* component and interoperates with the *Experiment Manager*. The *Experiment Manager* component provides all the necessary functionality to create, read, update, and delete (CRUD) UXAmI experiments. The *Experiment Live Session Manager* handles live streaming of the experiment engaging all the available cameras, as well as recording of videos and observer's annotations so that they are available for post-processing, while also timestamping the session.

The *Inspection Projects* component facilitates the creation and maintenance of an inspection project for an evaluation target (applications and systems of the AmI environment). It is the component responsible for providing all the back-end functionality needed by UXAmI Inspector (Section 5.3). It interoperates with the *AmI Modeller* component to retrieve information about systems and applications in the AmI environment and their context, as well as with the *Community Manager* component to retrieve guidelines with tags that match the ones specified for the evaluation targets. Moreover, it provides CRUD functionality for inspection problems and inspection projects.

The *Community Manager* component supports the back-end functionality of the UXAmI Online Community (Section 5.4), instantiating the reward scheme and regulating access to the community contents according to the current user role (e.g., administrator, users of various membership levels). Also, it provides the necessary functionality for the maintenance and management of the community content. It interoperates with the *Inspection Projects* component as described above and with the *Users Management* component for acquiring the needed user profile information.

Events Manager constitutes the register for all the events originating from *AmI Reporter*. It also provides functionality for events' acquisition based on specific criteria, such as the set of events that belong in a specific timespan, or all the events of type system response, etc. It interoperates with the *Experiment Sessions* component, which employs the specific information.

AmI Modeller is the component responsible for modelling the information acquired from the AmI environment through the *AmI Reporter*, as well as all the entities of the UXAmI tools, defining also their interrelationships. The core relational scheme of UXAmI is depicted in Figure 47, omitting intermediate tables and pertinent relationships and focusing on the fundamental entities of events, actors, contexts and users. Events can be of three main types, namely interactions that represent user input, actions that relate to agents' information and responses referring to systems' and applications' feedback. Actors in an AmI environment can be humans, applications or agents, according to the type of source or destination of an UXAmI event. Contexts represent a space in the AmI environment where a system may reside, and are structured in a hierarchical manner, according to which each subspace declares the space it belongs to through the *parent_id* field. Moreover, for each context a geojson[11] file is assigned, describing its location and the 2D geometry of the space (floorplan).

---

[11] http://geojson.org/

Figure 47. UXAmI core scheme

The *Users Management* component provides back-end functionality for the management of users of all UXAmI systems (i.e., evaluators, administrators, and members of the Online Community). Finally, *AmI Reporter* is the intercommunication component of the UXAmI system with the AmI environment, incorporating heterogeneous services for the acquisition of the necessary information required by the UXAmI tools, and interoperates with *Events Manager* and *AmI Modeller*. *AmI Reporter* runs as an independent service and exposes a RESTful API for the population of information of the UXAmI core entities referred in Figure 47. This information originates from external agents, which are responsible to perceive the necessary information from the AmI environment and interpret this information to the appropriate format defined by the *AmI Reporter* API. Such an external agent has been implemented in the context of the current thesis, in order to use UXAmI Observer to carry out user-based experiments in the Living Room area of the Home Simulation Space of the FORTH-ICS Ambient Intelligence Research Facility. To this end, the agent subscribed to a REDIS channel [12] where the AmI environment agents and

---

[12] https://redis.io/

applications broadcasted messages, listened to the messages and interpreted them according to the *AmI Reporter* API.

The implementation of the UXAmI core is based on the Laravel PHP framework[13], while the realization of the Observer and Inspector web applications, as well as of the UXAmI Online Community are based on the Angular framework[14].

5.2 UXAMI OBSERVER

UXAmI Observer aims to support evaluators in carrying out user-based evaluations, be them laboratory task-based experiments, in situ evaluations, or long-term experiments. In a nutshell, the tool aggregates data regarding user's interaction with systems and applications in AmI environments and presents them through multiple views, such as timelines, charts, and diagrams. In task-based experiments the evaluator has to define the tasks and participant characteristics, whereas long-term experiments can be unstructured, employing users that are already registered in the system (e.g., the inhabitants of an actual AmI environment). Furthermore, the evaluator can view a user session live and provide annotations for it, or review the recorded data and further process them after the experiment. In brief, the tool provides two views for an experiment: (i) a view of each interaction session, named Timeline, and (ii) insights from the entire experiment, based on all the users that are involved in it throughout the experiment period. Figure 48, Figure 49, and Figure 50 illustrate the main tasks that an evaluator can carry out with the tool in the form of Hierarchical Task Analysis (HTA) diagrams (Stanton, 2006).

UXAmI Observer supports evaluators by providing automated calculations, assistance for manually inputting information regarding the experiment, as well as statistical and usage information. The sections below are structured as follows: section 5.2.1 presents the concept of experiment and how it is represented in the tool, while sections 5.2.2 and 5.2.3 present the timeline and insights views respectively. Experiments can also be compared, allowing evaluators to have an overview of the evolution of specific targets (systems and/or applications), as described in Section 5.2.4. Section 5.2.5 discusses the relation of UXAmI Observer with the UXAmI framework. UXAmI Observer was developed following an iterative approach, starting with low-fidelity (Lo-Fi) prototypes that were developed with the Evolus Pencil tool[15] and are presented in Appendix E. Subsequently, UXAmI Observer has been evaluated by three UX experts, as reported in Section 5.2.6. Finally, conclusions and discussion on future developments of the tool are provided in section 5.2.7.

---

[13] https://laravel.com/
[14] https://angular.io/
[15] https://pencil.evolus.vn/
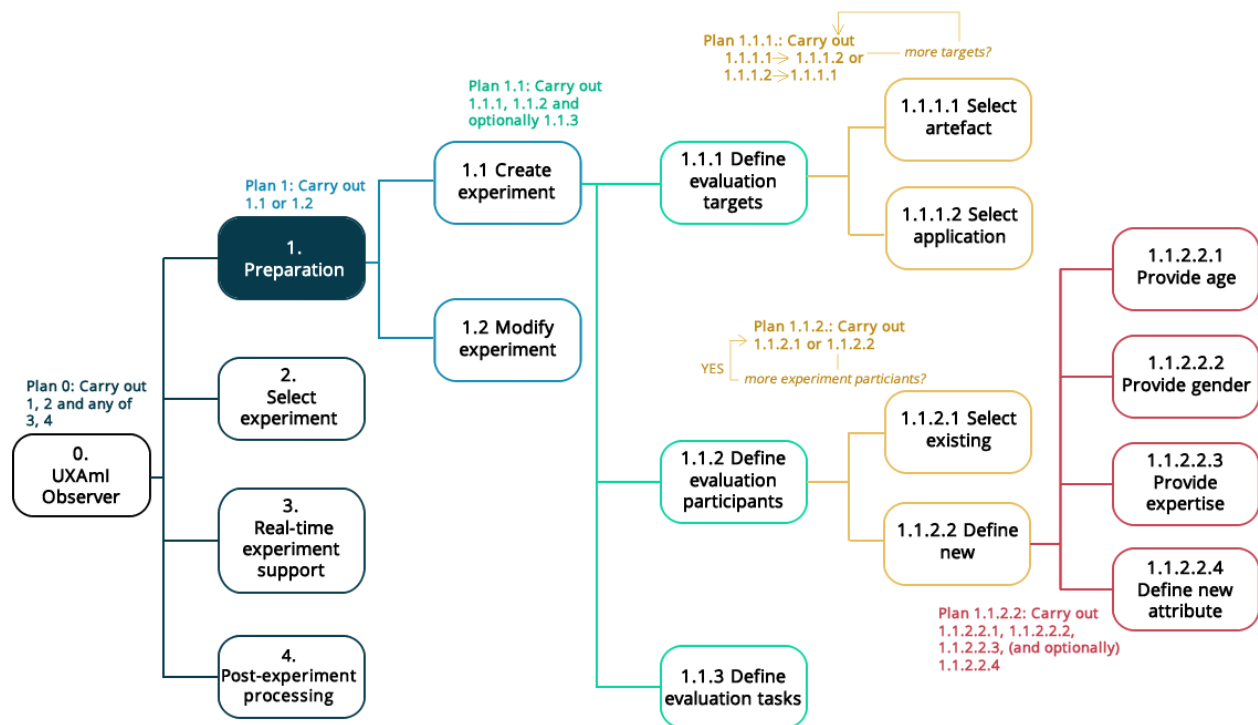
Figure 48. UXAmI Observer HTA diagram: Analysis of the "Preparation" task
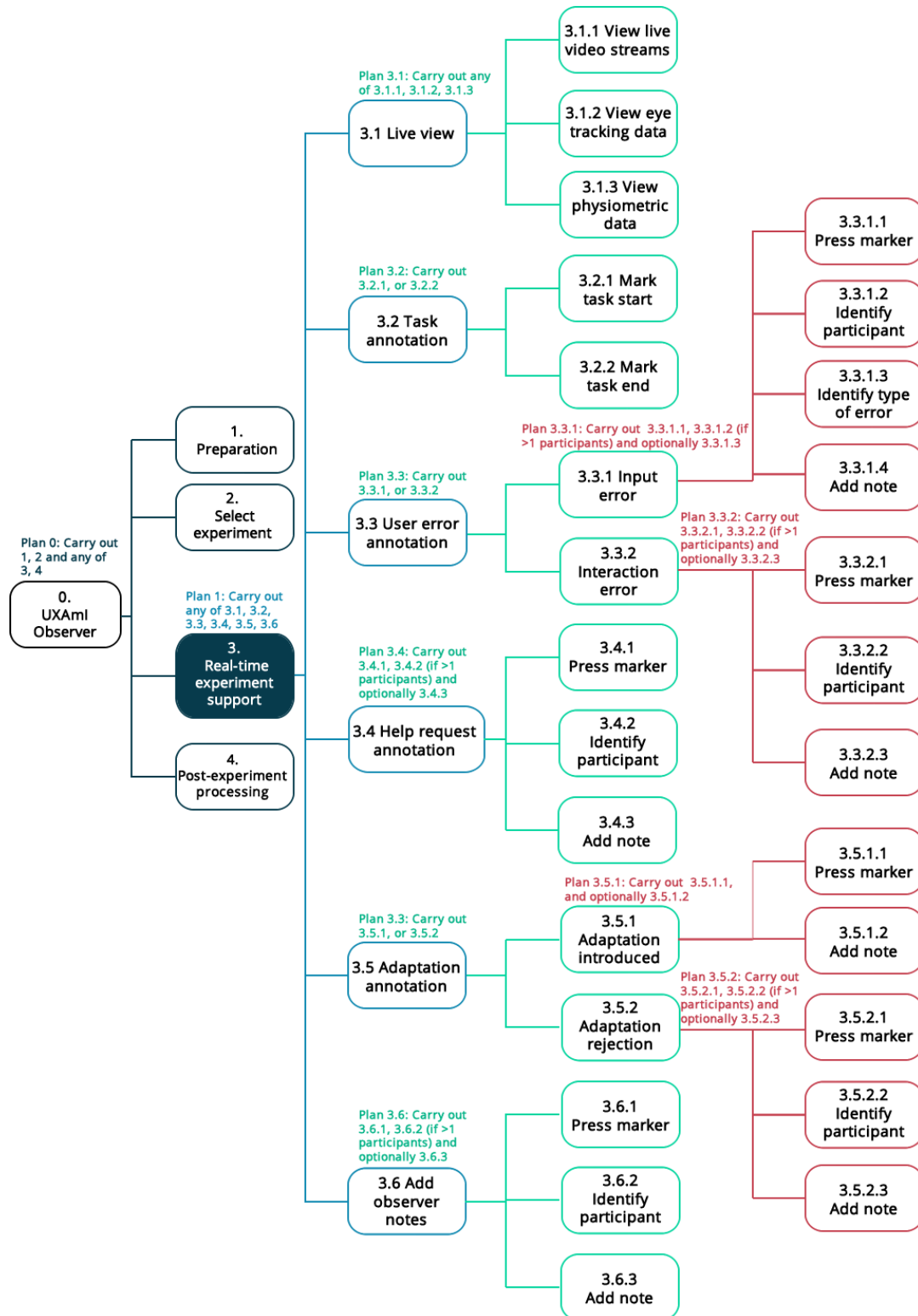
Figure 49. UXAmI Observer HTA diagram: Analysis of the "Real-time experiment support" task
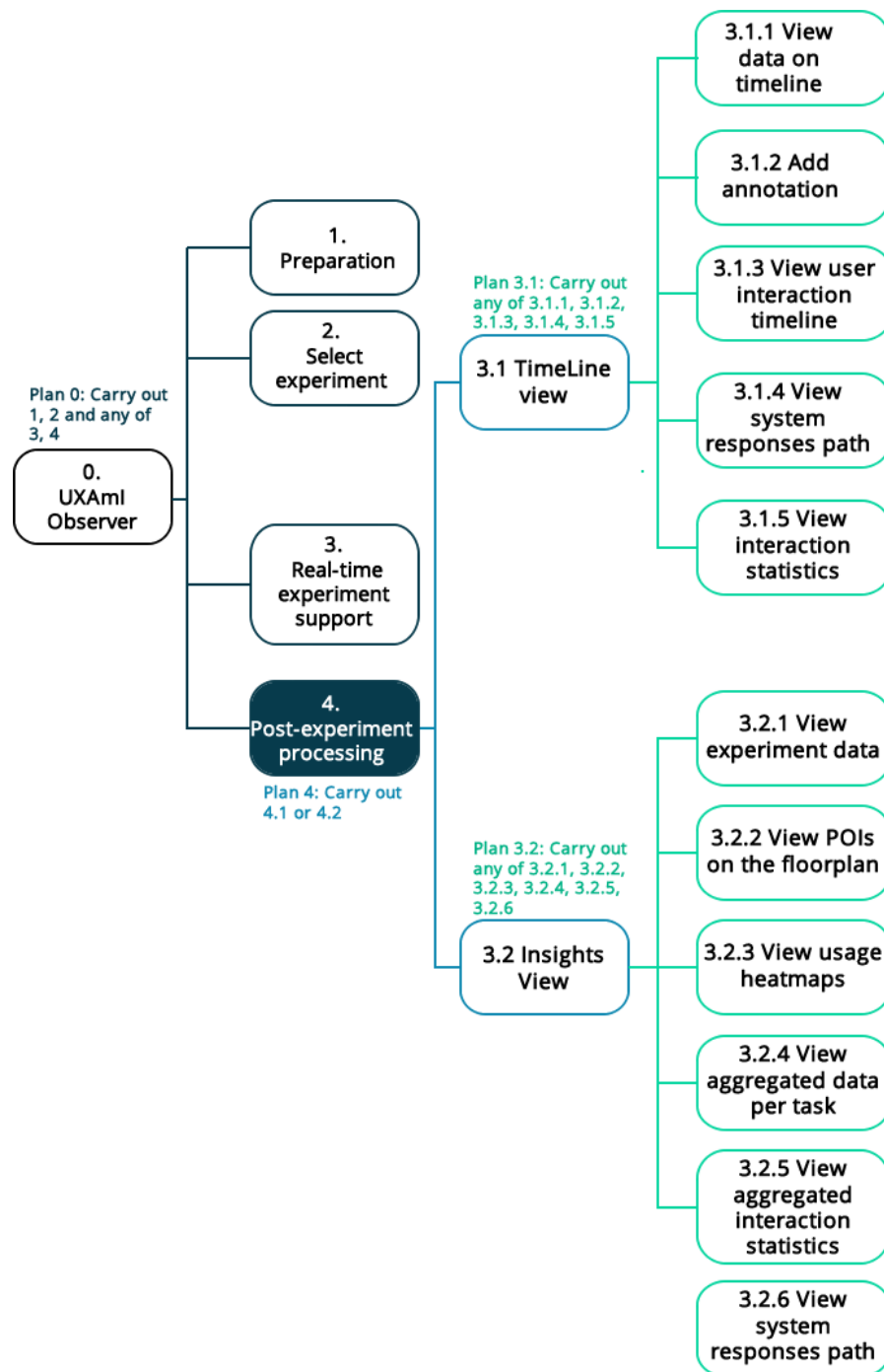
Figure 50. UXAmI Observer HTA diagram: Analysis of the "Post-experiment processing" task

5.2.1 EXPERIMENTS IN UXAMI OBSERVER

An experiment in UXAmI Observer is mainly defined by the evaluation targets, facilitating evaluators in organizing the acquired data according to the AmI spaces, artefacts, or applications that they wish to study. More specifically, an experiment includes (Figure 51):

- The evaluation targets, that is the involved artefacts, applications and the relevant contexts (AmI spaces where the artefacts are located).
- A name to facilitate its identification.
- A description of the experiment and its goals, as well as a photograph.
- The evaluators involved and their expertise, as indicated by the evaluators themselves in their profile. Evaluators expertise is rated on scale from 1 to 4 (1: not knowledgeable, 2: passing knowledge, 3: knowledgeable, 4: expert)
- Confidence in the evaluation experiment, as rated by the evaluators on a scale from 1 to 5.
- Tasks (if any).
- Participants.
- Sessions, that is, usages of the system by one or more participants concurrently.

When creating a new experiment, the evaluator has to first define the evaluation targets, provide a name for the evaluation and a short description, as well as a representative photograph. The process of selecting targets in UXAmI Observer is identical to the first steps of identifying evaluation targets in UXAmI Inspector (Section 5.3.1) and involves either selecting an entire AmI space and subsequently refining selections to specific artefacts and applications in the space, or selecting a pervasive application and subsequently refining the artefacts to which the application runs and which constitute the evaluation targets. Once the experiment has been created, the evaluator can add tasks and participants. Defining a task is optional, refers to task-based experiments, and requires providing a short description for each task (e.g., "Turn the TV on"). Adding participants is mandatory, and is achieved by selecting from a list of existing participants, or by adding new users to the system through defining their age, gender, and computer expertise level. According to the experiment carried out, the evaluator can proceed to defining further participant attributes by providing a name for the attribute, as well as its potential values, through identifying the values' scale (e.g., 1-5) and providing a string to describe each point of the scale. For instance, if one would like to define a new attribute for computer expertise, the following could be defined:

- Attribute name: computer attitude
- Attribute scale: 1-3
- Attribute scale values: negative; neutral; positive.
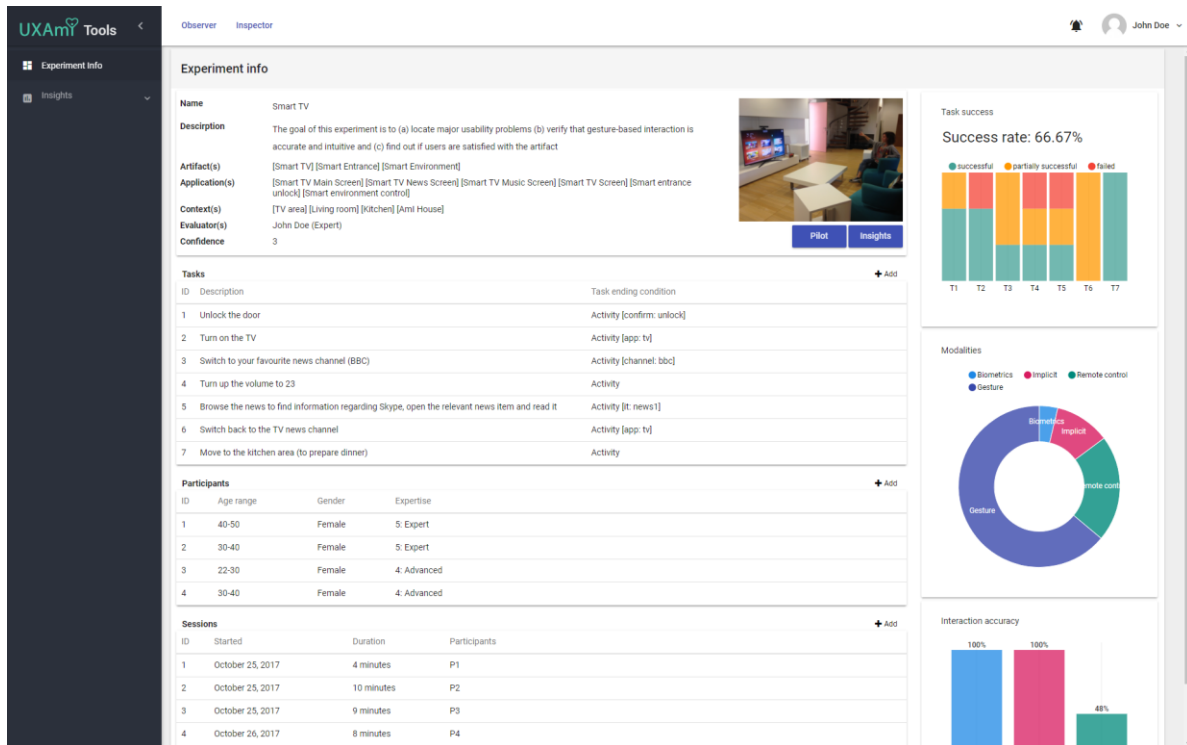
Figure 51. UXAmI Observer: Experiment information

To initiate recording of a session for a participant, the evaluator has to select the "Live" button, which becomes available upon placing their mouse over the corresponding line in the participants' table (Figure 52). Each participant may take part in more than one sessions, according to the current experiment needs.

Figure 52. UXAmI Observer: Initiating session recording

As soon as the live session ends, it is added to the sessions' list and becomes instantly available for the evaluator to review all the data that has been recorded, as well as the automatic calculations provided by UXAmI Observer, through the session Timeline (Figure 53).

Figure 53. UXAmI Observer: Activating session timeline

Additionally, a pilot session can be recorded for the specific experiment. Typically, at least one pilot session needs to be executed before actually carrying out a user-based experiment (Nielsen, 1994b). This pilot serves as a "test for the test", allowing evaluators to test the scenario given to participants, any recordings they plan to make, the instructions they plan to give and the questions to ask. The pilot test results are not employed in any statistics calculated for the experiment, however in UXAmI Observer pilots can be used to train the tool to identify when a task successfully ends. More specifically, in task-based experiments one needs to know when a task ends in order to calculate statistics per task (e.g., task duration, number of errors per task). The evaluators can mark when a task ends either through the live view by pressing the appropriate marker, or in the session timeline view (section 5.2.2). When viewing the pilot timeline, evaluators can mark the ending condition of a task, by clicking on the respective system response and identifying the task for which this response is the ending condition. Therefore, for all the sessions pertaining to this experiment, the system proposes as ending condition of a task the first instance of the condition as it was defined for the pilot test. Evaluators can always change this suggestion and define a different ending condition for any specific user session.

Through the experiment screen the evaluator can also choose to view the experiment insights, including the aggregated statistics and information regarding all the sessions that have been executed for the experiment. A preview of three indicative statistics is readily available through the experiment screen, and in particular a bar chart representing the task success rate score in total and per task (if the experiment includes tasks), a pie chart illustrating the usage (%) of interaction modalities, and a bar chart presenting interaction accuracy per interaction modality.

5.2.2   TIMELINE

The timeline contains information about an experiment session that can be clustered under four main themes: (i) session timeline, with all the recorded points of interest (POIs) marked along a horizontal timeline (ii) interaction timeline, with explicit indications of task start, task end, user input commands and implicit interactions, system responses and adaptations, ordered in a vertical timeline according to the time of their occurrence, (iii) system responses path, indicating all the system responses during the experiment, and (iv) interaction statistics.

Topmost is available the session timeline (Figure 54), aiming to give to the evaluator an overview of the entire session and all the data that have been recorded, synchronized with the session videos. Up to three videos are supported, ideally one illustrating the screen of the system with which the user interacts. If an eye tracking service has also been used for the experiment, the evaluator can enable annotation of the video with the eye tracking data. Also, all the data pertaining to the experiment are marked on the

horizontal timeline, including adaptations that have been introduced by the system, adaptation rejections, input error POIs, emotion POIs, interaction errors, help requests and observer notes. The user can select to play the video, while a red vertical line runs across the horizontal timeline to indicate the current time, facilitating association of POIs with the video displayed above.
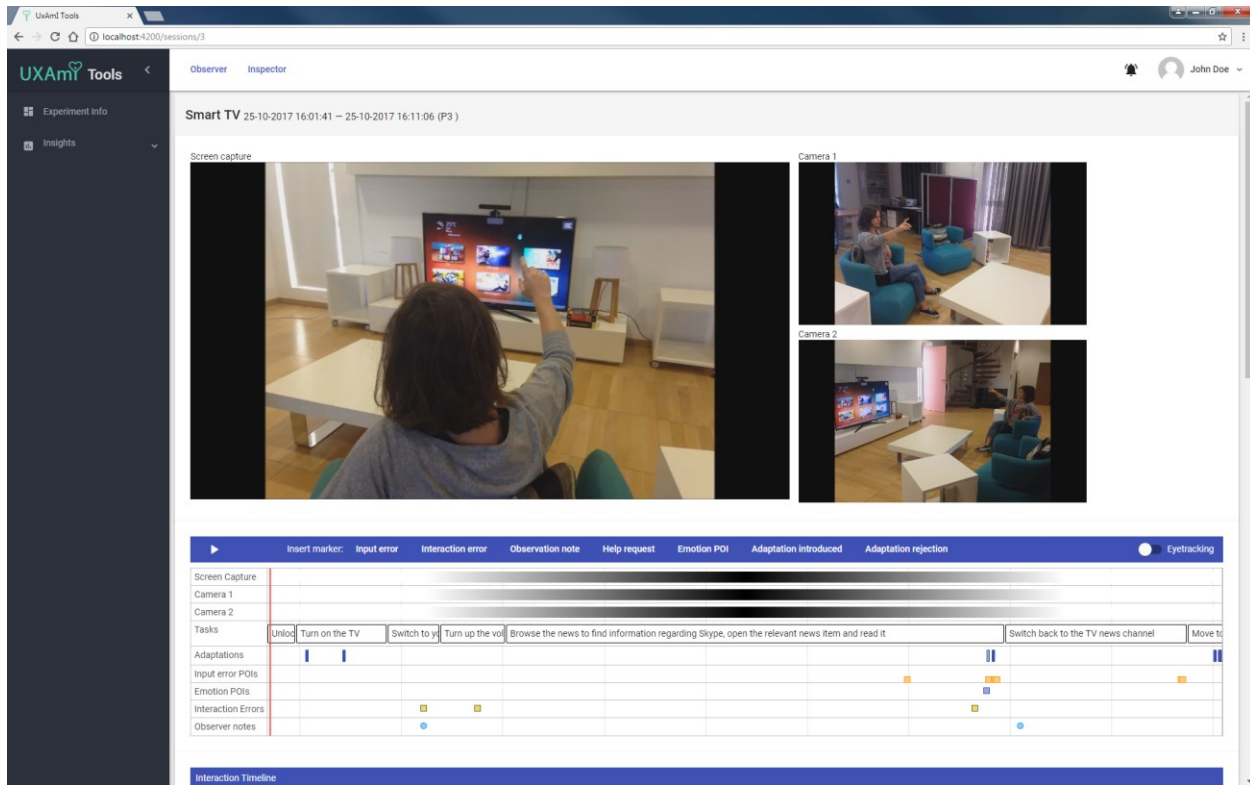
Figure 54. UXAMI Observer – Timeline: Session timeline

The POIs annotated on the horizontal timeline can be based on data manually provided by the observer or on automatically acquired information. Regarding manual input, the evaluator can annotate any of the aforementioned POIs during the live execution of the experiment, or during post-processing by selecting the corresponding marker indication above the horizontal timeline. For each marker a dialog is displayed to facilitate the evaluator in providing the required details (Figure 55), which may include the participant (if more than one users participate in one session), notes, and the type of input error in case the evaluator is marking such an error. The list of participants is populated only with the participants that are relevant for the current session, while the list of input errors types includes only the interaction modalities supported by the systems involved in the experiment. Table 22 summarizes the data that may be provided for each different type of POI on the timeline.
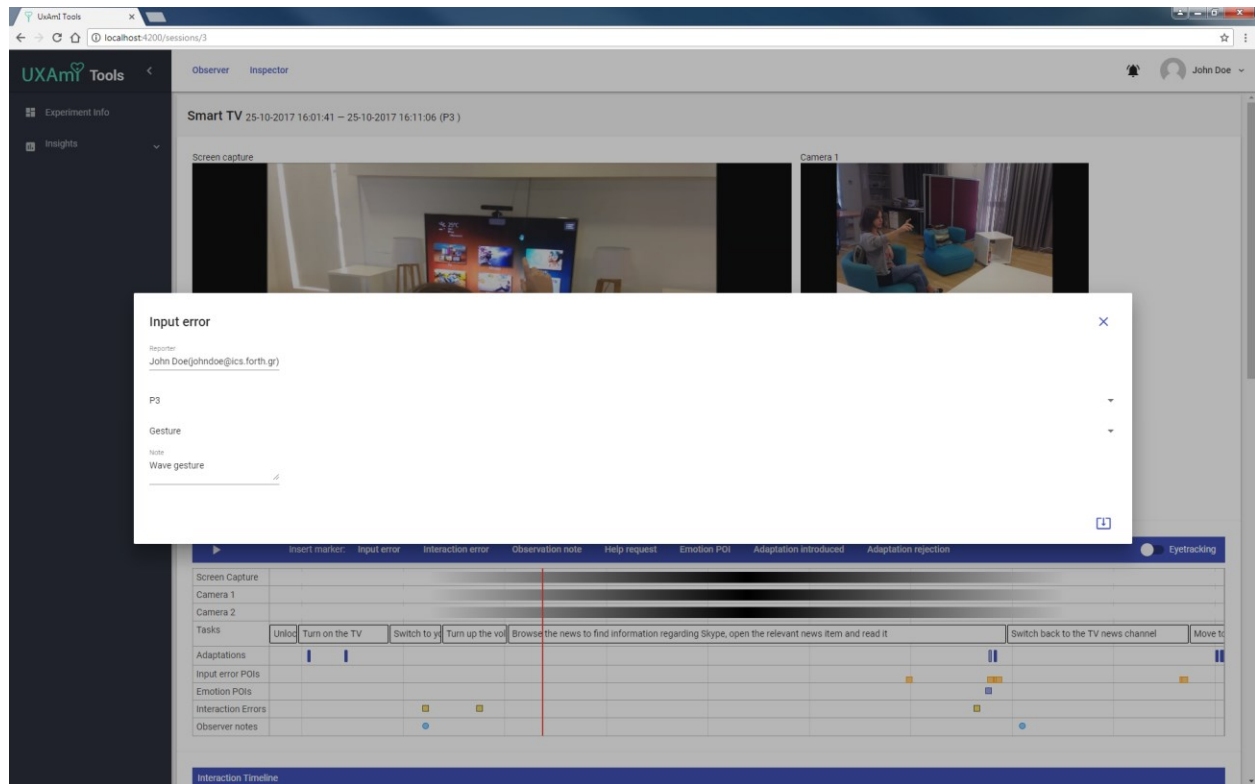
Figure 55. UXAmI Observer: Manual POI recording

Table 22. Data associated with each POI of the UXAmI Observer timeline

	Error type	Participant(s)	Notes
Input error	✔	✔	✔
Interaction error		✔	✔
Emotion POI		✔	✔
Help request		✔	✔
Adaptation introduced			✔
Adaptation rejection		✔	✔

Besides manually acquired data, UXAmI Observer includes in the timeline the following automatically detected points: adaptation insertion and adaptation rejection, input error, emotion and implicit interactions. Adaptation insertion points are identified based on events received from the reasoning

agents of the AmI environment, agents that will certainly be involved before any adaptation. Cook et al. (2009) identify the contributing technologies in an AmI environment and identify that the AmI algorithm perceives the state of the environment and users with sensors, reasons about the data using a variety of AI techniques, and acts upon the environment using controllers in such a way that the algorithm achieves its intended goal. A typical information flow in an AmI environment, as described by Augusto and McCullagh (2007) and illustrated in Figure 56 (left), clearly involves reasoning mechanisms after any sensed interaction and before any decision making. Augusto et al. (2010) discuss the role of vision and sensor networks and their interaction with high-level reasoning mechanisms (Figure 56 right), and report that the information acquired from vision and sensor networks is transferred to high-level reasoning modules for knowledge accumulation in applications involving behaviour monitoring, or for reacting to the situation in applications based on ambient intelligence and smart environments. UXAmI Observer takes advantage of this structure that is inherent in AmI environments and - as illustrated in Section 5.1 – "listens" to information propagated by the reasoning agents, and accordingly marks an adaptation insertion POI on the timeline.
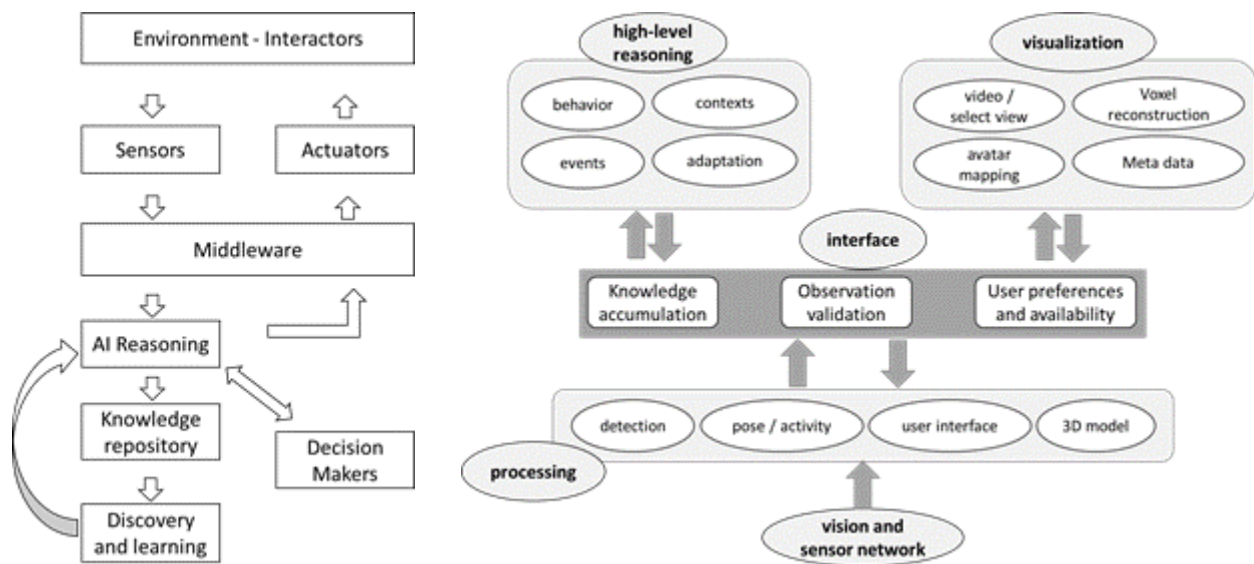
Figure 56. (Left) Typical information flow in an AmI environment (adapted from Augusto & McCullagh, 2007); (Right) The role of vision and sensor networks in interaction with high-level reasoning and visualization (adapted from Augusto et al., 2010)

Adaptation rejection points are inferred by monitoring the state of the system that was affected by the adaptation and checking if this state is changed by a user, according to the pseudo-code described in Figure 57. Two main challenges had to be addressed in this context, related on the one hand to accurately identifying the systems that are affected by the adaptation and those which respond to the user input, and on the other hand to defining the cut-off point after which a change in the system will not be

considered an adaptation rejection. Regarding the first concern, UXAmI Observer identifies as affected systems the ones that are detected to change state after the information propagation by a reasoning agent and before any user interaction or other agent event. It should be noted that more than one systems may be considered as relevant, as a decision of a reasoning agent may cause adaptation to more than one systems (e.g., lights are turned on in the kitchen, and interaction with the cooking assistant is switched to speech-based). Following the same rationale, all the systems that change state after a user input event and before any other user input or agent event are considered to be affected by a user input. A potential counterexample of the described rationale would be system responses that interleave the expected flow due to network problems or other delays or incorrect behaviours. This on the one hand constitutes a problem that should be detected by the evaluator, therefore it is expected that misjudgements by UXAmI Observer will draw the evaluator's attention towards locating problematic behaviours, an identification which is facilitated by the vertical timeline of events that is presented below. On the other hand, whenever an adaptation rejection is detected, it is appropriately annotated on the vertical timeline and the tool asks evaluators to confirm this inference. Regarding the cut-off point after which a change in the state of the affected systems will not constitute an adaptation rejection, it was decided to be determined by the number of user input commands and the time that has elapsed after the adaptation. A simple scenario illustrating the need for a cut-off point is the following: "*The environment detects that John is stressed and starts playing his favourite music songs. John is ok with this, but after some time he wishes to turn on the TV, so he turns the music off".* The cut-off points of five minutes or five user input commands have been arbitrarily defined, meaning that if the state of an affected system is changed after five minutes have elapsed or five user input commands have interceded, this change will not be considered as a rejection of the adaptation. Yet, a system affected by the adaptation is no longer examined if its state is changed again due to another decision of a reasoning agent. For simplicity purposes, Figure 57 illustrates the algorithm for a single system affected, while actually all the systems affected constitute entries in a table that is updated according to the Adaptation Rejection Detection algorithm. Although the initial tests that have been carried out with UXAmI Observer indicate that the defined cut-off points are reasonable, further testing with users living and interacting in actual AmI environments are required. Nevertheless, it should be noted that evaluators are fully empowered to either decline a suggestion of the system as an adaptation rejection, or manually add an adaptation rejection of it was omitted by the tool's inference mechanism. Future versions of the tool will also learn from evaluators' answers regarding adaptation rejection and will adapt the initial thresholds accordingly (e.g., if evaluators tend to indicate that a user action was an adaptation rejection beyond the five minute or five interactions threshold, then these will be expanded).

Adaptation Rejection Detection Algorithm (for each adaptation introduced)

begin

identify *currentState* of *systemAffected* (system that changes state after a reasoning agent propagates information)

while ((*timeElapsed* < 5 minutes) AND (*userinputCommands* < 5)) do

    if (newAdaptationIntroduced = TRUE) then

        identify *new_currentState* of *new_systemAffected* (system that changes state after a reasoning agent propagates information)

        if ((*new_systemAffected* = *systemAffected*) AND (*new_currentState* ≠ *currentState*)) then

            break;   // *stops inspecting for the current adaptation event*

        end if

    end if

    If (user provides *userinputCommand*) then

      identify *new_currentState* of *new_systemAffected* (system that changes state after user input)

      if ((*new_systemAffected* = *systemAffected*) AND (*new_currentState* ≠ *currentState*)) then

        adaptationRejected : TRUE;

        break;

      end if

      userinputCommand += 1;

    end if

end while

end

Figure 57. Rejection detection algorithm

Input error POIs are inferred based on the sequence of user inputs and system responses. More specifically, when UXAmI Observer detects at least two consequent user input commands without a system response, these are annotated as potential input errors, since one of them was potentially erroneous and not recognized by the AmI system that the user intends to interact with. Input commands are acquired by the tool through the information propagated by the corresponding AmI environment agents, as illustrated in Section 5.1. Although this rationale is effective for systems that support single user interaction with one system at a time, it is evident that this is not always true for multi-user

interactions with multiple systems in an AmI environment, where input commands are not given directly to a specific system, instead they are acquired through the environment sensors. For instance, one user might be waving in front of a television to interact with it and another user might be providing voice commands to the heating system, making it possible to receive two consecutive user inputs before a system response. Furthermore, although not common, it is possible that interaction with a system may require the combination of two input commands to trigger a system response. To compensate this behaviour that might lead to incorrectly suggested input error POIs, UXAmI Observer learns from the evaluator's responses regarding the correctness of the suggestion, and adapts accordingly future suggestions. For instance, in case that a system requires two consecutive input commands, once the evaluator indicates that this was not an input error the combination of these two input commands will not be suggested as a potential input error if it is followed by a response of the specific system.

Last, two more automatically acquired points refer to emotions and implicit interactions. Emotions are received by the information propagated by the corresponding emotion detection agents, if available. Implicit interactions pertain to information related to emotions and detection of user location. In the lack of the corresponding agents in the environment, emotion and implicit interaction POIs will not be automatically annotated. To facilitate evaluators in identifying whether a POI was automatically calculated or provided by a human observer, the label "Indicated by UXAmI Observer" is included in the POI details panel for all the automated measurements.

Below the experiment timeline follows the vertical interaction timeline, a scrollable panel which includes the following point annotations: task initiation, task ending, user input, implicit user interactions, system response, and adaptation insertion, ordered according to the time of their occurrence. The first time that the evaluator will view the interaction timeline for a session all task initiation events will be found at the top of the timeline, and all task ending events at its bottom (Figure 58 Left). By selecting however a system response, the evaluator can define for which task this response is the ending condition (Figure 58 Right), which will update the timeline accordingly (Figure 59). By selecting a task ending point on the vertical timeline, the evaluator can define whether this task was completed with success, partial success, or if it has failed. Based on these indications, the tool generates the task success rate score and the relevant chart (presented in section 5.2.3).
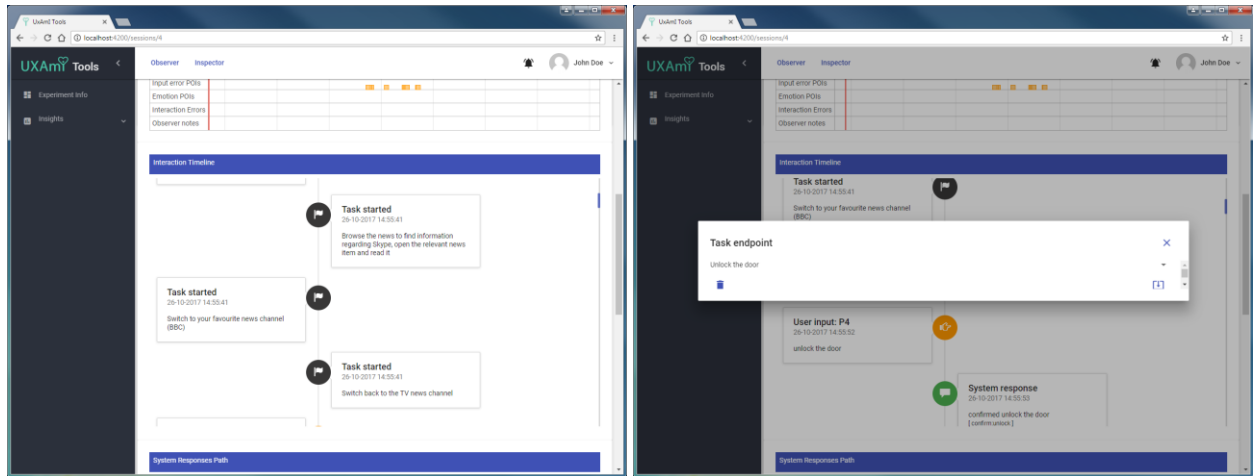
Figure 58. UXAmI Observer: (Left) Initially all task initiation events are placed at the top of the timeline; (Right) Defining a task endpoint
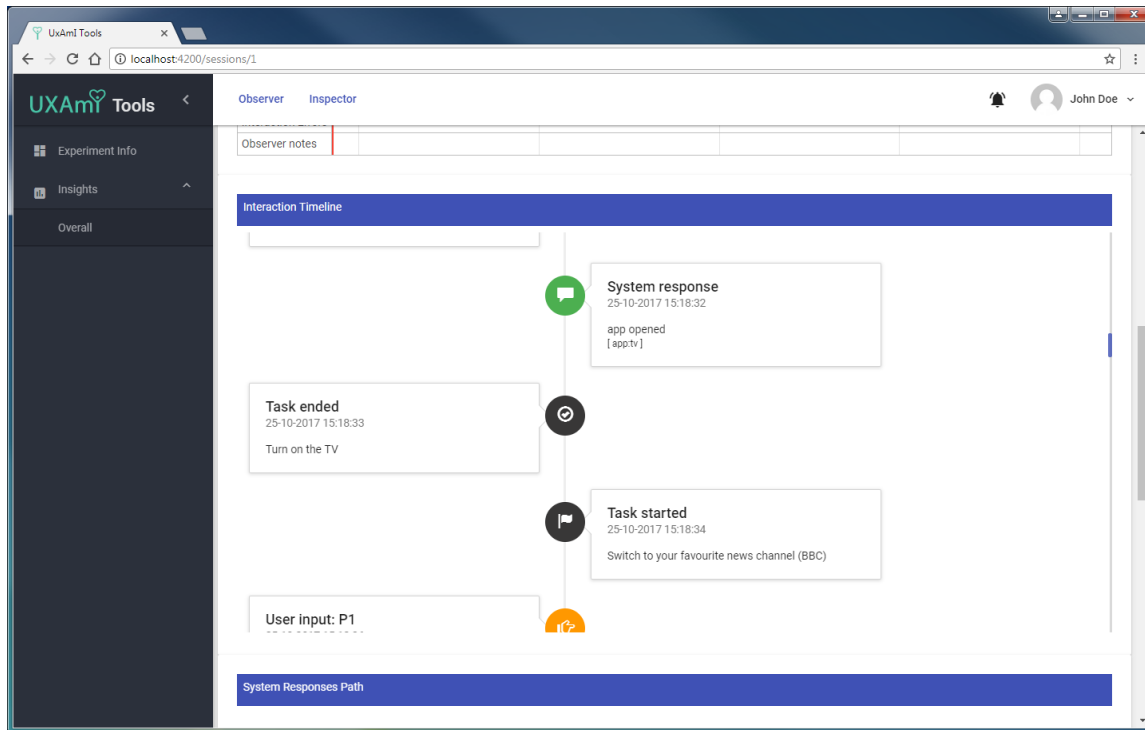
Figure 59. UXAmI Observer: Vertical timeline with appropriately ordered task ending and task initiation annotations

Each annotation on the timeline includes an area with all the relevant information for the specific event, as explained in Table 23. Since all events on the vertical timeline are timestamped, they are in complete synchronization with the horizontal timeline, therefore when the evaluator clicks on an event of the

vertical timeline, the horizontal timeline – including the videos – moves to the specific timestamp to facilitate direct association of the events and a deeper understanding of the interaction.

Table 23. UXAmI Observer: vertical timeline events information

Timeline annotation	Time	Participant	Additional information
Task started	✔		Task description
Task ended	✔		Task description
User input	✔	✔	Input type Input information as propagated by the interaction agent
Implicit interaction	✔	✔	Implicit interaction information as propagated by the corresponding agent
System response	✔		Status information as propagated by the application
Adaptation	✔		Adaptation information as propagated by the corresponding agent

As already mentioned, when at least two consecutive user inputs without an intermediate system response are detected, these are annotated as input error POIs. In the vertical timeline, an explicit question is addressed to the evaluator (Figure 60), asking them to confirm whether this is actually an input error. The same confirmation approach is applied for adaptation rejections. The horizontal timeline is updated accordingly, based on the evaluator's responses.

The next structural element of the timeline screen is the system responses path (Figure 61), aiming to shed light to the session from the perspective of the application. The evaluator can therefore have an overview of what has been done in the user's interaction with the system, without focusing on agent or user actions. Although this path is linear for each single session, this is not the norm for the entire experiment, where – as discussed in Section 5.2.3 – the evaluator can have an overview of the various system paths that have been employed by all the users.
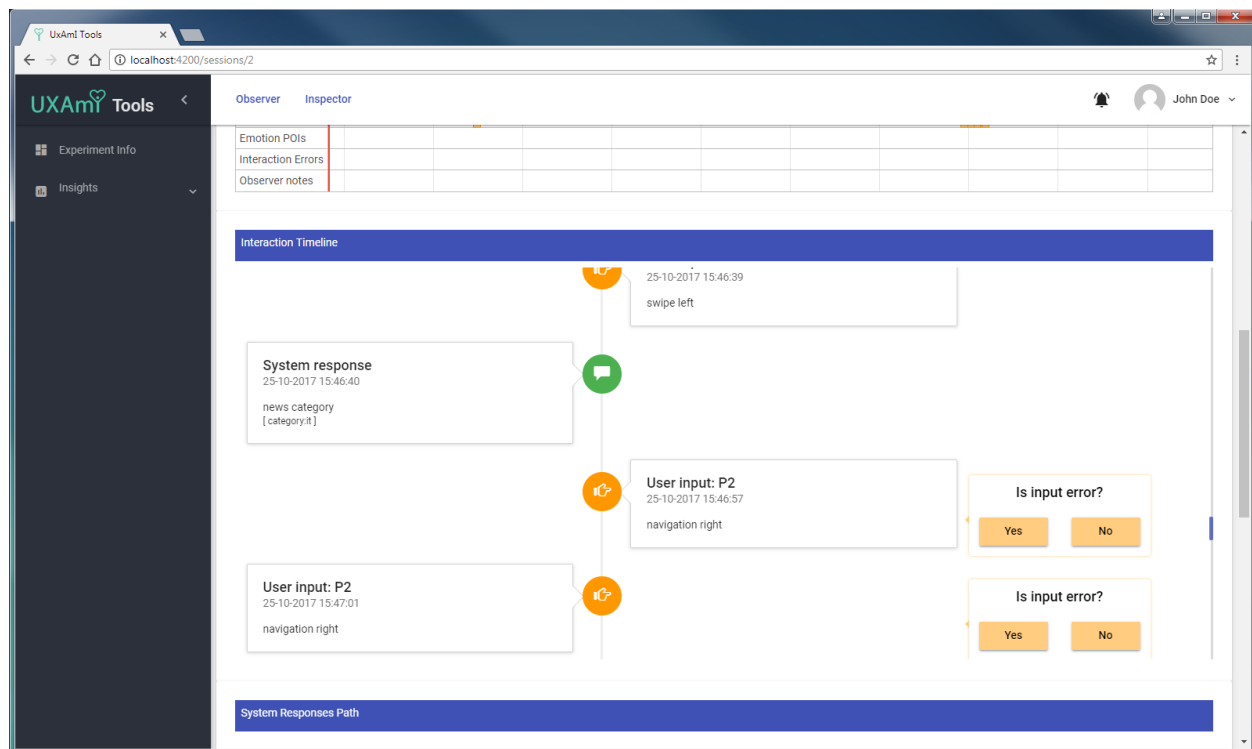
Figure 60. UXAmI Observer: Prompt to confirm input error POI
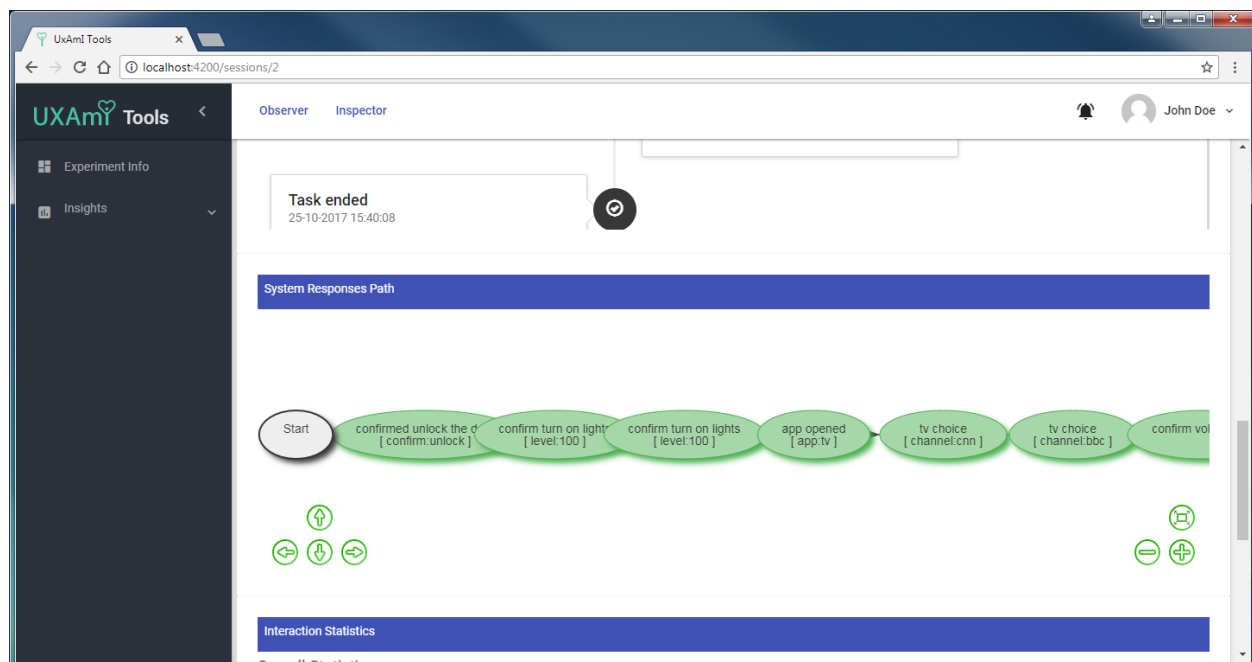
Figure 61. UXAmI Observer: System responses path

The last constituent of the timeline screen is the interaction statistics, which displays statistics information per participant besides the overall session statistics in multi-user tests (Figure 62). The statistics provided are:

- Number of user interactions.

- Number of user input errors.

- Number of system responses.

- Number of interaction errors.

- Number of adaptations introduced.

- Number of adaptations rejected.

- Number of emotion POIs.

- Usage percentage per input modality.

- Accuracy percentage per interaction modality.

- Number of errors over time, with the possibility to change the time units illustrated in the chart, so as to explore the user's error behaviour over larger periods of time (in the case of long-term experiments).
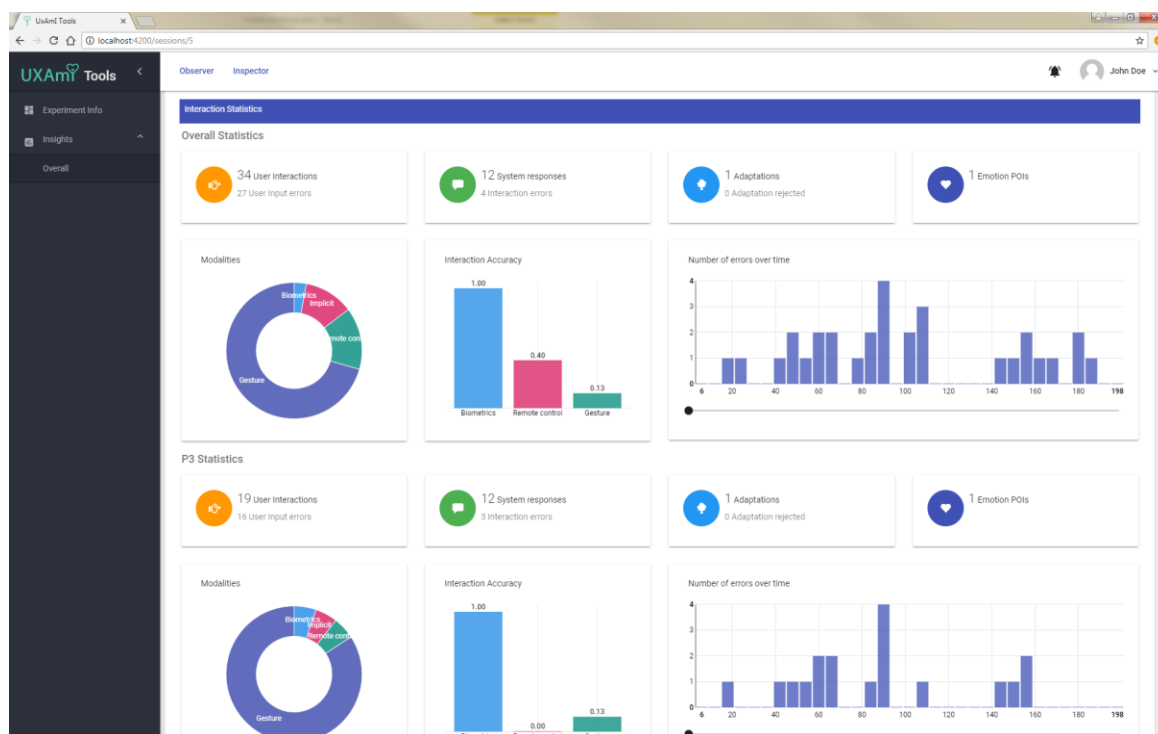
Figure 62. UXAmI Observer: Interactions statistics for a session involving more than one users

5.2.3 EXPERIMENT INSIGHTS

Insights aim to aggregate information from all the participants of an experiment, thus facilitating the evaluator towards more generalized observations and conclusions. The insights information provided is focused on five main concepts: (i) overview of experiment details, (ii) points of interest annotated on the floorplan of the relevant AmI environment, (iii) usage information, (iv) interaction statistics, and (v) system responses path.

Experiment details refer to information about the participants, clustered and presented according to the enlisted attributes (typically age, gender, and expertise). Moreover, the specific evaluation targets (artefacts/systems and applications) of the current experiment are presented. The experiment details information is available on the top right corner of the screen (Figure 63).
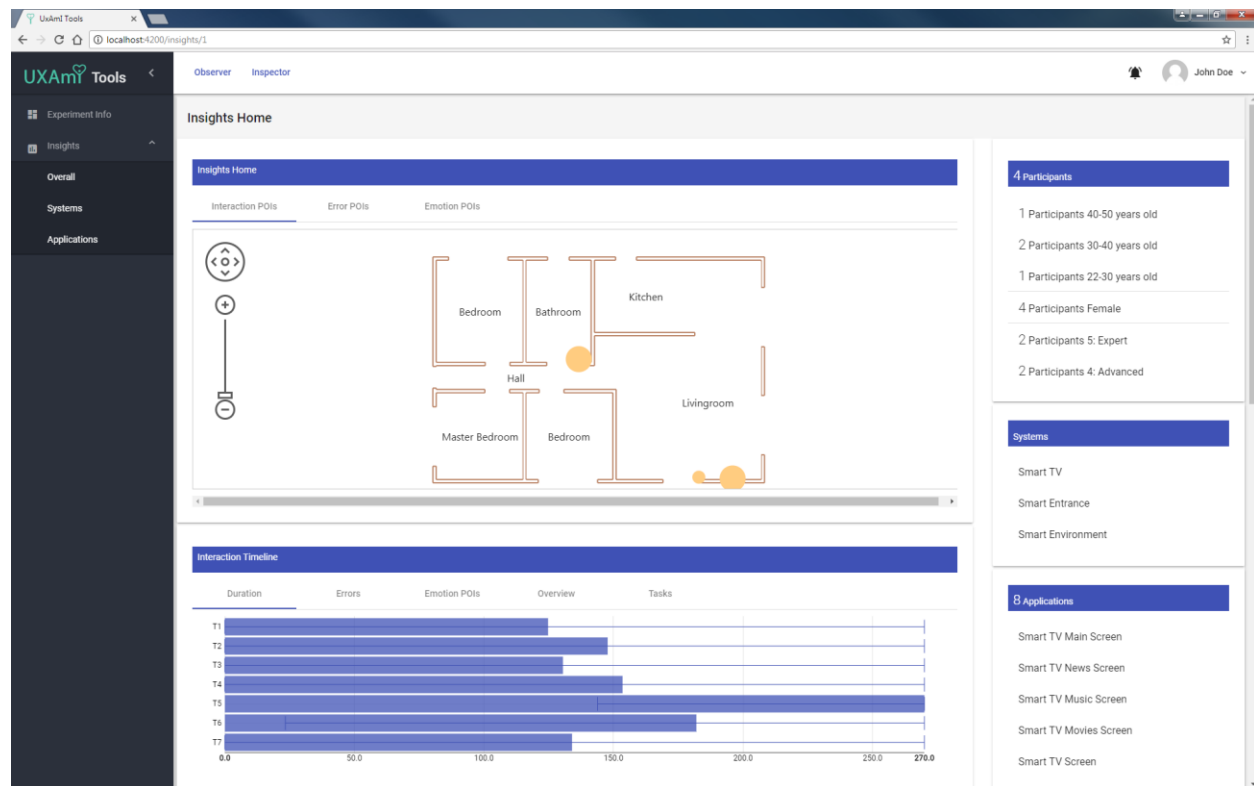
Figure 63. UXAmI Observer – Insights: experiment details and floorplan

The floorplan information involves presenting interaction POIs, error POIs, and emotion POIs on a representation of the AmI space (Figure 63). Interaction POIs occur based on the artefacts with which the user interacts, error POIs represent points where input errors mostly occur, while emotion POIs refer to points where the user has been reported by the corresponding agent to have specific psychophysiological

measurements (e.g., stress, high skin conductance, fast heart rate). The floorplan is constructed using coordinates for each room, stored in UXAmI modeller (section 5.1), a process that needs to be carried out only once for each AmI space. Each POI is then annotated as a bubble marker on the floorplan location where it occurred, with a size corresponding to its frequency (three sizes are always available: small, medium, and large). POIs therefore are determined by two parameters: their position and their size. The position can be acquired either using a user localization agent of the AmI environment, or inferred according to the location of the system with which the user interacts. The size of the POI depends on the frequency of its occurrence, and is calculated dynamically, by receiving the minimum and the maximum value of occurrences and dividing them in three quantiles. The floorplan visualization is available topmost in the insights page and aims to assist the evaluator in obtaining an overview of the user's interaction in the environment and to detect where users mostly interact with artefacts and application, where errors happen mostly and where users are more stressed.

Usage information aims to reveal how the specific systems and applications have been used during an experiment. For long-term experiments it refers to displaying interaction heat maps, by illustrating the number of usages per hour and per day of a week. For short-term task-based experiments, this information includes duration per task, errors per task, emotion POIs per task and an overview of all the above for all the tasks of the experiment. The first three are provided in the form of bar charts, where each bar represents the average value per task (e.g., average duration per task) and is accompanied by a line indicating the standard deviation (Figure 64). The overview panel features four area charts (Figure 65), each illustrating how the parameters of duration, user input errors, user emotion POIs, and adaptations applied evolve over the various tasks involved in the experiment.
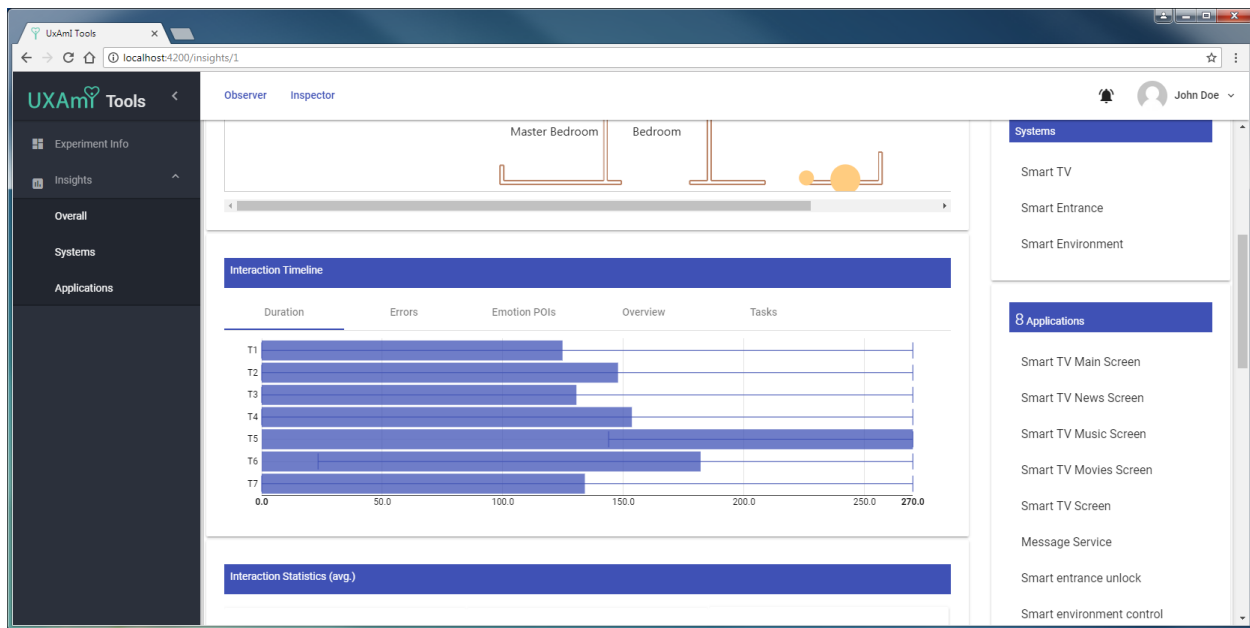
Figure 64. UXAmI Observer – Insights: Task duration
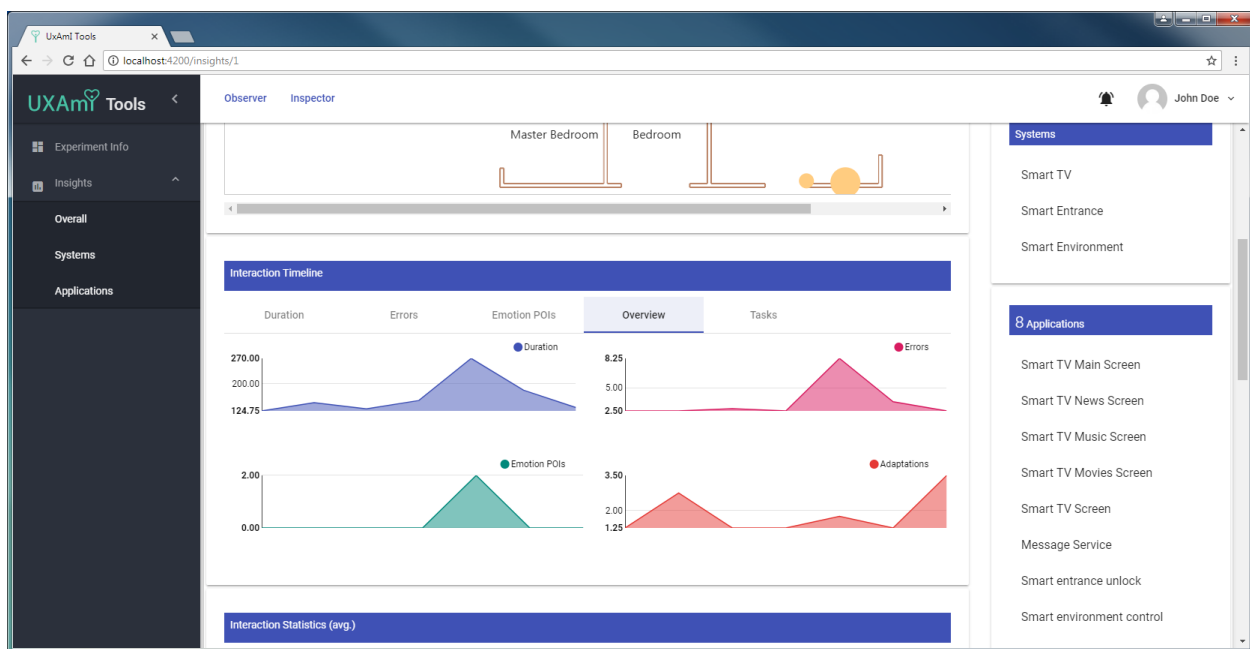
Figure 65. UXAmI Observer – Insights: overview of usage analysis for task-based experiments

The interaction statistics are similar to the ones presented in the session timeline view, with the difference that they are calculated over all the system usages by all the experiment participants and include (Figure 66): pie with input modality usage percentage, pie with percentages of implicit interactions, number of total user interactions and input errors, number of implicit user interactions along with the number of

relevant adaptations introduced and number of those adaptations that was rejected, bar chart with accuracy per input modality used, bar chart with the number of errors over time featuring customization over the time units displayed, number of system responses and number of interaction errors, total number of adaptations and number of rejected adaptations, as well as number of POIs related to users' detected emotions. A chart illustrating task success per task is also included, employing stacked bars with three colours. The chart shows the number of successful, partially successful and failed executions per task, providing also the overall task sucesss rate score, calculated according to Equation 2 (Nielsen, 2001).

$$TS = (S + (0.5 * PS)) / T \qquad (2)$$

S is the number of successful task executions, PS is the number of partially successful task executions, and T is the total number of task executions.

Finally, the system responses path illustrates all the responses from the applications involved in the experiment aggregated from all the participants (Figure 67). Therefore, the path is not linear (as in the session timeline view). The goal of this component is to facilitate the exploration of all the possible paths that have been followed to retrieve specific information (e.g., find news item through browsing the categories or through search, select a channel through a menu or incrementally through the up/down command), and to assist evaluators in clarifying if and how users employ the various offered paths during their interaction with the applications.

Figure 66. UXAmI Observer – Insights: interaction statistics

Figure 67. UXAmI Observer: Insights – system responses path
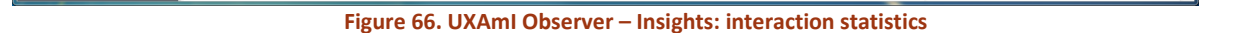
If more than one systems have been employed in an experiment, and more than one applications, the aggregated insights information is also available per system and per application, through the left menu options.

5.2.4 COMPARING EXPERIMENTS

Experiments that refer to the same evaluation targets can be compared, allowing the evaluator to verify if the user experience was improved during the various iterations. Experiments can be selected for comparison, where they are placed in chronological order side by side. In the experiment comparison screen, the evaluator can view insights and charts stemming from the analysis of all the experiment sessions, as they are presented in the insights view. As illustrated in Figure 68 and Figure 69, the experiment comparison screen includes (i) charts for task success, modalities employed, implicit interaction analysis, interaction accuracy, number of errors over time, and (ii) insights regarding the total number of system responses, interaction errors, user interactions, user input errors, adaptations applied, adaptations rejected, as well as the number of implicit user interactions accompanied by the number of relevant adaptations and adaptation rejections.
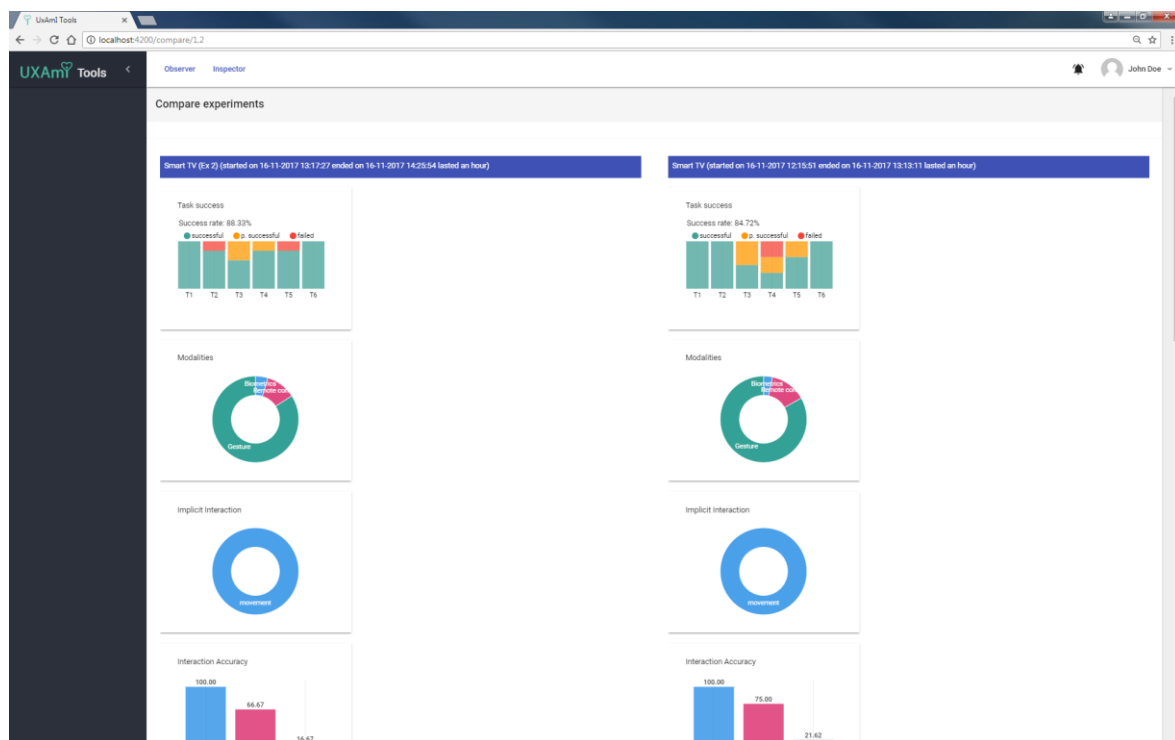
Figure 68. UXAmI Observer: Comparing experiments (part A)
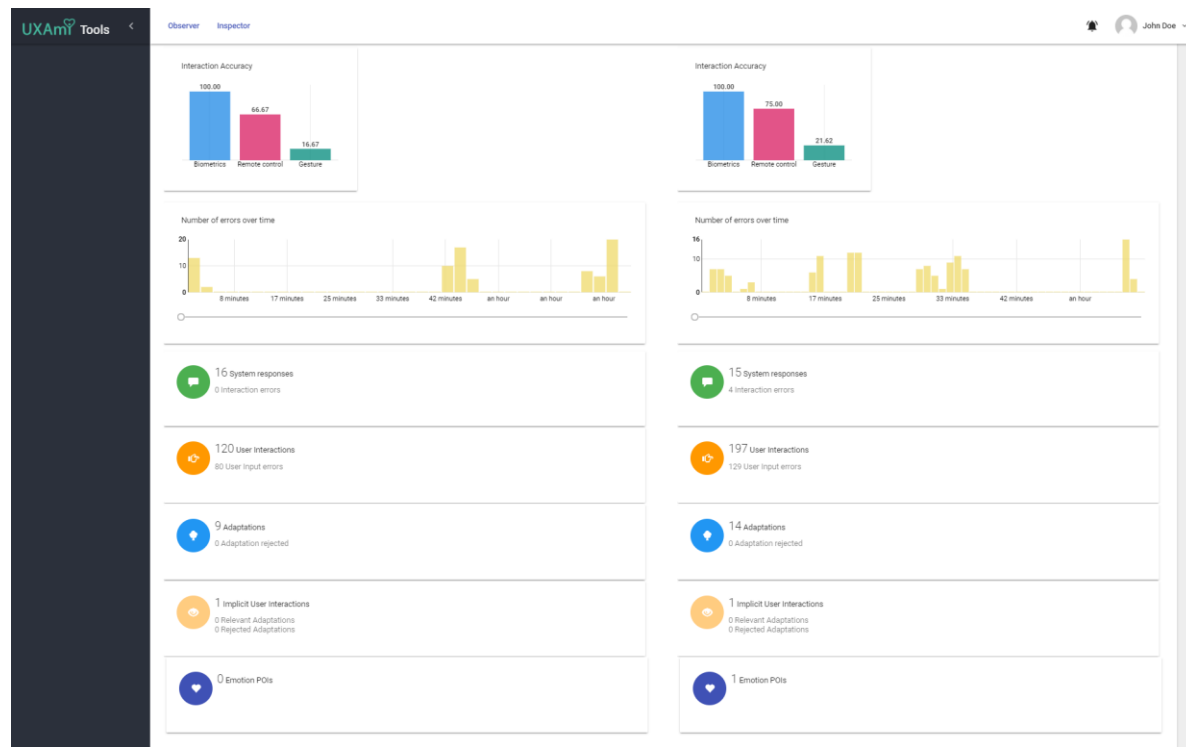
Figure 69. UXAmI Observer: Comparing experiments (part B)

5.2.5 COMPATIBILITY WITH THE UXAMI FRAMEWORK

UXAmI Observer is a tool aiming to facilitate user-based experiments and the acquisition of automated measurements and relevant insights, following the approach advocated by UXAmI Framework. In summary, the framework suggests 30 fully automated measurements, a considerable proportion of which (28 in total) has been implemented by the tool, as shown in Table 24. It should be noted that five metrics (the ones marked with ✔*) have been implemented and the relevant information is available in the various views supported by the tool, however to better facilitate evaluators' understanding, special markers need to be added to the timelines, signifying that a change in the employed platform has been detected or a collision with another user's actions. Two UxAmI metrics have not been implemented and constitute the objective of future work: system failures, which is straightforward and will be easily addressed, and undiscovered functionalities of the system. The latter needs to be further explored regarding how it can be acquired, given that the entire set of functionalities of a system cannot be known without any instrumentation, which was one of the primary concerns and drivers during the implementation of UXAmI Observer.

Table 24. UXAmI Framework constructs implemented in UXAmI Observer

UXAmI Framework Constructs		Implemented in UXAmI Observer
INTUITIVENESS		
Awareness of application capabilities	Functionalities that have been used for each system	✔
	Undiscovered functionalities of each system	Future work
Awareness of the interaction vocabulary	Percentage of input modalities used	✔
	Erroneous user inputs (inputs that have not been recognized by the system) for each supported input modality	✔
	Percentage of erroneous user inputs per input modality	✔
ADAPTABILITY AND ADAPTIVITY		
Adaptation impact	Number of erroneous user inputs (i.e., incorrect use of input commands) once an adaptation has been applied	✔
Appropriateness of recommendations	Percentage of accepted system recommendations	✔
USABILITY		
Effectiveness	Number of input errors	✔
	Number of system failures	Future work
Efficiency	Task time	✔
Learnability	Number of interaction errors over time	✔
	Number of input errors over time	✔
	Number of help requests over time	✔
Cross-platform usability	After switching device: number of interaction errors until task completion	✔*
	Help requests after switching devices	✔*
	Cross-platform task time compared to the task time when the task is carried out in a single device (per device)	✔*
Multi-user usability	Number of collisions with activities of others	✔*

	Percentage of conflicts resolved by the system	✔*
Implicit interactions	Implicit interactions carried out by the user	✔
	Number of implicit interactions carried out by the user	✔
	Percentages of implicit interactions per implicit interaction type	✔
Usage	Global interaction heat map: number of usages per hour on a daily, weekly and monthly basis for the entire AmI environment	✔
	Systems' interaction heat map: number of usages for each system in the AmI environment per hour on a daily, weekly and monthly basis	✔
	Applications' interaction heat map: number of usages for each application in the AmI environment per hour on a daily, weekly and monthly basis	✔
	Time duration of users' interaction with the entire AmI environment	✔
	Time duration of users' interaction with each system of the AmI environment	✔
	Time duration of users' interaction with each application of the AmI environment	✔
	Analysis (percentage) of applications' used per system (for systems with more than one applications)	✔
	Percentage of systems to which a pervasive application has been deployed, per application	✔
APPEAL AND EMOTIONS		
Actionable emotions	Detection of users' emotional strain through physiological measures, such as heart rate, skin resistance, blood volume pressure, gradient of the skin resistance and speed of the aggregated changes in the all variables' incoming data.	✔

The framework also proposes 25 metrics that can be calculated with automation support (i.e., metrics that need manual input by the evaluator but can be assisted with automatic calculations. UXAmI Observer has fully implemented six of them and provides visualizations for eight more (e.g., implicit interactions, task deviations, completed tasks, etc.) without however having implemented markers for each one of

them to allow evaluators provide specific input (e.g., if the system response to an implicit interaction was appropriate), which will be addressed in future development iterations of the tool.

5.2.6 EVALUATION OF THE UXAMI OBSERVER

UXAmI Observer has been evaluated by three UX experts following the heuristic evaluation approach (Nielsen, 1994a). More specifically, each evaluator inspected the interface alone against the heuristic evaluation guidelines (Table 25).

<div align="center">Table 25. Heuristic evaluation guidelines</div>

1.	Visibility of system status	The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
2.	Match between system and the real world	The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order
3.	User control and freedom	Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
4.	Consistency and standards	Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
5.	Error prevention	Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
6.	Recognition rather than recall	Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
7.	Flexibility and efficiency of use	Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
8.	Aesthetic and minimalist design	Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

9. Help users recognize, diagnose, and recover from errors	Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution
10. Help and documentation	Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Each problem that was identified was correlated with one or more guidelines. The problems reported by each individual evaluator were aggregated into a single report, removing the duplicates and merging identical problems. Subsequently, each evaluator provided their severity rating for each one of the problems in the unified list, while a final severity rating for each problem was calculated as the means of the individual evaluators' ratings.

Ratings are given by the evaluators, according to their judgement on the frequency, the impact, and the persistence of the problem, and may range from 0 to 4, as follows:

- 0 = I don't agree that this is a usability problem at all
- 1 = Cosmetic problem only: need not be fixed unless extra time is available on project
- 2 = Minor usability problem: fixing this should be given low priority
- 3 = Major usability problem: important to fix, so should be given high priority
- 4 = Usability catastrophe: imperative to fix this before product can be released

The final evaluation report of UXAmI Observer included 33 issues, 21 of which were minor or cosmetic (rating <=2), such as alignments, label names, or font sizes, which are straightforward to be addressed. The remaining 12 problems are reported in Table 26, in order of severity.

Table 26. UXAmI Observer heuristic evaluation results

Problem	Guidelines	Severity
Typically, in an iterative process an evaluator would start with heuristic evaluation and then proceed with user-based evaluation. The system could provide interoperability with the Inspector tool, so that evaluators have an overview of all the evaluation efforts for a given evaluation target.	[2]	3.33

In a user-based experiment involving many users, evaluators explore the impact of specific user attributes over the acquired the results. Since the evaluator provides user categories, it would be useful to have results presented per user category attribute range (e.g., results for older vs. younger users).	[2], [7]	3.00
Statistical analysis functionality could be included, or at least the possibility to export data in a spreadsheet format.	[2]	3.00
In the system responses area, it is not visible how often a path was followed.	[1]	3.00
User testing also involves questionnaires – it would be good to have some embedded, as well as tools to analyse them.	[2]	2.67
Charts could constitute illustrations in the usability evaluation report that the evaluator will prepare, so it would be useful to be able to export them to images.	[2], [7]	2.67
Regarding the evaluation report, the tool could provide a template pre-filled with data to facilitate its preparation. This template could be structured along the AmI attributes suggested by the UXAmI framework.	[2]	2.67
The tool could facilitate selection of variables from the framework, if one does not want to explore their full set.	[2], [3]	2.33
It is not easy to view many events in the vertical area in the interaction timeline of a session – it could be larger.	[1], [7]	2.33
The task success metric is usually calculated in user-based evaluations, a metric which is missing in the tool.	[2]	2.33
The chart illustrating the number of errors over time does not include measurement units, making it difficult to understand.	[1], [7]	2.33
The system responses path in the insights view is very small for the evaluator to have a good overview of the paths (in a zoomed-out mode but still readable) – it could be larger in height.	[1]	2.33

In summary, the most important problems revolved around common practices in usability testing and suggestions of functionality to be added, such as evaluation report templates, statistical analysis, or questionnaires support. All the issues that have been brought to surface by the heuristic evaluation – minor and major ones – will be addressed in future UXAmI Observer versions.

5.2.7 CONCLUSIONS AND FUTURE DEVELOPMENTS

This section has presented UXAmI Observer, a tool which implements the majority of automated measurements suggested by UXAmI Framework. The tool addresses evaluators conducting user-based experiments and supports them through visualizations of the data acquired for each experiment participant, as well as with aggregated statistical and usage information from all the participants. In summary, the tool supports:

- Task-based and free usage user-based experiments.
- Detailed analysis of a participant's interaction and aggregated insights from all the participants.
- Automatically acquired measurements regarding adaptations applied and adaptation rejection, user input errors, detected user emotions and implicit interactions.
- Manual annotation of POIs by the evaluator and verification of the automated measurements.
- Synchronized view of all the POIs (manual or automatic) with the session videos.
- Powerful visualizations of the interaction timeline, interaction statistics, system responses path, system usage statistics, and floorplan-based POIs.
- Comparisons among experiments.

An important contribution of the tool is that it does not require any instrumentation, in contrast to previous efforts in the field of usability evaluation automation. More specifically, neither the developers of the application, nor the evaluator is required to invest effort towards instrumenting the software to collect usage data. Other innovative features of the tool include:

- The discrimination between input errors and interaction errors, with input errors being automatically detected.
- Detection of adaptations inserted and rejected.
- Implicit interactions recognition.
- Association of POIs with the space floorplan.

Future work will address the issues that have been highlighted from the heuristic evaluation, and in particular: provide insights per user category, support questionnaires, facilitate exporting data and images, as well as to support evaluation report templates with pre-filled data. Furthermore, the tool will be redesigned to provide interoperability with the UXAmI Inspector tool, while it will be possible in future versions to select which metrics of the framework will be used in each experiment. Finally, although some tests have already been carried out, further tests with more users involved in long-term experiments are required to test the validity of the tool's inference mechanisms and to evaluate its usability.

5.3   UXAMI INSPECTOR

This section presents UXAmI Inspector, a tool which accompanies the UXAmI framework and aims to facilitate expert-based reviews. More specifically, the tool falls in the category of Tools for Working with Guidelines and is mainly addressed to evaluators who organize expert-based reviews, as well as the experts themselves who carry out the reviews. The key challenge that it aims to address is that of retrieving the appropriate sets of guidelines for the system(s) that will be evaluated, a concern that stems from the volume of guidelines which was already large in the past millennium (Vanderdonckt, 1999) and has become even larger nowadays, given the rapid evolution of the HCI domain (Shneiderman, 2017). In addition, the complexity of retrieving the appropriate guidelines increases in AmI environments, as the interaction paradigm has evolved from a single user in front of a desktop to multiple users interacting with numerous devices employing various novel interaction techniques. As a result, new guidelines have been developed to address evaluation needs in new contexts, and come into play in evaluation inspections in AmI environments.

In sum, the tool facilitates the evaluator in selecting evaluation targets (systems and/or applications) by displaying the list of systems and applications registered in the AmI environment. Once the evaluation targets have been selected (Section 5.3.1), UXAmI Inspector retrieves all the sets of guidelines that match the evaluation targets, following a tag-based approach (Section 5.3.2). The evaluator can select the guidelines against which they wish to evaluate the system(s) and application(s), and create a new inspection project. Experts can then proceed with reporting specific problems of the evaluation targets (Section 5.3.1), by describing each problem, binding it with specific guidelines violated, and assigning to it a severity rating. Figure 70 illustrates the HTA diagram (Stanton, 2006) explaining the tasks a user can carry out with the tool.

UXAmI Inspector was developed following an incremental prototyping process, starting with low fidelity paper-based prototypes (Appendix F). Screenshots from the final fully functional prototype are provided in section 5.3.1, which describes the processes of creating a new inspection project and carrying out the evaluation. The tool has been evaluated following the heuristic evaluation approach, which is described in Section 5.3.3. Finally, Section 5.3.4 summarizes the key points regarding the tool and its benefits, discusses how the tool interoperates with the UXAmI Online Community, and concludes by providing directions for future work regarding UXAmI Inspector.
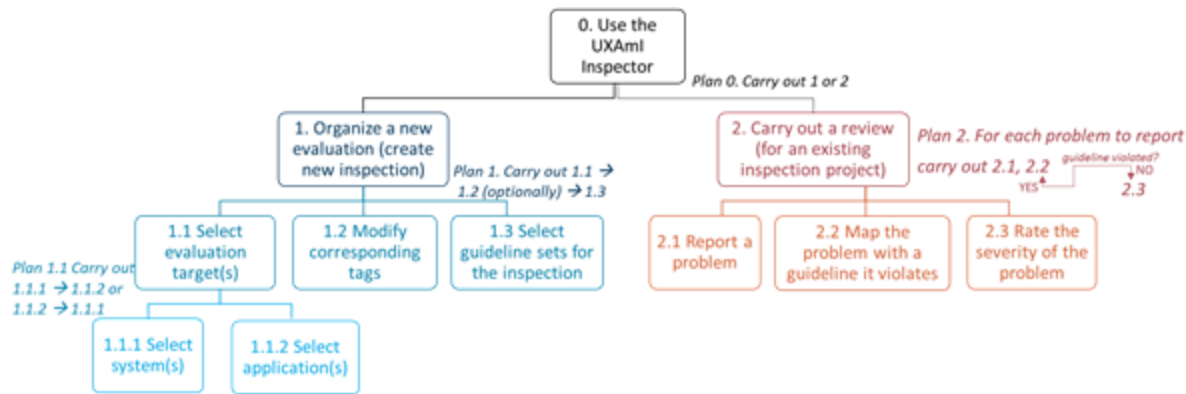
Figure 70. UXAmI Inspector Hierarchical Task Analysis

5.3.1 ORGANIZING A NEW INSPECTION AND CARRYING OUT A REVIEW

The first step towards carrying out an expert-based review is to define the evaluation target, a process in which evaluators are assisted by the tool. More specifically, the evaluator can set-up a new inspection by selecting the "Start new inspection" button in the UXAmI Inspector main screen (Figure 71). The process of setting up an inspection involves selecting the evaluation targets, reviewing the tags with which they are described, selecting which guidelines from the suggested ones will eventually be used in the inspection, and saving the aforementioned selections. This process is facilitated through a step-by-step wizard, featuring three concrete steps: (i) select main evaluation target, (ii) refine selection by indicating specific subsystems or applications, (iii) review evaluation targets, select guidelines and save selections to a new inspection project.
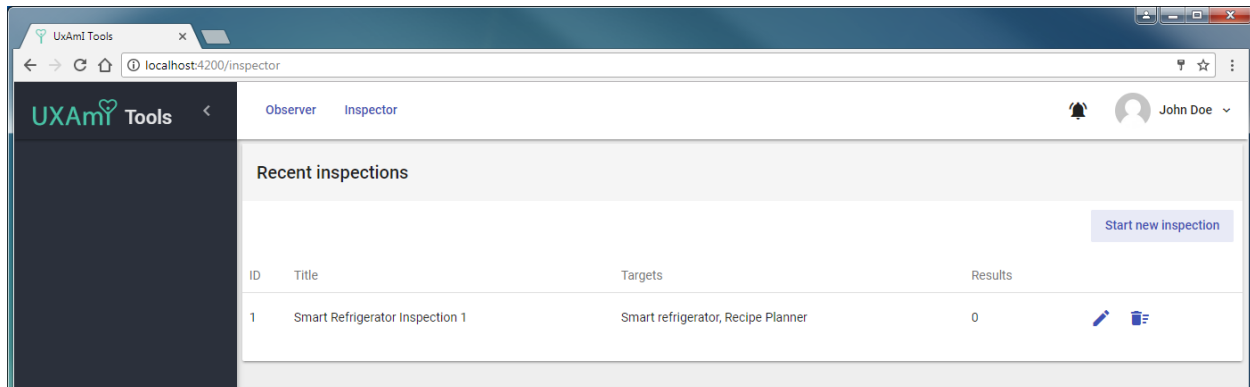
Figure 71. UXAmI Inspector main screen

As a first step, a main evaluation target has to be selected (Figure 72). The evaluation target may be defined either following a space-oriented approach or a pervasive approach, based on the fact that a pervasive application may run in several systems located in various spaces. As a result, two lists are

available to the user: one with the available spaces and one with the pervasive applications. In the space-oriented approach, the evaluation main target may be an entire AmI environment (e.g., AmI home) with its systems and applications, or a specific space in the AmI environment (e.g., Living room in the AmI home). Therefore, the "Available Spaces" list includes all the available AmI spaces and sub-spaces, while for each one an indication of the number of the artefacts and applications therein is provided. Likewise, the list of pervasive applications lists all the applications that exist in the AmI environment that are deployed in more than one devices, while for each application the number of artefacts and relevant spaces is indicated. Further refinement of the evaluation targets, and in particular the exact artefacts and exact applications of these artefacts that will be inspected, is achieved in the second step of the wizard.
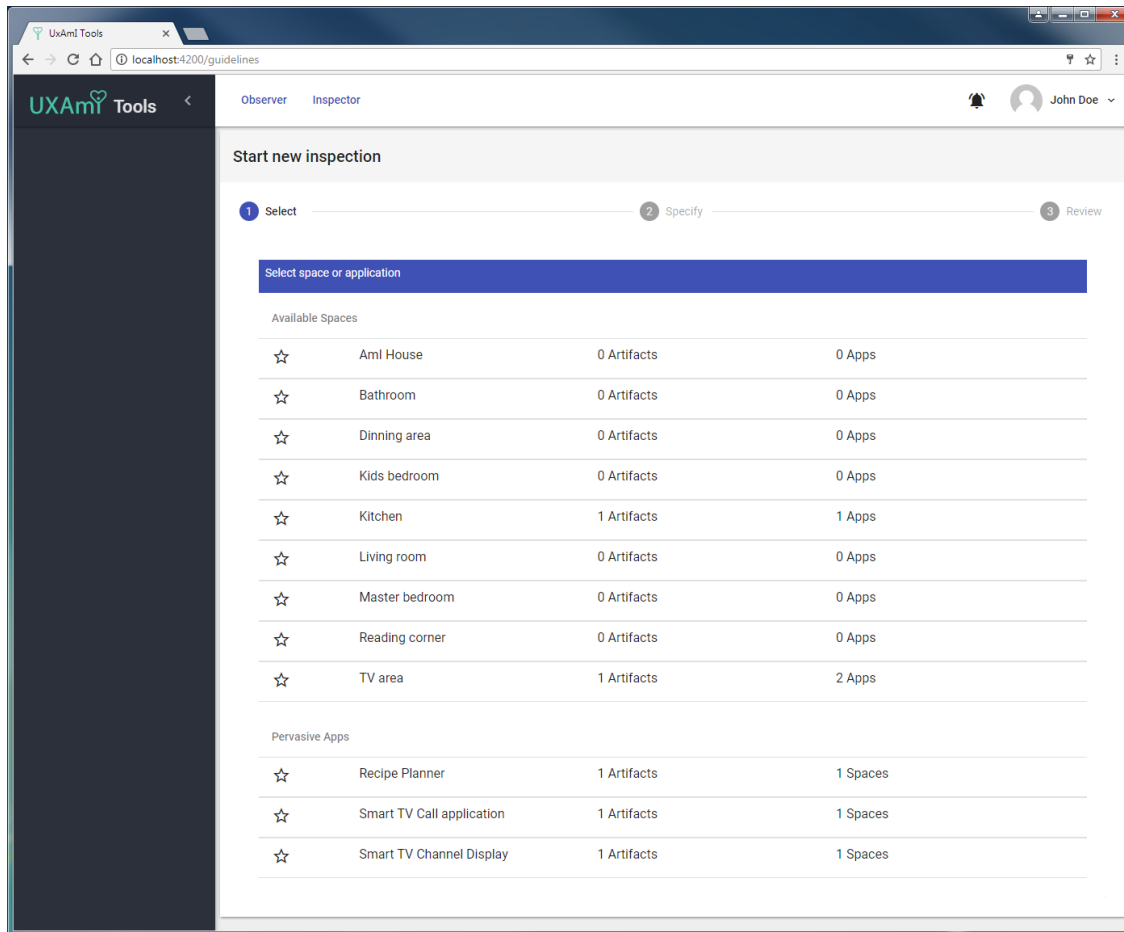
Figure 72. UXAmI Inspector: Create new inspection (step 1)

Having selected the primary evaluation target (space or pervasive application), the user is transferred to the next step to further refine their selection. If a space-oriented selection was made in the first step, then all the artefacts of the selected space are presented in the second step, with an indication of how many applications are deployed in each artefact. The user can either select an entire artefact as evaluation

target, and therefore all its applications, or expand the artefact to view and select specific applications (Figure 73). If a pervasive application was selected in the first step, then the user is presented with the list of artefacts on which the selected application runs, in order to specify the exact evaluation targets (Figure 74). Eventually, the user will end up with a specification of the exact applications that will be evaluated, having identified the exact systems in which these applications run.
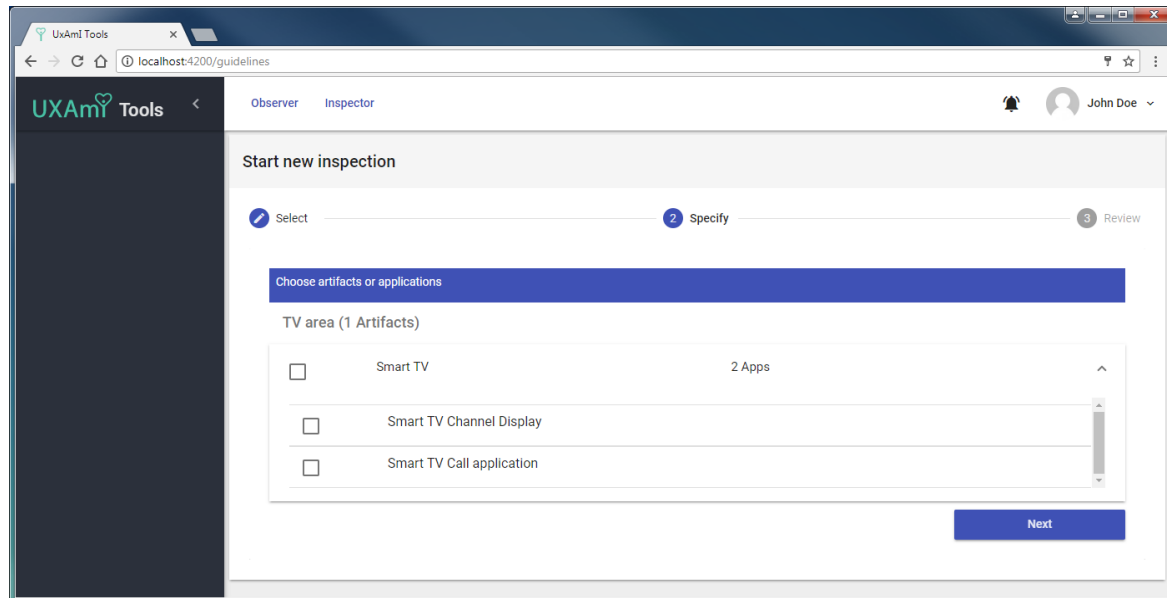
Figure 73. UXAmI Inspector: Create new inspection (step 2a) following the space-oriented approach
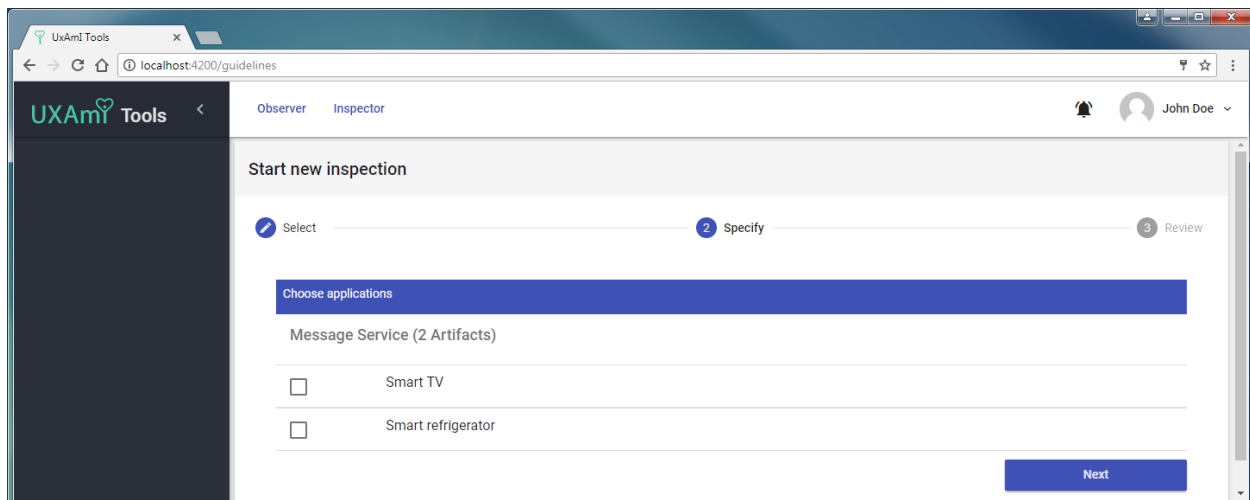
Figure 74. UXAmI Inspector: Create new inspection (step 2b) following the pervasive application approach

In the third step, the user is presented with their selections (systems and applications), as well as the tags that have been associated with each one of them (Figure 75). During this step, the evaluator can remove any tags that are considered out of scope for the inspection being planned or add new tags. Below the list of systems/applications and tags, the guideline sets that match the specific evaluation targets (Section 5.3.2) are displayed. In the case that conflicting guidelines are included in the suggested guideline sets, these are annotated with an exclamation mark icon, which displays information about the confict when selected (Figure 77). The user can select a guideline set to view the included guidelines (Figure 76) and eventually select the guideline sets that will be used in the current inspection. The creation process is considered complete once a name has been given to the inspection project and at least a set of guidelines has been selected. It should be mentioned that at any step of the process, it is possible to go back to the previous steps and change one's selections.

Figure 75. UXAmI Inspector: Create new inspection (step 3)
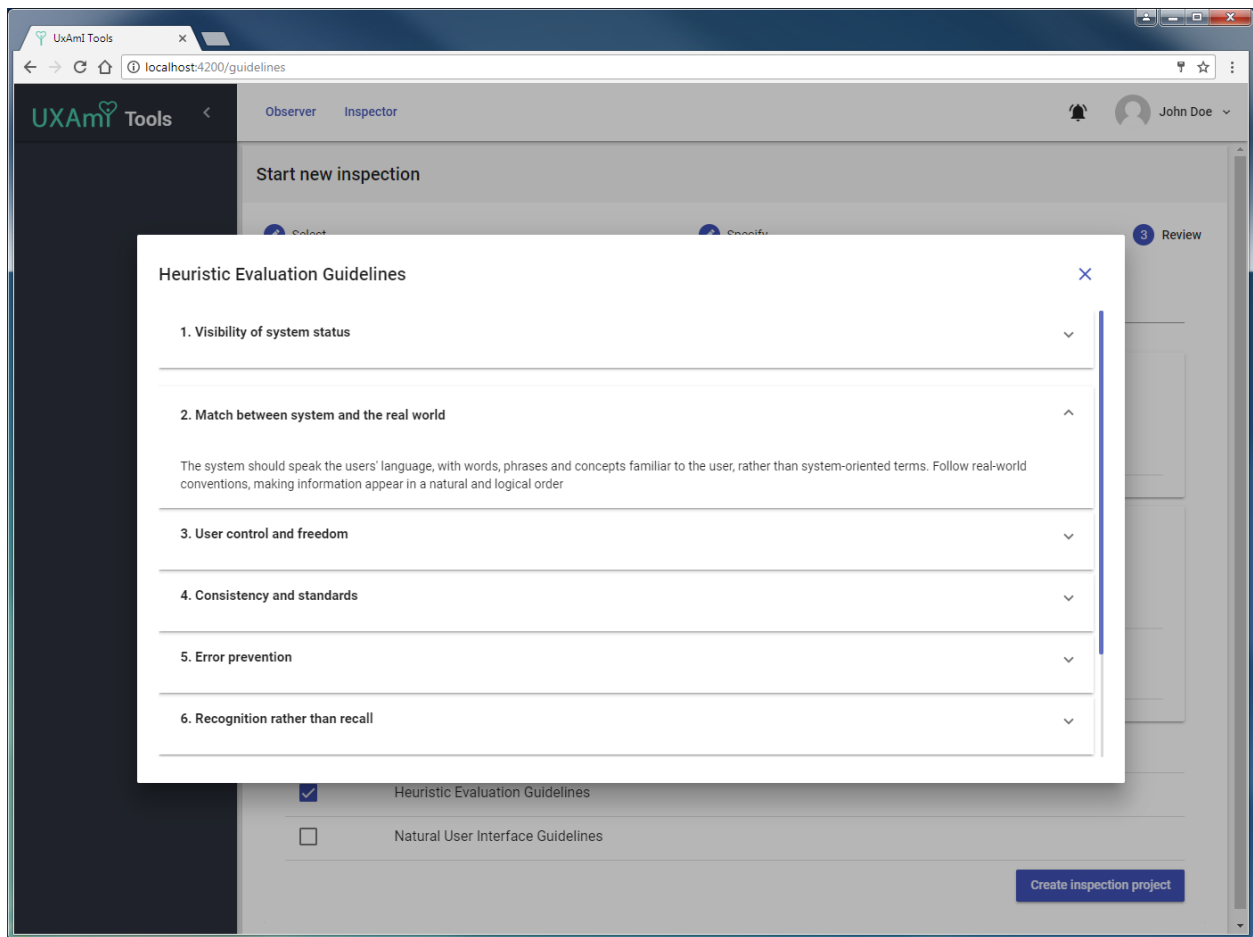
Figure 76. UXAmI Inspector: Create new inspection (step 3) – Reviewing guidelines
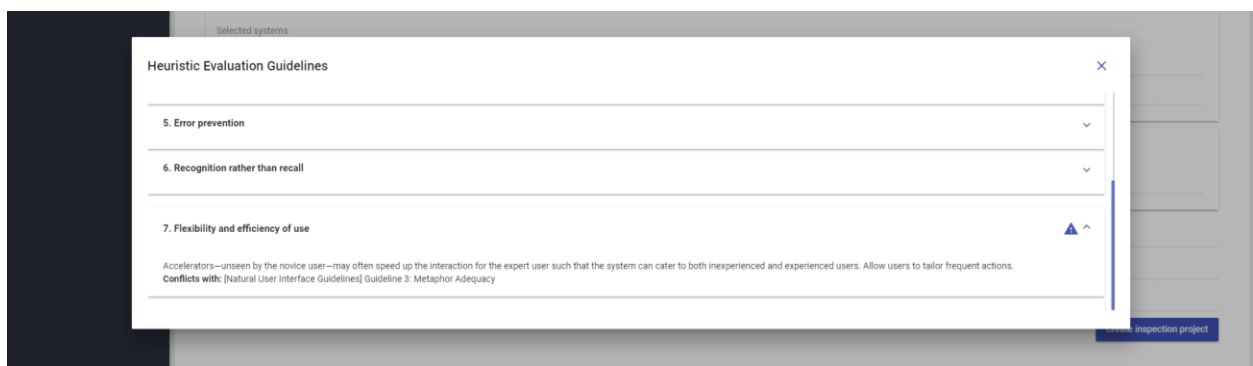
Figure 77. UXAmI Inspector: Information on conflicting guidelines

Once the third step is completed, a new inspection project is added to the relevant list (Figure 78), featuring its title, the list of evaluation targets, as well as the number of problems that have been identified (results column). The user can delete an inspection project, or edit it to review the project details. Then, the evaluator can proceed with the inspection process and report problems directly from the project inspection details page (Figure 79).
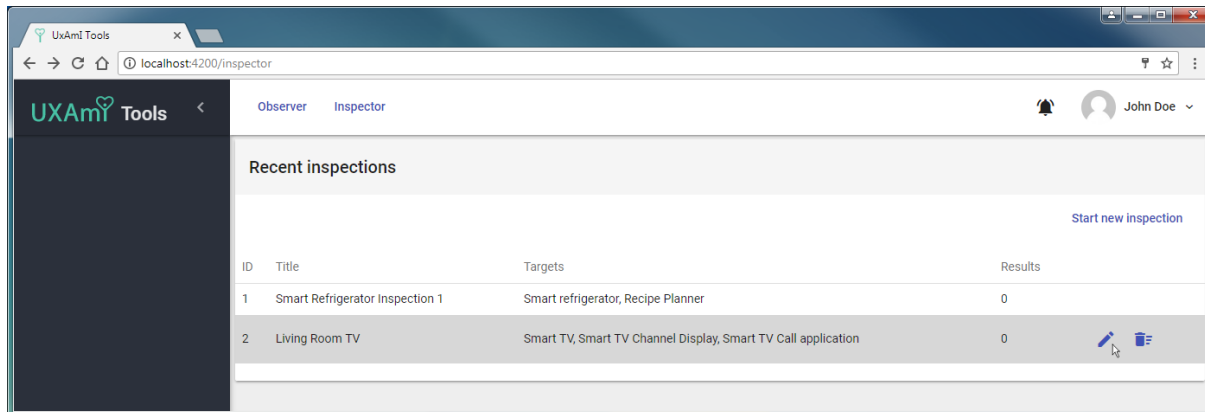
Figure 78. UXAmI Inspector: List of inspection projects
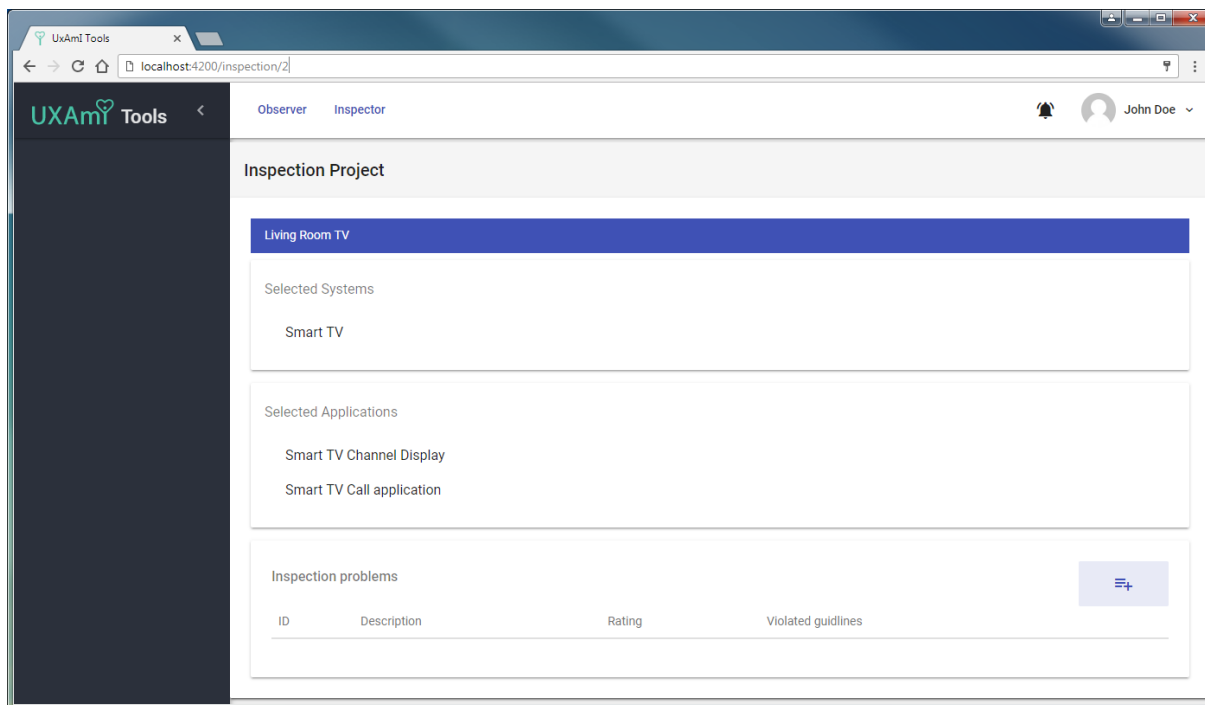
Figure 79. UXAmI Inspector: Inspection project details

Upon selecting the ("+") icon in the inspection problems area, a new screen is displayed for reporting a problem, requiring the following input from the evaluator: short problem description, severity rating, guidelines that are violated (Figure 80). The sets from which the evaluator can select the violated guidelines are the ones that the evaluator has selected as appropriate for the current project among those that UXAmI Inspector suggested according to the evaluation targets. The user can select to expand a guideline set to view its guidelines (and eventually select one), while guidelines can also be expanded to view their description. The evaluator can indicate that a guideline is violated by clicking the blue exclamation mark icon, next to its title. As soon as a guideline is selected, it is added to the summary of guidelines that the problem violates (Figure 80). The user can remove any guidelines from the summary by simply clicking the "x" icon next to the guideline title.

Figure 80. UXAmI Inspector: Reporting an inspection problem – indication of violated guidelines

Once an inspection problem is saved, it is added to the list of problems reported in the context of the inspection project. Each inspection problem in the list features its description, rating, and number of violated guidelines (Figure 81).
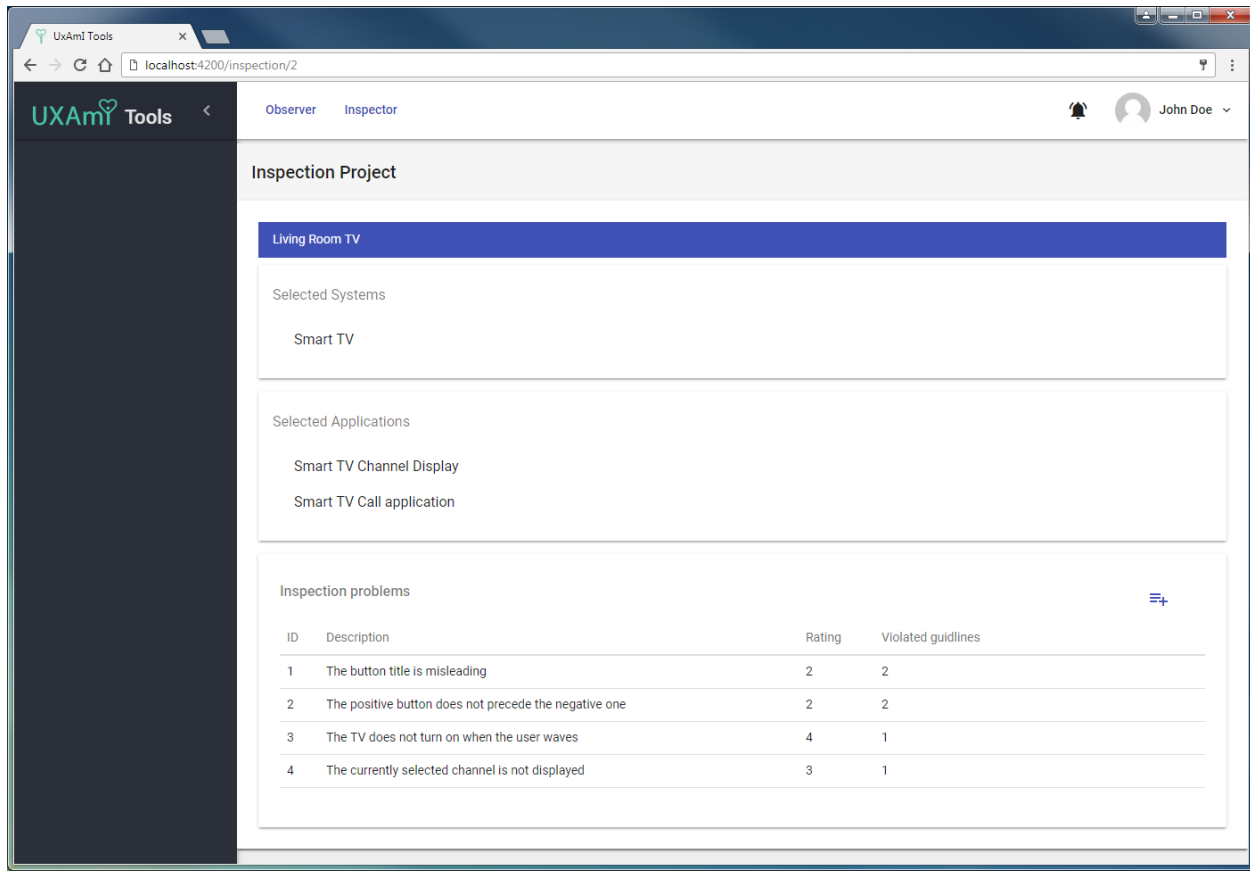
Figure 81. UXAmI Inspector: Inspection project details – with inspection problems reported

5.3.2 MATCHING GUIDELINES TO EVALUATION TARGETS

An important consideration in the context of the UXAmI Inspector was to facilitate the as-automatic-as possible retrieval of guidelines relevant for the specified evaluation targets. To this end, two challenges were raised: (i) how to acquire guidelines and (ii) how to classify guidelines so as to appropriately match them with the evaluation targets, challenges which were met through crowdsourcing and tag-based classification.

Crowdsourcing is an online, distributed problem-solving and production model that has emerged in recent years, harvesting intellect distributed in the crowd, deriving its wisdom not from averaging solutions but from aggregating them (Brabham, 2008). The term, which was first coined in 2006 by Jeff Howe and Mark Robinson in the June issue of *Wired* magazine (Howe, 2006), has already seen wide application in practice and is yet to receive intense attention from the scholars (Zhao & Zhu, 2014). Although crowdsourcing was initially used in a business context, this is no longer the norm, as it has already been extensively employed

in scientific and engineering fields (Zhao & Zhu, 2014). Towards implementing this crowdsourcing approach, UXAmI Inspector interoperates with the UXAmI Online Community (Section 5.4), where members are expected and motivated to contribute guidelines that can be used in the context of expert-based reviews.

Despite their power, crowdsourcing systems face three key challenges: how to recruit and retain contributors, how to combine their contributions, as well as how to evaluate users and contributions (Doan, Ramakrishnan, Halevy, 2011). The challenge of recruiting and retaining users has been addressed in the context of the UXAmI Online Community through exploring the motivators of user participation and knowledge sharing in OCs and studying state of the art practices in rewarding user contributions (Section 5.4.1), and by designing an adaptive reward scheme to promote content contribution and active user participation in the UXAmI OC (Section 5.4.2). With regard to combining user contributions, a technologically assisted moderation approach has been foreseen (Section 5.4.3). On the one hand, users are prompted to relevant contributions (e.g., tags or guidelines) when uploading content, allowing them to check for themselves and eliminate duplicate materials. On the other hand, potentially relevant guideline sets are identified and clustered for moderators to review and approve the material that will be eventually made available to the OC members. Last, evaluation of users' contributions is achieved through moderation (Section 5.4.3), as well as through helpfulness votes that promote content of high quality, and through violation reports that indicate if any content or user violate the terms and conditions of the OC (Sections 5.4.3and 5.4.4). As all activities in the community are directly related with the reward scheme that has been designed, in general positive contributions are endorsed, while inappropriate content or activity in the OC is discouraged.

As illustrated in Section 5.1, all the UXAmI framework tools share a common database, therefore when a guideline set is approved by the UXAmI OC moderators it becomes instantly available in UXAmI Inspector. Each guideline set stored in the UXAmI adopts the structure illustrated in Figure 82, and in particular it includes one or more guidelines (each featuring a title, a short description, and pointers to conflicting guidelines), one or more references, as well as one or more tags. Systems, on the other hand are also described by tags, therefore UXAmI Inspector retrieves as relevant for the current inspection all the guidelines that have a tag matching one of the target systems' or application' tags (see Figure 83 for an example).
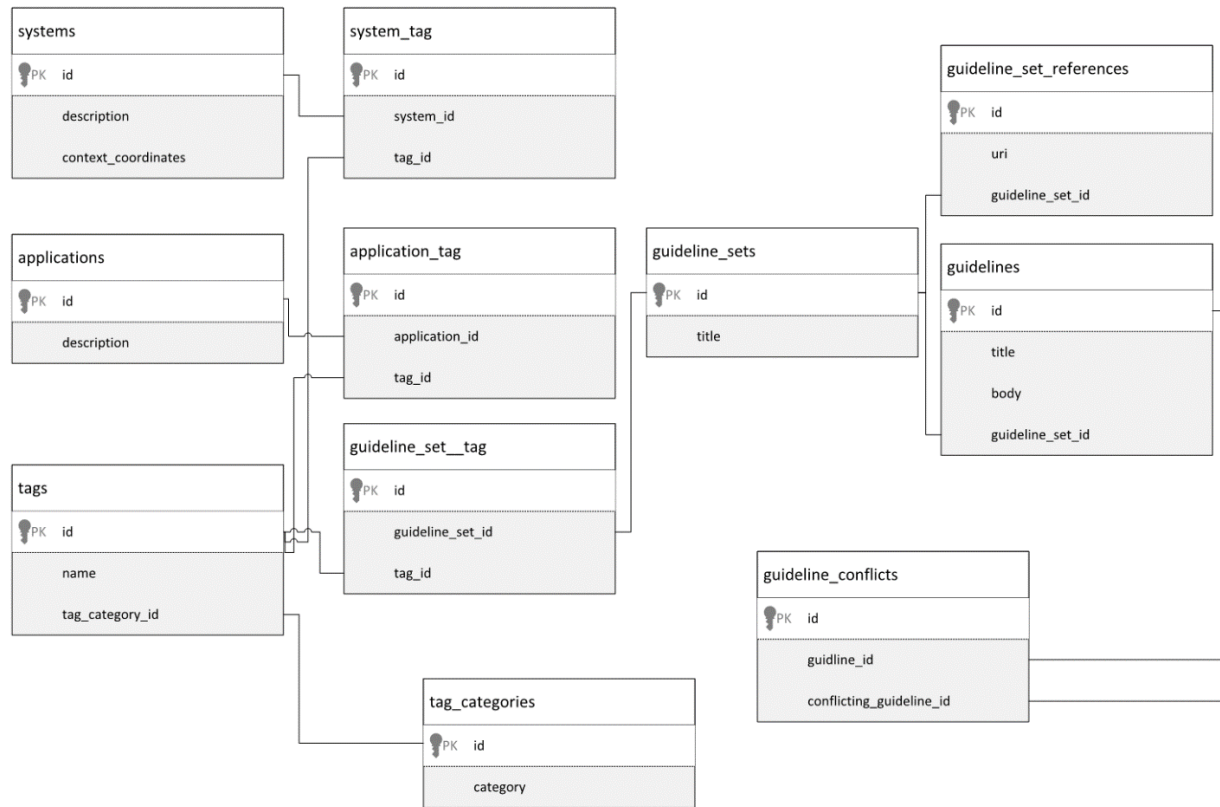
Figure 82. UXAmI DB schema representation for guideline sets
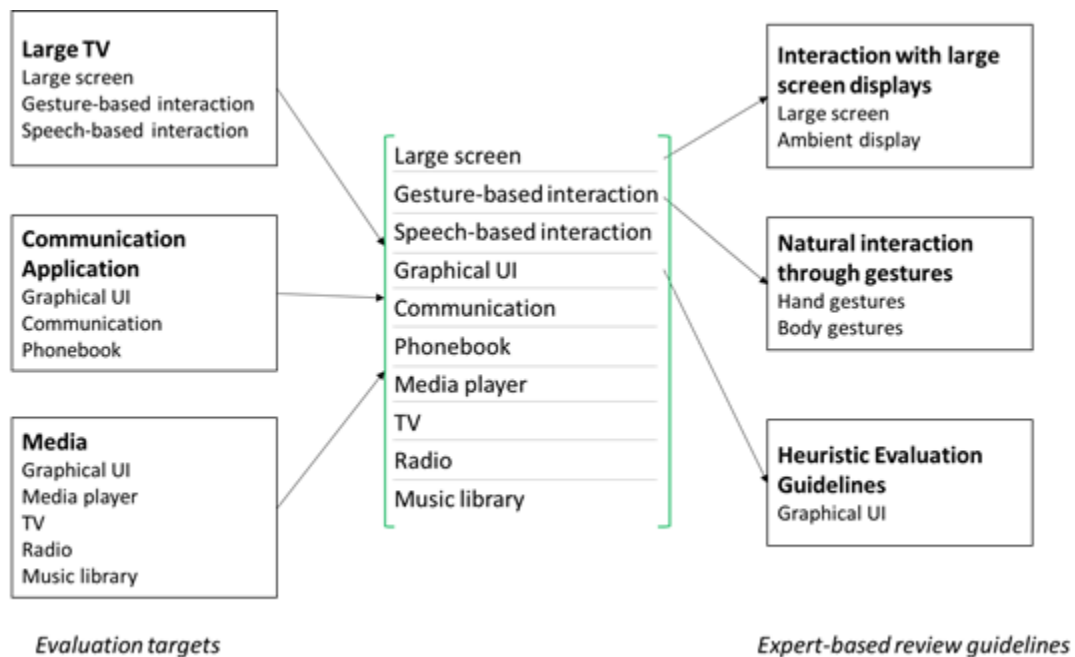
Figure 83. UXAmI Inspector tag matching example

Folksonomies and social tagging have become quite popular in the web, inspired by the need to associate metadata with web objects in a flexible manner, avoiding problems imposed when fixed taxonomies are employed for metadata description (Gupta, Li, Yin, & Han, 2010). Folksonomies, unlike formal taxonomies, do not have explicitly defined relationships between terms, organized under a hierarchy; instead all terms belong to a flat namespace (Gupta, Li, Yin, & Han, 2010). Several benefits have been reported for folksonomies, such as that they are inclusive (reflecting the vocabulary of the users) and current (by rapidly reflecting changes in terminology and world events), they offer discovery, they are democratic (support users' way of thinking) and self-moderating, they offer a low cost alternative and foster usability (Kroski, 2005). Despite their many benefits, folksonomies suffer from specific weaknesses related to the problems inherent to all uncontrolled vocabularies such as ambiguity, polysemy, synonymy, and basic level variation (Spiteri, 2007). Approaches that have been suggested towards alleviating problems related to tags include (Guy & Tonkin, 2006) the education of users and the improvement of systems to allow "better" tags. Users' education refers to providing a set of helpful heuristics to users to promote good tag selection, based on relevant research that has been carried out, indicating that the major problems with tags stem from misspells, badly encoded tags (e.g., groupings of words such as TimBernersLee), tags that do not follow convention in issues such as case and number, or single-use tags that appear only once in the database (Guy & Tonkin, 2006). On the other hand, system improvements include spell-checking, tag suggestions and synonyms (Guy & Tonkin, 2006).

An approach standing in between taxonomies and folksonomies, is that of hybrid taxonomy-folksonomy (Kiu & Tsui, 2011). Structured tags is such a hybrid method, according to which tags are provided along predefined categories (Bar-Ilan, Shoham, Idan, Miller, & Shachak, 2008), a concept that is similar to tagsonomies (Sommaruga, Rota, & Catenazzi, 2011), which feature the controlled combination of a predefined top-down classification and a bottom-up classification defined by users. The results of evaluating structured tags indicate that this hybrid approach is positive in terms of guiding users, however tag categories need to be clear so that they don't confuse users (Bar-Ilan et al., 2008). In general, three hybrid taxonomy-folksonomy approaches exist (Kiu & Tsui, 2011): (i) folksonomy-directed taxonomy, where taxonomy and folksonomy co-exist, folksonomy serves as a pool of candidate terms to enrich the taxonomy, keep the taxonomy up-to-date and allow finding of new terminology (synonyms, popular language) and concepts; (ii) taxonomy-directed folksonomy, which provides choices or suggestions to users from controlled set of terms/tags in form of drop-down menus, check boxes, type ahead or tree view, and enables more consistency and better support for findability; (iii) folksonomy hierarchies/ontologies, which may be user-powered, having a small population make the contribution, and automatically derived through statistical or clustering algorithms.

UXAmI Inspector adopts the structured tags approach and embraces the taxonomy-directed folksonomy approach to facilitate tagging and eliminate some of the aforementioned problems of tags, such as misspells, or badly encoded tags. A very simple taxonomy has been adopted, urging users to provide tags classified under the following main categories: target users, context of use, systems employed, interaction techniques, and application domain. This tags classification is employed in describing the systems and applications residing in an AmI environment, the guidelines uploaded in the UXAmI community, as well as the projects uploaded in the UXAmI OC. As discussed in section 5.4.3, users of the UXAmI OC are assisted during the process of contributing content, so as to grasp the meaning of each category and therefore avoid misconceptions or inappropriate tagging. Additionally, moderators of the OC are expected to act as a safety net towards proactively correcting and aligning user contributions. Future work will explore more sophisticated algorithms for combining tags and potentially expanding the initial taxonomy, with the aim to reduce the effort required by moderators, especially in a flourishing community with content abundance.

5.3.3 EVALUATION OF THE UXAMI INSPECTOR

UXAmI Inspector has been evaluated by three UX experts following the heuristic evaluation approach (Nielsen, 1994a), having each evaluator inspected the interface alone against the heuristic evaluation guidelines (Table 25 in Section 5.2.6). Each problem that was identified was correlated with one or more guidelines. The problems reported by each individual evaluator were aggregated into a single report, removing the duplicates and merging identical problems. Following, each evaluator provided their severity rating for each one of the problems in the unified list (ratings are explain in Section 5.2.6). The final evaluation report includes one rating for each problem, calculated as the means of the individual evaluators' ratings.

In total, seventeen problems were identified, ranging from minor (severity rating: 1) to major (severity rating: 3.33). Table 27 below lists the identified problems in order of severity, with a reference to the guidelines that are violated.

Table 27. UXAmI Inspector heuristic evaluation results

Problem	Guidelines	Severity
Inspection by multiple experts is not currently supported, whereas it is a common practice.	[2], [7]	3.33
Identification of a problem should also include relevant screenshot(s).	[2], [6]	3.33

Since a large number of guideline sets may be retrieved as relevant, functionality for adding guideline sets to favourites should be provided, while these guidelines should be displayed topmost in the list of suggested guidelines.	[7]	3.33
Additional functionality is required for the evaluator, so as to be able to aggregate the results from different evaluators to one inspection report and edit the results (e.g., when two different descriptions refer to the same problem, one of them should be kept).	[2], [7]	3.00
Once all problems are aggregated in one report, experts could individually rate their severity, while the evaluator (inspection administrator) should be able to review the individual ratings and the automatically calculated average.	[2], [7]	3.00
Problems should be able to be organized in clusters, so as to facilitate reading the report, especially for more complex systems or systems with many problems.	[2], [7]	2.67
Automatically producing an evaluation report would be useful.	[2], [7]	2.33
The tool is useful for any heuristic evaluation, and it should be allowed to create inspections not only for AmI spaces / applications, but also for any other project. One could add manually their inspection project and adequately describe it through tags along the specific categories supported by UXAmI Inspector, in order to retrieve the relevant guidelines.	[7]	2.33
New inspection – step 1: stars usually denote favourites. Better replace them with checkboxes.	[2]	2.33
When typing a tag, auto-complete would facilitate eliminating spelling errors.	[5]	2.33
New inspection: Completed steps could have a different colour than the currently active one.	[1]	2.00
The icon for contradictory guidelines (in the last step of the evaluation target definition process) resembles the one for violated guidelines (in the inspection process).	[5]	2.00
It is not possible to change the evaluation settings (targets, applications, and tags) once an inspection has been saved.	[2], [3], [7]	1.67
It would be useful if the evaluator could invite specific experts (e.g., by e-mail) to undertake an inspection.	[2], [7]	1.67
It should be possible to automatically order problems according to their severity.	[2], [3], [7]	1.67
New inspection – step 2: it would be better if no scrollbar was used in the mini-panels; instead users could scroll through (Even a long list) via the browser scrollbar.	[3]	1.67
The term pervasive application might not be straightforward for everyone.	[2], [10]	1.00

In a nutshell, the most important problems that were identified pertain to the actual inspection process, which currently supports one single evaluator, whereas it should support inspections by multiple experts and administration facilities for the evaluation coordinator. It is also notable that evaluators found the tool in general useful for usability inspections and suggested extending its usage to other contexts beyond AmI as well.

5.3.4   CONCLUSIONS AND FUTURE DEVELOPMENTS

UXAmI Inspector is a tool facilitating evaluators in setting up and carrying out expert-based reviews of AmI systems and applications. The main facilities of the tool can be summarised in the following points:

- It includes a step-by-step wizard for the specification of evaluation targets, automatically retrieving artefacts and applications residing in the AmI environment.
- It suggests guidelines relevant to the specified evaluation targets, relieving evaluators from the need to carry out extensive research for locating them and at the same time supporting evaluators in carrying out more thorough inspections by taking into account guidelines that might otherwise be neglected.
- It highlights conflicting guidelines, bringing them to the evaluator's attention.
- It supports the inspection process, and in particular recording problems, associating them with specific guidelines that are violated, and rating their severity.

The benefits of the proposed tool refer to reducing the time and effort needed to prepare an inspection, while at the same time broadening the guidelines that may be taken into account during the evaluation process. Evaluators are assisted throughout the process, however – recognizing their expertise as well as the fact that each inspection may have different goals – they are not forced to follow the suggestions of the tool. Instead they are empowered to either add parameters to systems/applications that were not used in the systems' initial description as it is retrieved by the UXAmI Inspector, or remove parameters that are considered beyond the scope of the current inspection and eventually they are authorized to select those guidelines that better suit the current inspection needs.

Other approaches in the field have attempted to suggest guidelines through questions asked to the designer (e.g., is there a standard set of icons used?) (Henninger 2000), selections made from the designer regarding the interaction style that is used (Vanderdonckt, 2001), or interoperation with development environments in order to provide component-specific guidelines (e.g., "Field labels should be followed by a colon: ") (Grammenos et al., 2000). To our knowledge, UXAmI Inspector is the first tool to automatically

suggest high level evaluation guidelines (i.e., not directly related to any specific interaction style or input control), without the need for explicit input by the evaluator.

UXAmI Inspector employs a crowdsourcing approach to retrieve guidelines and adopts a structured tag-based classification to facilitate the automatic retrieval of guidelines for the specified evaluation targets, and is therefore directly related with the UXAmI Online Community. The design of the UXAmI OC and the proposed reward scheme has addressed several challenges related to tag-based approaches. In the context of the current thesis however, as the OC has not yet been deployed, the tool has included a predefined set of guidelines (with predefined tags) to demonstrate its usage.

Future developments of UXAmI Inspector include addressing the evaluation comments and namely to expand the functionality provided so as to support concurrent inspection by multiple inspectors and management of the inspection results (e.g., to unify them and organize them in classes). Further, as suggested by the evaluators, the tool will be expanded to support inspections beyond the AmI context. Finally, the tool will be further extended so as to become independent of the UXAmI OC, allowing its operation within a single organization or even by a single individual. To this end, it will be extended with a guidelines management module, to add guidelines and provide tag-based metadata, preserving the structured tag-based classification approach.

5.4   UXAMI ONLINE COMMUNITY

During the last few years, social networking has evolved to a fundamental daily activity for many individuals and a new frontier for business marketing. Three reasons reported amongst others for building social networks and communities (Howard, 2009) are: (a) enhancing and sustaining intellectual capital, since well-led and well-managed communities provide members with access to the same kind of state-of-the-art research and thinking that one expects to find in the best university departments; (b) increasing creativity and cross-fertilization, ensuring at the same time the validity of contributions, since the community will purge erroneous material; (c) improving decision-making processes with "epistemic communities". Besides "traditional" social networking, a recent trend concerns professional network services and domain-specific network services, which are focused on interactions and relationships of a business nature around a specific target domain. These online communities (OCs) have the potential to become a platform for collective intelligence and open innovation (Leimeister, 2010), a medium for knowledge collaboration (Faraj, Jarvenpaa, & Majchrzak, 2011), as well as a trustworthy decision-support tool (Bulmer & DiMauro, 2009).

Following this approach, the UXAmI Online Community described in this section aims to become a medium for UX professionals to exchange knowledge and promote collaboration. The UXAmI OC embraces the evaluation approaches advocated by the UXAmI framework and is intended to be used as a medium to educate UX professionals in adopting the framework. The tight connection of the UXAmI OC with the framework is not only exhibited in terms of the knowledge shared, but also by providing to the community members the UXAmI framework tools. Further, a bilateral relationship between the UXAmI Inspector and the OC has been established, as the tool uses content that has been uploaded to the community and has been approved by the community moderators. More specifically, the guideline sets published in the OC are the ones used by the tool in order to suggest the most appropriate guidelines for a given evaluation target (Section 5.3.2).

An important concern for any online community is how to design it in order to stimulate regular contributions and cultivate effective collaborations. Following literature findings regarding participation motivators, as well as literature reported usability and sociability factors to reading, contributing, collaborating and leading in an online community (Preece & Shneiderman, 2009), the proposed professional network aims at providing high quality services to UX engineers, and at the same time actively engage them in collaborative and community-building activities. The UXAmI OC is built on three fundamental pillars, which are further analysed below: content contribution, active member participation and collaboration, and self-sustainability. Each one of them is crucial for creating an online community that will manage to attract a critical mass of users, to keep users interested in the community and foster social interactions and feelings of belonging, making therefore users responsible for maintaining a flourishing community and actively supporting it.

The main content that will be provided through the online community is projects uploaded by the community members, questions and answers, as well as guidelines for evaluating interactive and/or intelligent systems. Although the community stems from the need of exchanging information about the UXAmI framework, members are allowed to contribute any project relevant to interactive systems, as long as they provide accurate descriptions of their projects, in terms of targeted users, devices, and contexts of use. Questions and answers are important not only in terms of providing useful information, but also as a means of collaboration and strengthening bonds. Last, a significant category of content is sets of guidelines that can be used in the context of expert-based evaluations. The guidelines contributed by members of the OC are reviewed and approved by other OC members according to their membership level (see section 5.4.2).

Participation and collaboration is fostered through the Q&A functionality, through followers of OC members or specific projects, as well as by means of groups, and mentoring activities. Becoming a follower of a specific project or person ensures receiving notifications and updates, while groups serve the purpose of creating sub-communities focused to specific interests, with the potential of deeply engaging specific OC members and strengthening their bonds with the community. Mentoring refers to the possibility of an experienced OC member to provide guidance to other members of the community and assist them in carrying out evaluations in the context of their published projects.

The third constituent, self-sustainability is pursued through helpfulness votes offering higher visibility to high quality content, and moderation activities of OC members. Moderation involves reviewing uploaded projects, questions, and answers, reviewing and approving suggested guideline sets and tags to existing guideline sets, as well as handling violation reports. The need for professional moderators may not become obsolete through this approach, however this is a practical issue that can be examined once the OC is deployed, depending on the number of members and contributions.

The life cycle of an online community involves several stages from its birth, growth, maturity, and (hopefully not) death. Keeping alive an online community requires among others the contribution of good quality content in quantities adequate for the members' needs. To this end, this section carries out a literature review with the goal to identify what makes an online community sustainable and what motivates members towards sharing their knowledge and actively participating in the community. Despite the fact that the exact motivators for participating in an online community may vary among individuals, a well-designed reward scheme has the potential enhance users' intrinsic motivations towards participating in the community and contributing content, supporting thus the prosperity and longevity of the community (Ntoa, Margetis, & Stephanidis, 2017). As a result, this section studies state of the art approaches and best practices for rewarding user contributions and proposes an adaptive reward scheme, aiming to promote content contribution, active participation, and self-sustainability.

In summary, section 5.4.1 provides a literature review regarding motivators of user participation and knowledge sharing for OCs, as well as state of the art approaches towards rewarding user contributions. The reward scheme proposed in the context of UXAmI OC is described in section 5.4.2, while OC approaches to minimize problems related to crowdsourced content are discussed in section 5.4.3. Finally, representative mock-ups of the UXAmI OC, illustrating how the reward scheme is incorporated, are available in section 5.4.4. Section 5.4.5 concludes the discussion on the UXAmI Online Community and provides directions for future work.

5.4.1 MOTIVATING AND REWARDING USER PARTICIPATION AND KNOWLEDGE SHARING IN ONLINE COMMUNITIES

Given the importance and proliferation of online communities, several models have been proposed in an effort to predict users' intention to use and to continue using an online community. These models, built upon existing well established approaches towards predicting usage of eServices, such as the Technology Acceptance Model, the Information System (IS) success model, and the Theory of Reasoned Action, are reviewed in this section. Moreover, this section carries out a review of factors that have an impact on knowledge sharing attitudes of online community members.

In terms of acceptance and intention to use, information quality (IQ) has been identified as an important constituent of perceived usefulness, while system and service quality influence both perceived ease of use and perceived usefulness of the community (Lin, 2007). Information and system quality (SQ) also affect indirectly a member's loyalty in the community, as it was found that they influence member satisfaction (Lin, 2008; Zhang, 2010). Loyalty was also found to be affected by sense of belonging, which in turn is influenced by trust (Lin, 2008), IQ, SQ, fulfilment of needs and emotional connection (Zhang, 2010). User satisfaction and sense of belonging, have been reported to be determined by appraisal factors, namely pleasure, awareness (i.e., the degree to which a user stays informed and current with others' activities through the use of Social Networking Sites – SNSs), and connectedness (i.e., the degree to which a SNS helps users stay connected and maintain social relationships and ties) (Lin & Chau, 2014). Sociability, status, and social influences are factors that have also been highlighted to directly or indirectly influence intention to use an online community (Li, 2011). Other important factors include the perceived playfulness, ease of use, and usefulness of the site (Pai & Yeh, 2014), enjoyment (Lin & Lu, 2011), and internet self-efficacy (Wang, Chung, Park, McLaughlin, & Fulk, 2012).

Knowledge sharing constitutes a form of user participation in an online community, and several studies have explored the factors that influence knowledge sharing intentions, given its importance in communities based on User Generated Content (UGC). Privacy is an important factor that has reported to be explicitly correlated with disclosures, controlled by privacy policy consumption and privacy behaviours (Stutzman, Capra, & Thompson, 2011). Trust, which is partially affected by privacy, has been found to have a direct positive effect on knowledge sharing (Lin, Hung, & Chen, 2009) and psychological safety (Zhang, Fang, Wei, & Chen, 2010), as well as on the quality of information shared (Chang & Chuang, 2011). According to technology acceptance studies, trust further influences one's sense of belonging, which was also found to affect knowledge sharing intentions (Chai & Kim, 2012), and is - among others - influenced by one's familiarity with the community members (Zhao, Lu, Wang, Chau, & Zhang, 2012). Social

interactions within the members of a virtual community was also found to have a positive effect on the quality of content contribution (Chang & Chuang, 2011). Another constituent of sense of belonging is the social networking sharing culture (fairness, identification, and openness), which was also identified as a significant factor affecting knowledge sharing intention (Pi, Chou, & Liao, 2013; Chai & Kim, 2012).

Several intrinsic motivators have been also reported to affect knowledge sharing intentions in an OC, such as altruism and enjoying helping (Chang & Chuang, 2011; Zhao, Stylianou, & Zheng, 2013; Liao, To, & Hsu, 2013), but also reciprocity (Cheung, Lee, & Lee, 2013), outcome expectations (Chiu, Hsu, & Wang, 2006; Chen & Hung, 2010) and satisfaction with the online community (Jin, Zhou, Lee, & Cheung, 2013; Chang, Hsu, Hsu, & Cheng, 2014). Furthermore, members' self-efficacy is a parameter with both direct and indirect effects on knowledge sharing behaviour in professional online communities (Lin, Hung, & Chen, 2009; Chen & Hung, 2010), as well as their experience in practice and tenure in the occupation (Wasko et al., 2009). Finally, another intrinsic motivator with profound impact on knowledge sharing is reputation, which was found to have positive effects on the quality of shared knowledge in virtual communities (Chang & Chuang, 2011) and electronic networks of practice (Wasko, Teigland, & Faraj, 2009), as well as a motivator towards providing meta-information and joining social structures (Nov, Naaman, & Ye, 2010).

Given the multifaceted nature of motivators for active user participation and content contribution, several efforts have attempted to provide a classification of motives. Sun, Fang and Lim (2012) provided the following classification of motives to share knowledge in virtual communities: personal motives, which can be further classified into extrinsic motives (extrinsic reward, reputation/image) and intrinsic motives (sense of self-worth, learning, enjoyment in helping others); and social motives, including community advancement, social identity, reciprocity, and sense of belonging. On the other hand, Fugelstad et al. (2012) have studied three general classes of motivations that might influence user participation in online communities: general volunteer motivations, pro-social behavioural activity, and community-specific motivations in the context of a movie recommender system based on user ratings. The results of the study indicated that different motivations, and different histories of pro-social behaviour, led to different patterns of behaviour. For instance, people with more volunteer experience were found to be more likely to edit a movie, while people with higher community involvement were more likely to invest effort in the community by visiting the Question and Answer (Q&A) forums. Similarly, different community-specific reasons for joining the community predicted different patterns of behaviour.

The role of extrinsic and intrinsic rewards has also been extensively studied, with some studies focusing on intrinsic, other on extrinsic, and other on both, while there seems to be a contradiction regarding the impact of extrinsic rewards. Belous (2014) carried out a survey with 897 bloggers, Wikipedians, forum

participants and website writers and identified that although a wide variety of rewards may influence one's decision to share online, the strongest motivators are intrinsic rewards and self-efficacy. Examining the impact of extrinsic motivation in relation to task complexity, Sun et al. (2012) found that when task complexity is low, extrinsic motivation has a positive and significant influence on continuance intention, while when task complexity is high, extrinsic motivation has no effect on it. Economic gains have been identified as a motivator of participation in peer-to-peer knowledge sharing communities, along with the enjoyment of the activity and the sustainability of the community itself (Hamari, Sjöklint, & Ukkonen, 2016). On the contrary, a study carried out with data collected from a large social Q&A site with millions of registered users indicated that extrinsic rewards such as virtual organizational rewards undermine the effect of enjoyment in helping others towards sharing knowledge for active members, while the effect of self-efficacy is undermined by reciprocity (Zhao, Detlor, & Connelly, 2016).

Several studies have considered the impact of specific rewards in online communities. One type of rewards includes reputation (or status markers) and is based on historical information of a user's contributions (Kraut & Resnick, 2011). Such rewards may include experience points, virtual wealth, level in the community and ranking (best answer rate) and may have an impact on the quantity of contributions, not their quality however (Lou, Fang, Lim, & Peng, 2013). Such a reputation system provides track records of knowledge contributors' past activity, thus promotes trust to knowledge seekers and also acts as a positive feedback to users' competence (Lou et al., 2013). Rewards may also include attention, recognition, commendations, compliments, and praise, and they can be extremely powerful incentives as long as they are public, infrequent, credible, and culturally meaningful (Tedjamulia, Dean, & Albrecht, 2005). Finally, another mechanism for rewarding users is via privileges, which may include access permissions to the activities of the community (such as read-only access, permission to post content, privilege of moderating other's content, uploading a personal photo to their profile) and can serve as a status symbol or a validation of a member's competence and loyalty to the community (Kraut & Resnick, 2011). The impact of gamification in applying the aforementioned rewards through points, levels, leader boards and badges has also been reported to positively increase contributions (Hamari, 2017) and user performance (Mekler, Brühlmann, Opwis, & Tuch, 2013).

Given the above, a popular reward mechanism may involve a point system, according to which the community members are rewarded for their contributions. A point-based system, featuring four membership levels has been employed by Farzan et al. (2008). The use of membership levels aimed to provide some benchmarks for users to know how they stand in relation to how many points they have, assisting them in setting goals and seeing a change in status as a reward for their site activity. The main characteristics of the reward system were that:

- it was determined which content generated the most activity on the site and that content was rewarded with higher amounts of points
- moving out of the first point class to the second was made very easy with the aim to motivate users in order to get started; moving from the second level to third required more work, encouraging an active level of participation; while the hardest level to achieve was the fourth with the rationale that the majority of users should be in the third class, keeping them motivated to contribute
- the user can compare themselves with everyone in their network in the online community, so as to be able to make easy comparisons with other people on the site
- with the aim to provide users with the personal benefit of increased reputation, because of the social focus of the site, users with the most points were granted greater visibility. This was incorporated into the site be revealing point-related information on different parts of the system. For instance, the ten users with the highest number of points were shown on the homepage of the community. On the other hand, the number of points and the class label of every user are always shown on the name badge on users' profile page and anywhere on the site where a list of users is shown.

Other interesting features that have been employed by various reward and motivation systems in online communities, involve:

- Making recommendations to new users during their sign-up process. The recommender system exploits external social media to produce people and profile entry recommendations for new users (Freyne, Jacovi, Guy, & Geyer, 2009). The results of the relevant user study showed that users who received recommendations at sign-up created more social connections, contributed more content, and were on the whole more engaged with the system, contributing more without prompt and returning more often. In general, two main types of recommendations are the most prominent in social recommender systems, and in particular recommendation of social media content and recommendation of people (Guy, 2015).
- Establishing an individual identity and profile (Berlanga, Bitter-Rijpkema, Brouns, Sloep, Fetter, 2011) and allowing the user to develop relationships with other users in the community (one would do a favour to a friend, but not for anonymous people) (Bretzke & Vassileva, 2003).
- Self-organizing sub-community. The sub-community creation feature facilitates subscribers to organize their own sub-communities (Lui, Lang, & Kwok, 2002). In other words, a community can be composed of many sub-communities that are owned by particular subscribers. This feature can benefit both the subscription system and the subscribers, since for instance the subscription system can support decentralized marketing and delegate some management tasks to the sub-communities owners. Employing sub-groups in communities has been claimed to increase

identity-based attachment to the group, enhancing thus the user's bonding with the community (Tausczik, Dabbish, & Kraut, 2014).

- Peer recommendation. Peer recommendations can be constructed based on peer evaluations, reviews, and ratings (Lui et al., 2002), an approach which can act as a tool to assist in identifying trustworthy members in the community. Users' rating is a complicated process and does not merely depend on individual review quality alone, instead it is influenced by social factors, such as the herding effect (i.e., users' awareness of previous votes on a review) (Sipos, Ghosh, & Joachims, 2014). On the other hand, as users become more popular in the community, they produce more reviews and better reviews, yet their ratings become more negative and more varied (Goes, Lin, Au Yeung, 2014). The impact of negative reviews is also a controversial issue, recently indicated as a parameter affecting negatively the quality of future contributions of the community member who received negative reviews (Cheng, Danescu-Niculescu-Mizil, & Leskovec, 2014).
- Number of membership levels. To ensure that people will not be confused by the hierarchy of the memberships, the number of the membership levels should not be greater than six; to distinguish the users with different participation levels, the number should not be less than three (Cheng & Vassileva, 2005).

Finally, a noteworthy approach is that of adaptive rewarding mechanisms. Cheng and Vassileva (2006) proposed such a mechanism to measure the quality of user contributions, control the overall number of contributions in the community, and motivate users to contribute high-quality resources. On one side, the mechanism aims to encourage users to rate contributions, thus ensuring decentralized community moderation. To this end, each user receives a limited number of rating points to give out, while the users with higher membership levels receive more points to give out, which makes them more influential in the community. On the other side, the mechanism intends to influence the individual users' actions of contributing by adapting the rewards using a model of the current needs of the community and a model of the users' individual reputation in contributing quality resources. In more details, rewards are adapted for different forms of participation for individual users to the user's current reputation (based on the quality of their contributions so far) and the current needs of the community. The individual rewards for each type of action are displayed in personalized motivational messages, which the user sees at login, outlining what the community expects from the user in terms of quantity and quality of contributions. Also, the timeliness of a contribution (according to the needs of the community) is an important factor that impacts the granted rewards. An evolution of the aforementioned mechanism (Vassileva, 2012) proposes that adaptations should be threefold with the aim to: (i) optimize system behaviour with respect to all the users in the system, by increasing the number and quality of contributions, binding users in social ties, enticing users to coming to a common goal, and making the community self-sustainable; (ii) provide

rewards for particular individually weighted actions; and (iii) visualize the community adaptively to emphasize particular incentives.

5.4.2 THE PROPOSED REWARD SCHEME

The goal of the proposed reward scheme is threefold: (i) to motivate users towards knowledge sharing; (ii) to encourage active engagement, not only through content contribution but also through participating in the community and rating content, keeping thus the community active and providing at the same time a quality-control management by the community itself; and (iii) to foster high quality contributions, laying the foundations for a sustainable community. Taking into account the knowledge contribution motivators identified in the literature, state of the art reward schemes, and best practices of well-established online communities, the proposed reward scheme:

- features a system of points and membership levels, accompanied by specific privileges for each membership level
- employs reputation mechanisms, through top contributors and badges
- fosters the adaptive rewarding approach, by providing adaptations to rewards, recommendations, and content promoted to each user
- embraces rating content by endorsing its helpfulness.

The initial reward scheme has been evaluated through a co-design workshop approach (Sanders & Stappers, 2008), where participants provided their reviews through immersive activities (such as voting, word racing games, throw-and-catch the ball game, and brainstorming). The workshop highlighted that facilities for creating a complete user profile that were initially foreseen as a first membership level reward should be provided for all users. An important concern was raised towards the impact of negative ratings, as well as an initially foreseen mechanism limiting the number of ratings that a participant would be allowed to contribute (rating points to give out). This mechanism was initially proposed to discourage community members from providing inaccurate negative reviews and ratings, however it was considered too complicated for users to comprehend, with ambiguous impact. As a result, it was decided not to be included in the reward mechanism. Additional features that were suggested were annual awards with an offline ceremony, as well as free registrations to UX events, which will be considered in future versions of the community.

Overall, the mechanism foresees four membership levels, each associated with specific privileges bound with the UXAmI community, targeted at further enticing users to participate in the community and offering at the same time high visibility and reputations. Users will be able to advance to higher

membership levels according to their points. As soon as a user registers, they are placed at level 1, being able to carry out all the fundamental activities of the community, such as view projects, questions and answers, add profile information (e.g., avatar, website, twitter or linked account), create their own project, post a question, or post answers. Advancing to the next level only requires a minimum activity, such us completing profile information, uploading posting a question and/or a couple of answers, or creating a project. Level 2 offers more substantial privileges, such as creating groups, uploading evaluation guidelines for expert-based reviews, rating content (e.g., projects, answers, users), and gaining access to the UXAmI Inspector tool. Moving from level 2 to level 3 requires active participation, through rating and content contributions. Level 3 members enjoy privileges indicating that they are valuable to the community, such as becoming a mentor to a project and gaining additional visibility, access to the electronic version of the UXAmI framework, and the ability to moderate tags posted by the community members for specific guidelines, and review sets of guidelines posted. The goal is to keep the majority of users at level 3, in order to be motivated and to contribute to the community. Therefore, moving from level 3 to level 4 requires substantial content contributions and rating activity of good quality. Level 4 members constitute the elite of the community and act as community moderators, reviewing guidelines, tags, violation reports, and gaining access to the UXAmI Observer tool. Figure 84 illustrates an overview of the membership levels and privileges of the proposed reward scheme. Users will advance levels according to their points, and in particular: from 0 to 199 points users are ranked at Level 1, from 200 to 499 points at Level 2, from 500 to 999 points at Level 3, and from 1000 points onwards users will reach Level 4.
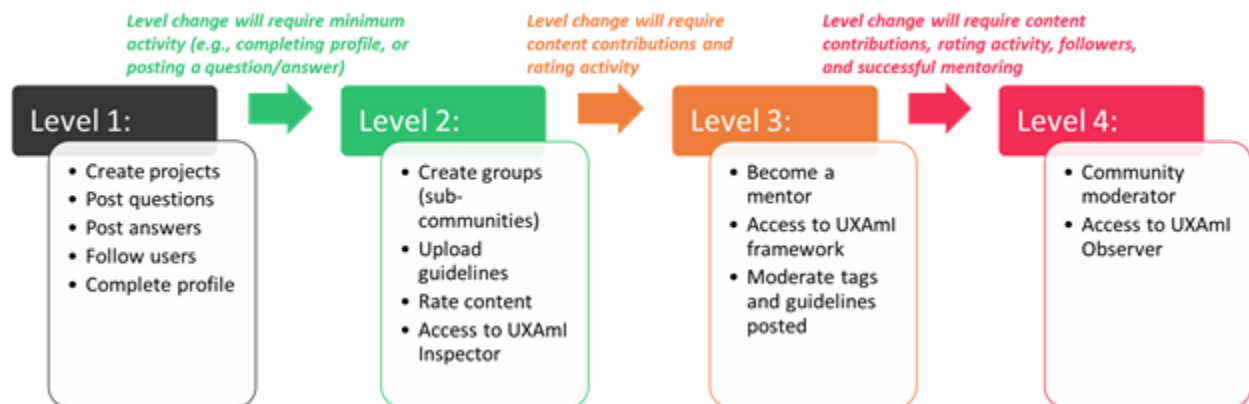
Figure 84. Membership levels of the UXAmI Online Community

Content rating activity is important in terms of keeping the content quality high and promoting content with high value, so that it can be easily accessed by a wide audience. Although it is common practice to provide ratings ranging from 1 to 5, with ratings below 3 being negative and 3 representing neutral

feelings, taking into account recent literature that highlights the negative impact of negative ratings and also the results of the co-design workshop, a quite simple rating mechanism is provided. Users are able to vote positively a project or an answer by indicating that it was helpful, as well as to follow users who they believe are influential. Furthermore, they are able to report projects, answers, questions, or community members who violate the terms and conditions of the OC (e.g., intellectual property violation, offensive material). Correctly reporting inappropriate content or individuals exhibiting inappropriate behaviours is awarded with status points.

Reputation is an important component of the reward scheme, aiming to enhance users' value and recognisability in the community. Reputation is achieved by means of top contributors list, badges, as well as by indication of one's membership level, number of followers, number of helpfulness votes for answers and projects, number of groups moderated, and number of projects mentored. The list of top contributors (hall of fame) is refreshed on a weekly basis, so as to provide a chance to members of any level to be included in the list. It displays the top contributors based on the points accumulated the current week. Badges are awarded to users according to their contribution in the community, as listed in Table 28. A user may collect one or more badges of each type, and view them in their collection which is available through their profile page.

Table 28. Badges awarded to the UXAmI OC members

Category	Badge
Project-related	Most-viewed project (of the week / month)
	Most helpful project (of the week / month)
	50 project followers reached
	100 project followers reached
	200 project followers reached
	500+ followers reached
	5 projects contributed
	10 projects contributed
	25 projects contributed
	50 projects contributed
Answer activity	Response promptness for a specific question (first to answer)
	10 prompt responses
	50 prompt responses

	100 prompt responses
	Most helpful response for a question (also displayed first)
	10 most helpful responses
	50 most helpful responses
	100 most helpful responses
Followers	100 followers reached
	200 followers reached
	500 followers reached
Groups	Medium size group owner (with at least 50 users)
	Large size group owner (with at least 100 users)
	Extra-large size group owner (with at least 300 users)
Mentoring	Successful mentoring of 5 projects (success is determined by mentor's rating score, which should be larger than 7/10)
	Successful mentoring of 20 projects
	Successful mentoring of 50 projects
Moderating	Moderation of 25 guideline sets
	Moderation of 50 guideline sets
	Moderation of 100 guideline sets
	Moderation of 100 tags
	Moderation of 250 tags
	Moderation of 500 tags
	Handling 25 violation reports
	Handling 50 violation reports
	Handling 100 violation reports

With the aim to provide incentives for timely and high quality contribution, the rewarding mechanism is adaptive in awarding points for timely answers, number of followers, and violating the terms of the OC. Timely answers to questions are awarded with a higher amount of points, the maximum being 25 points for any answer provided within the first day, decreasing exponentially according to Equation 3, where x is the number of days within which the answer is provided. No reward points are awarded if the question is answered after 25 days have passed. A graphic representation of the points awarded according to the timeliness of answer is presented in Figure 85.

$$f(x) = {}^{25}/_x \hspace{8cm} (3)$$
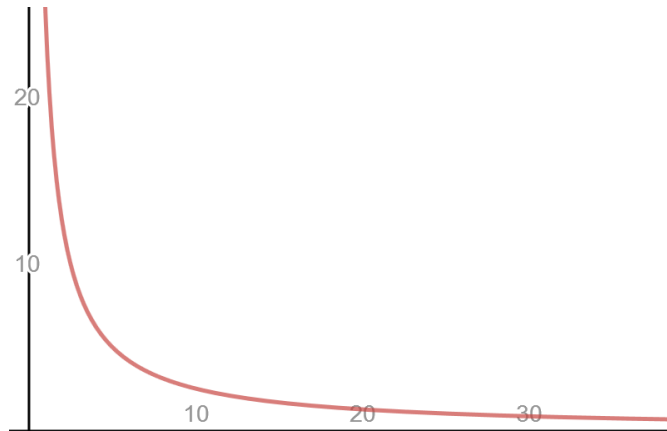
Figure 85. Graphic representation of points awarded according to the timeliness of an answer

The number of points awarded to a user who is followed by other members of the OC, or whose projects are followed, depends on the membership level of their follower. More specifically, when a user acquires a new follower, they are awarded as many points as their follower's membership level. When a project acquires a new follower, the project owner is awarded with half as many points as their follower's membership level. The same equation is applied for awarding points when a new OC member joins one's group. Correspondingly, the same number of points is deducted if a follower or group member is lost.

Violating the terms and conditions of the community is a behaviour that should be effectively deterred. Members who violate terms are deducted status points following an aggressive policy regarding repeated misbehaviours, therefore the number of points removed is analogous to the number of violations n, according to Equation 4.

$$f(n) = -n \times 100 \hspace{8cm} (4)$$

Through their activity in the community users accumulate points, advancing therefore to higher membership levels. In summary, points are awarded for registering to the OC, logging in, creating projects, posting answers, being followed, receiving helpfulness votes for a project or an answer, creating a group, gaining members to one's group, correctly reporting a violation, contributing approved new guideline sets, providing new correct tags for a guideline set, successfully mentoring a project, moderating guidelines and tags, as well as moderating questions and answers. There are also cases where points are lost, such as long-time of inactivity, violating terms and conditions, and falsely reporting violations. The exact points to be gained or lost are listed in Table 29.

Table 29. Points per user activity

User activity	Points
Registration to the OC (once)	+100
Login to the OC (once daily)	+1
Add profile information	+20
No log-in for fifteen consecutive days	-5
No activity at all over a month's time	-30
Creating a project	+20
Receiving a helpfulness vote	+1
Reaching 50 helpfulness votes for a project or answer	+50
Reaching 100 helpfulness votes for a project or answer	+100
Acquiring a new project follower: points will depend on the follower's membership level *(fl)*	*+fl/2*
Losing a project follower: points will depend on the follower's membership level *(fl)*	*-fl/2*
Reaching 50 followers for a project	+25
Reaching 100 followers for a project	+50
Reaching 500 followers for a project	+100
Reaching 1000 followers for a project	+200
Losing 50 followers for a project	-25
Losing 100 followers for a project	-50
Losing 500 followers for a project	-100
Losing 1000 followers for a project	-200
Providing an answer to a question: points will depend on the number of days *(d)* within which the answer was provided	*+25/d*
Acquiring a new profile follower: points will depend on the follower's membership level *(fl)*	*+fl*
Reaching 50 profile followers	+25
Reaching 100 profile followers	+50
Reaching 500 profile followers	+100
Reaching 1000 profile followers	+200
Losing a profile follower: points will depend on the follower's membership level *(fl)*	*-fl*
Losing 50 profile followers	-25

Losing 100 profile followers	-50
Losing 500 profile followers	-100
Losing 1000 profile followers	-200
Creating a group	+50
Acquiring a new group member: points will depend on the user's membership level *(fl)*	*+fl/2*
Reaching 50 group members	+25
Reaching 100 group members	+50
Reaching 500 group members	+100
Reaching 1000 group members	+200
Losing a group member: points will depend on the user's membership level *(fl)*	*-fl/2*
Losing 50 group members	-25
Losing 100 group members	-50
Losing 500 group members	-100
Losing 1000 group members	-200
Correctly reporting a violation	+20
Contributing a new guideline set (approved by the moderators)	+20
Contributing a new tag for an existing guideline set (approved by the moderators)	+2
Successfully mentoring a project	+50
Moderating a guideline set	+10
Moderating a newly contributed tag	+1
Handling a violation report	+10
Moderating a question	+2
Moderating an answer	+2
Violating rules and policies of the OC: points will depend on the number of violations (n)	-n*100

Aiming to support identification, each user has a profile page which showcases their contributions to the UXAmI community. For instance, the profile page features all the user's projects, answers, questions, and badges, it indicates the membership level and provides prompts as to how to advance to the next level, while it also provides access to the user to privileged material. Last but not least, each user is able to receive personalised messages prompting them to answer questions, view projects and join groups relevant to their interests.

With regard to gamification, the UXAmI OC - through the proposed reward scheme - adopts the game element hierarchy of dynamics, mechanics and components (Werbach & Hunter, 2015), where dynamics are high-level features that provide motivation (e.g., narrative or social interaction), mechanics are the elements that drive player involvement (e.g., rules, feedback, rewards), and components are the specific examples of high level features (e.g., points, collections, virtual goods). Starting from the lowest level, the components of the gamification approach are instantiated through the individual features of the UXAmI reward scheme, and namely points, levels, badges, content unlocking, leaderboards, and teams. In terms of mechanics, the gamification approach that has been applied includes challenges communicated to users through prompts, competition to achieve the best score and appear in the top contributors list, cooperation through mentoring project and commenting facilities, feedback of one's membership status, resource acquisition through gradually gaining access not only to new functionality in the OC but to tools as well, and rewards materialized through the reward scheme. Finally, the main dynamics components that have been employed are progression of the individual (in terms of understanding and abilities) achieved with higher engagement in the OC, and relationships which are promoted through followers and subgroups, as well as through mentoring and commenting others' projects.

In summary, the proposed reward scheme aims to motivate users towards contributing content and encourage their active participation, by rewarding high quality contributions, as well as active and loyal members. Rewarding is achieved through providing privileges to active and loyal members and making these members visible to the entire community, thus enhancing their reputation. In addition, OC members of higher levels also act as safeguards for the community, as they are assigned with the privileged task of moderating content and checking for violation reports.

5.4.3   ASSISTING, COMBINING AND MODERATING USER CONTRIBUTED CONTENT

Besides a personal repository and a source for knowledge retrieval, the UXAmI Online Community aims to be used as a source for guidelines for the UXAmI Inspector Tool (Section 5.3). Users can contribute guidelines and describe them through tags, so that they can be appropriately matched by UXAmI Inspector with the evaluation targets. The challenges that reside in this approach include that the content should be appropriate and relevant, avoiding duplications, and that tags should be as error-free as possible. To this end, the UXAmI Online community assists users when providing tags, assures through assisted moderation that there are no duplicates and that the content is appropriate, and fosters high quality contributions through rewarding them, as already explained in the previous section.

To help users in correctly typing tags, an auto-complete functionality suggests words that match what the user has typed so far. Suggestions of the auto-complete functionality are based on tags that have already

been provided by other OC members and approved by moderators. A spell-checking Application Programmable Interface (API)[16, 17] can also be used towards eliminating misspelled words. Finally, users can view clarifications regarding the categories under which they are asked to provide tags, through a short explanation and examples.

An important concern in the design of the UXAmI OC was how to assist moderators in reviewing uploaded guidelines and tags and locating duplicates. At first, when a new guideline set is uploaded in the community, moderators are notified through the notifications section of the OC (bell icon in the top menu bar) and by email as well. Newly uploaded and not-yet-approved guideline sets are listed under a separate menu item in the moderator's profile page (Figure 97). As soon as a moderator decides to review a guideline, they can select the "REVIEW" button, indicating that they have undertaken this task, to save time from other moderators who might decide to review the specific guideline set as well. Then, the moderator can select to review the guideline set, read the guidelines as well as the provided references. Moderators have full editing capabilities on the uploaded guideline set, so as to eliminate possible typing errors. While reviewing a guideline that has been uploaded, they are able to also see other guideline sets that the system found to be similar to the one they are reviewing. To this end, pattern-matching algorithms for strings can be applied (Alfred, 2014), looking for patterns in the title of the guideline sets, the contained guidelines, and their references and calculating a similarity score. This however, constitutes a topic of future research before the deployment of the OC. This similarity match algorithm and suggestions to potentially similar guideline sets can also be made available to users, while uploading a guideline set. As a result, they will be able to review other similar content and avoid uploading again a guideline set that is already approved by the community. In case a user would like to suggest an improvement on an accepted guideline set, they are able to do so, and explain their improvement. Suggesting an improvement triggers a notification to content moderators, who can eventually accept or reject the suggested modification of the guideline set.

Figure 86 illustrates the moderation screen of a guideline set that has been suggested, featuring the guideline set details, information about the contributor and a direct link to their profile, and the similarity score. The design also facilitates the direct communication of the guideline contributor with the reviewer, so as to easily resolve any minor issues and answer questions.

Another concern regarding guidelines is that they often are in conflict, perhaps due to originating from different sources and addressing different perspectives (Massip, 2012). To assist UX engineers in appropriately selecting guidelines, the OC provides an option for reporting a conflict. When two guidelines are reported as conflicting, they are both annotated with an icon (exclamation mark) and details regarding the guidelines with which they are in conflict are provided (Figure 87). These details include the name and

---

[16] https://www.webspellchecker.net/web-services.html
[17] https://docs.microsoft.com/en-us/azure/cognitive-services/bing-spell-check/proof-text

set of the conflicting guideline, as well as the explanatory comment that was provided by the user who reported the conflict. Conflict reports are moderated, like newly contributed tags and guideline sets. Furthermore, next to a guideline set, all the projects that have employed it in the context of an evaluation are presented, ordered by the number of helpfulness votes received. As a result, the UX engineer can view examples of other projects and how the specific guideline set has been employed in practice.

Figure 86. UXAMI OC: Guideline moderation

Figure 87. UXAmI OC: Guideline view

5.4.4 HIGH-FIDELITY PROTOTYPES OF THE UXAMI OC

This section includes high fidelity mock-ups of the professional network for UX engineers and briefly discusses the design rationale. The UXAmI professional network features a simple and aesthetic design aiming to create a strong bond between UX engineering and community-oriented activities. The design ensures that pages oriented towards UX engineering information include details and prompts for community-building.

The network features a simple navigation menu, with options towards projects contributed by community members, guidelines for various design cases, as well as a Q&A section. To assist users who prefer searching over browsing, an advanced filtering mechanism and a search facility is provided. Finally, personalized options are readily available for each user, including notifications, personal profile, settings and help (drop-down menu through the arrow next to the user profile icon).

The home page of the UXAmI OC (Figure 88) serves a threefold purpose regarding the content displayed: to provide recommendations to the user regarding potentially interesting content (projects, members, or questions), to promote the most helpful projects, thus allowing users to easily select them and also to directly follow or vote them, as well as to display recent questions and groups in case the user is interested in them. With regard to the reward scheme, the indication of one's membership status is readily available on the right column, featuring prompts as to how to increase one's points and advance to higher membership levels. In addition, the list of top community members (hall of fame) is displayed in a visible place (above the fold), enhancing the reputation of active and high quality contributors. Each entry in the hall of fame illustrates the avatar of the user, their name, membership level and number of followers. Furthermore, the live feed aims to show activity from the user's network, acting both as a path to potentially interesting content and as a motivator to actively contribute to the community.

The "Projects" section includes UX evaluation projects contributed by the community members. For each project, the following information is available (Figure 89): upload date and contributor, target users, context of use, systems involved, interaction techniques, application domain, evaluation methods and number of helpfulness votes (if any). The user may follow or vote a project as helpful, while additional actions regarding a project (e.g., report, comment, mentor, or pick as interesting) are available from the project page, where one it is possible to view all the project details. Projects can be browsed according to the following categories: all projects, recently uploaded, most helpful, picks, recommended projects for the specific user, projects from the user's network, and projects that the user follows. Picks from moderators and recommendations per user have been introduced to counterbalance the effect of higher visibility of most recent or highest rated contributions. Recommendations will be based on the content

contributed by the user, such as projects, questions, answers, and guidelines, framing the user's current domains of interest. Finally, filtering options are available to facilitate users in reducing the number of results displayed.

A UXAmI Project is an entity with a wealth of information that must be delivered in a well-structured manner. Information about a UXAmI Project revolves around the system description, details about the evaluation approaches that were applied to the specific system, and reviews from the community. In more details, the information provided for a UXAmI project includes (Figure 90):

- Description of the project and summary of main attributes. Attributes are provided by project contributors, following the structured tag approach that has been discussed in Section 5.3.2, along the following tag categories: target users, environment, systems involved, interaction techniques, application domain and evaluation methods.
- Number of helpfulness votes (if any).
- Multimedia gallery showcasing the project through photos, videos, or images.
- Details about each evaluation approach and relevant material.
- Contributors and mentor (if any).
- Relevant files and publications from where users can retrieve additional information.
- Other relevant projects ("users also viewed").
- Comments that have been posted by the community members regarding this project. Comments may refer to questions or suggestions, while the project owner has the option to remove a comment from being visible.

In addition, several options are readily available to the user, including among others following the project, voting it as helpful, or reporting a violation for this project, as well as posting a public comment.

Figure 88. Home page of the UXAmI Online Community

Figure 89. UXAmI OC: Projects page

Figure 90. UXAmI OC: Project page

The guidelines section (Figure 91) lists all the guidelines that have been approved. The sub-menu on the right allows the user to view guidelines recently uploaded, those recommended for the user, as well as

those suggested by the user, and is accompanied by an option to submit a new guideline. Filters are also available to reduce the number of displayed guidelines. For each guideline, its title is displayed which also serves as a link to the guideline details, as well as the tags that have been provided by the users under the structured tags approach, along the following categories: target users, contexts of use, systems, interactions, and application domain. When no tags have been provided for a specific category, the displayed value is "any". The contributor of each guideline is also displayed next to the guideline, as a means of identification and reputation within the community.

Figure 91. UXAmI OC: Guidelines

The Q&A section (Figure 92) follows a simple structure, with the top of the page dedicated to the user's question (if they wish to post one) and the remaining of the page displaying questions that have been posted to the community. Asking a question requires only typing the question and optionally adding

material (e.g., documents, presentations, images) to clarify one's question. Questions that have been posted by other members of the community can be browsed through the right-hand menu, which provides direct access to all questions, those recently uploaded, questions in the user's field of interests, the ones from their network, as well as questions that the user has asked. Each question includes the profile photo of the user who asked, as well as their name, an indication of when it was asked, the question itself and the number of answers. The user can either select to answer the question, or click on the question to view its entire answers thread. Finally, the most helpful answers are promoted on the right side bar, as a means of increasing users' reputation and for their potential to assist other community members in retrieving useful information.

Figure 92. UXAmI OC: Questions page

The profile page of a member of the OC (Figure 93) features their membership level and contact information (on the right). Top most displays the OC member's photograph, name, and occupation, as well as their contribution in the OC regarding projects, guidelines, questions and answers. In addition, the user is able to follow or report the specific OC member by clicking the corresponding buttons. Next, follow details of the member's contributions, namely projects and questions (in a timeline representation), as well as their badges. For each project that has been uploaded by the OC member, the user will be able to view its title, date, a photo, and the tags that have been provided to describe the project. Also, the user can directly follow or vote the project as helpful. The option of reporting a project is only available in the project page, ensuring that before reporting users will have adequate knowledge of the specific project, thus discouraging hasty decisions and impulsive actions.
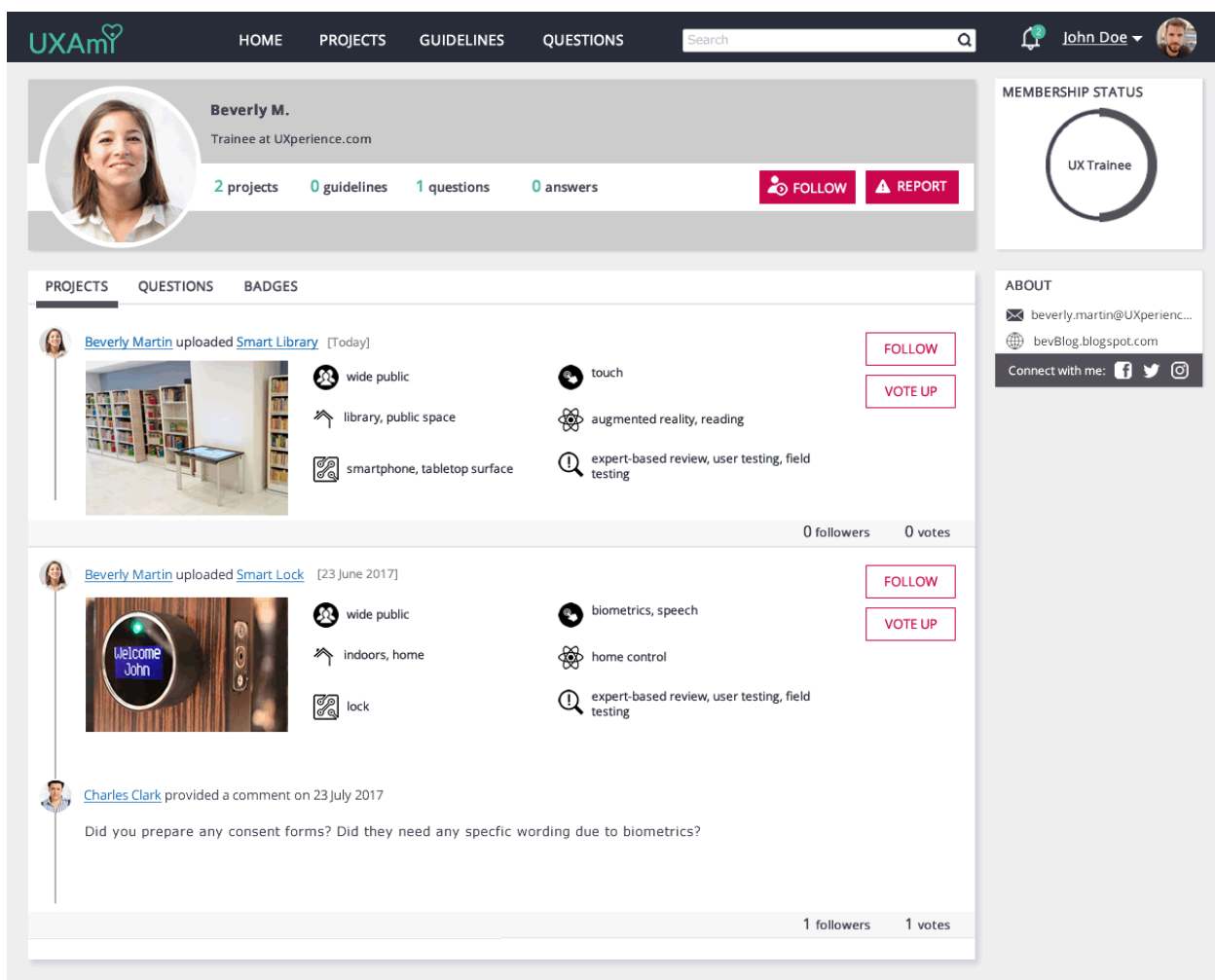

Figure 93. UXAmI OC: Profile page of a user at membership level 1 (as seen by others)

The profile page as users see it for themselves is similar, with a few differences however (Figure 94). First, below one's membership level status, a suggestion as to how to advance to higher membership levels is provided, with a link that will display the privileges of the next membership level. On the right side, one's most recent badges are available, or a link to the badges collection if the user has not yet collected any badges. The user profile area features two additional buttons, one for editing one's profile and one for directly adding a project. Next, three panels follow showcasing the user's most recent activity (e.g., projects, questions or answers uploaded). If less than three contributions have been made, the user is prompted to contribute more content, by adding a project or answering a question potentially in their field. The area below serves as a catalogue of one's contributions and a direct access to the privileges provided at each membership level. For the first membership level, the user's contributions are displayed in a timeline format (having aggregated all content contributions under one timeline), while the relevant submenu also includes a link to their badges.

The design of the OC also supports identification of one's membership level through colours, as each level has been associated with a specific colour: dark grey for level 1, royal blue for level 2, silver green for level 3, and teal (as the UXAmI logo) for level 4. As a result, when one's card is displayed next to a contributed project or guideline, it is accompanied by the colour that corresponds to their membership level.

The profile page of a user at the second membership level adopts the same structure and design rationale as the page of users at the first membership level, being enhanced however with the additional options provided at this level, and featuring the royal blue colour (Figure 95). The additional options include: (i) the tools panel providing access to UXAmI inspector and a prompt to upgrade level to access more tools, (ii) the "GROUPS" menu option, allowing users to view and manage their groups, a privilege provided at this level. The "Latest Activity" panel may feature:

- a contributed project, highlighting its title, date, and number of helpfulness votes
- guideline sets contributed, including their title, the included guidelines, date, and status (i.e., approved, under review, pending, or rejected), and
- the groups created, providing the group logo, title, date, and number of members.

The user timeline is actually structured as several mini-timelines revolving around a specific contributions. For instance, in Figure 95 the mini-timeline regarding the group that has been submitted includes the group creation event, as well as when each member joined the group. Each item in the main timeline can be displayed as recent, when an event has occurred in its mini-timeline (e.g., although a group may have been created two months ago, it will appear in the main timeline if a new member has joined or an existing member has left recently).

Figure 94. UXAmI OC: User profile page (membership level 1)

Figure 95. UXAmI OC: User profile page (membership level 2)

Groups mainly act as targeted discussion and information dissemination areas (Figure 96). Each group features a logo and a short description of its goals, as well as a discussions area. Discussions are based on a two-level hierarchy, as responses to a posted comment are linear and nested responses are not supported. Each comment in the discussions is accompanied by a photograph of its owner and a link to their profile, as well the date it was posted. Finally, the group page displays the group owner, as well as the group members, and provides direct access to the related actions of joining the community, following the owner and posting a comment (available to community members only).
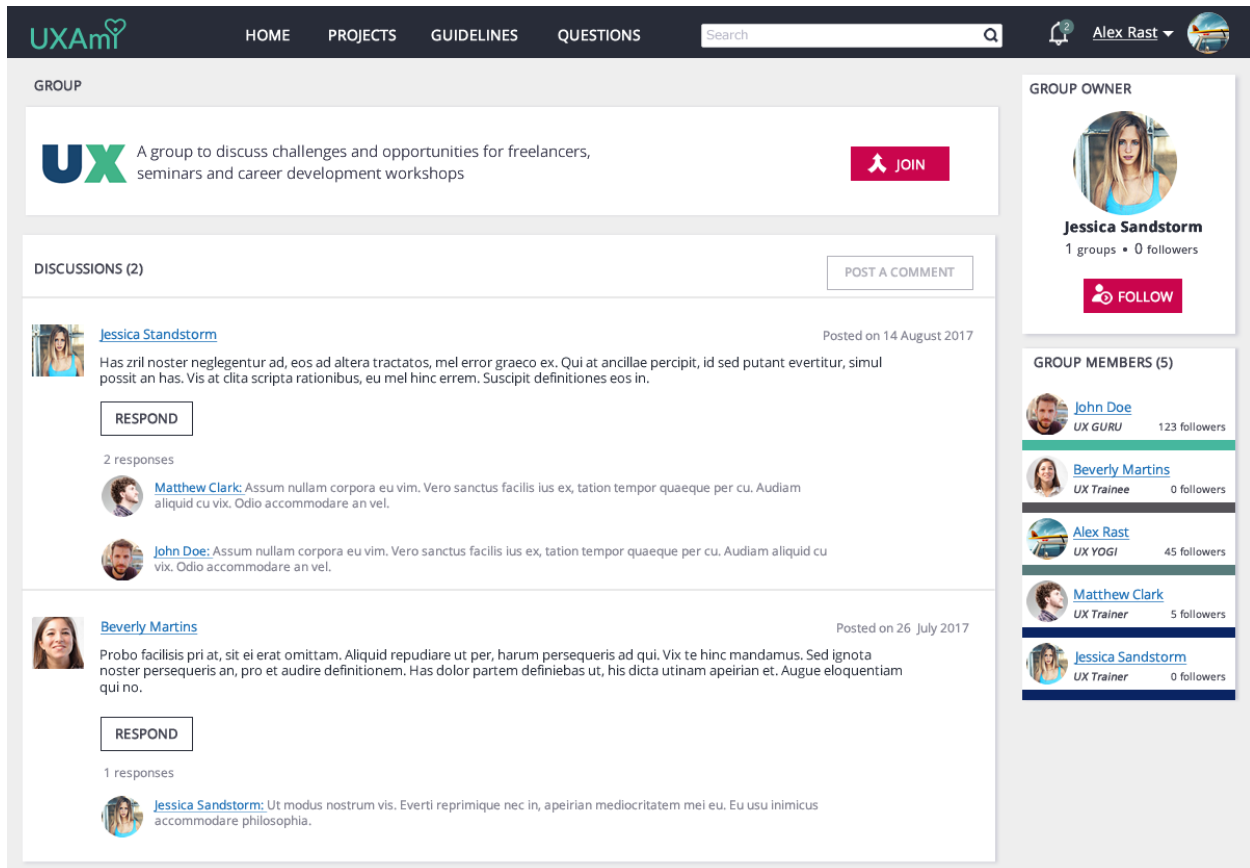
Figure 96. UXAmI OC: Group page

The profile page of a user at the third membership level (Figure 97) follows the silver green theme and provides access to the additional privileges available at the users of this level, namely access to the electronic version of the framework, as well the options of moderating guidelines and mentoring projects. The "Latest Activity" area may also feature at this level guideline moderation activity and mentoring activity. Figure 97 displays the guideline moderation tab, where the user can see all the guidelines that have recently been contributed, and guidelines that have not yet been approved, in a timeline format. If

the user decides to undertake the review of a guideline, they have to indicate so by pressing the "REVIEW" button. The review process has already been discussed in Section 5.4.3 and is illustrated in Figure 86.

Figure 97. UXAmI OC: User profile page (membership level 3)

Figure 98 illustrates the badges collection. Badges that have been collected are illustrated in colour, with an indication of how many times they have been earned. The remaining badges of the UXAmI OC are displayed in grey, allowing users to view the entire collection and explore how they can earn additional badges. Details for each badge are provided upon placing one's mouse over the badge.

Figure 98. UXAmI OC: Badges collection

The profile page of a user at the highest membership level features the teal colour employed by the UXAmI OC logo (Figure 99), and acts as a gateway to the privileges provided at this level, namely access to the UXAmI Observer tool, and community moderation activity. Figure 99 displays the mentoring tab of a user at the fourth membership level, where the projects that one mentors are displayed in the timeline format adopted throughout the UXAmI OC, facilitating direct communication with the contributors of the project that is mentored via messages.

Figure 99. UXAmI OC: User profile page (membership level 4)

5.4.5 CONCLUSIONS AND FUTURE DEVELOPMENTS

In conclusion, the UXAmI Online Community aims at becoming a knowledge resource for UX practitioners, while at the same time it can act as one's personal repository. In the context of designing the OC, this section has analysed the factors that motivate user participation in such communities and the motivators of knowledge contributions, given that an important denominator of the success of a community is the high quality and adequate quantity of content.

Furthermore, with the aim to promote content contribution and active user participation, this section has explored state of the art approaches and best practices for rewarding user contributions. Based on the

above, a reward scheme for the UXAmI OC has been designed, serving a threefold purpose: (i) to motivate users towards knowledge sharing (i.e., contribute evaluation projects and guidelines) (ii) to encourage active engagement, not only through content contribution but also through participating in the community and rating content, keeping thus the community active and providing at the same time a quality-control management by the community itself; and (iii) to foster high quality contributions, laying the foundations for a sustainable community.

In addition, this section has discussed how the UXAmI OC anticipates and addresses the main challenges related to the crowdsourcing approach and tag-based classification adopted, since the guidelines that are contributed by the OC members and approved by the moderators act as input to one of the framework's tools, namely the UXAmI Inspector (Section 5.3). Finally, high-fidelity mock-ups illustrating the basic concepts of the UXAmI OC, the implementation of the reward scheme, and the content moderation approaches have been provided.

Future work includes the complete implementation and deployment of the proposed community. During the further implementation of the UXAmI OC, several issues need to be investigated, such as the algorithm for the recommender system, advanced techniques for hybrid combination of folksonomies and taxonomies (see Section 5.3.2 for the related discussion), as well as the calculation of guideline sets' similarity score, through applying pattern-matching algorithms for strings.

5.5   PUTTING IT ALL TOGETHER – THE AMI RESIDENCE USE CASE

This section describes a use case of an AmI environment, namely an AmI residence, indicating how the UXAmI evaluation framework and the relevant tools can be used. The Ambient Intelligence residence consists of several interactive systems embedded in everyday objects. All the smart devices constitute part of the AmI home ecosystem, including for instance smartphones, tablets, laptops, PCs, televisions, mirrors, smart windows, art displays, and kitchen counters. The main goal of this AmI environment is to support its inhabitants in daily living activities, proactively recognising and responding to their needs, while learning from their behaviour.

The following sections describe the target users and the systems (Section 5.5.1), as well as the evaluation process with the UXAmI framework through a usage scenario (Section 5.5.2).

5.5.1   TARGET USERS AND SYSTEM DESCRIPTION

The Ambient Intelligence residence addresses any potential target user, including people who may live alone, couples, or families, with ages ranging from infants to seniors. Users are expected to be proficient

in using the system after some time. In general, the AmI environment supports various interaction modalities (touch, hand gestures, head and body gestures, eye gestures, and voice) and exhibits high adaptation according to each user and context of use. Furthermore, it provides output through three modalities, namely text, images and video, as well as audio. Since an Ambient Intelligence residence supports the daily activities of its inhabitants, a detailed description of the functionality and possible interactions and adaptations is a highly intricate task and beyond the scope of this thesis. However, in order to get a glimpse into such an AmI environment and better understand the evaluation process that will be applied, short descriptions follow, highlighting specific attributes and functionalities of the environment.

- **Alarms and notifications.** The AmI residence supports individual alarms and notifications for each resident. For instance, alarm clocks are set individually according to each user's agenda for the day. Then each resident is reminded the tasks of the day, while they can receive messages and notifications from other users (e.g., other house residents, co-workers, family, and friends). The smart agenda application is available through a wide range of devices, including smartphones, tablets, smart mirrors and televisions.

- **Environment control.** An important characteristic of the AmI residence is the advanced environment control facilities, including security of the residence entrances, house temperature and lighting, and devices' control. The house entrance is controlled through biometrics, allowing access only to the residents or authorised visitors (e.g., a caregiver for an elderly). Furthermore, all the entrances (doors and windows) can be controlled through any of the home devices or even remotely. On the other hand, house temperature is automatically adjusted according to the residents' preferences and the time of the day. For instance, when there is no one at home a minimum low temperature is maintained during winter. In the morning, a while before the first resident is awake, a temperature commonly accepted by all the residents is set, as the smart home prepares for waking up its inhabitants. Likewise, lighting levels are automatically adjusted according to the residents and their current activities, or any other lighting scenarios that have been defined (e.g., when the family is away for a long time, lights turn on and off sporadically, indicating activity for security reasons). Both temperature and lighting can be controlled through any smart device of the house, as well as remotely.

- **Devices control.** Authorised residents can control home devices on site or remotely, according to their profile. For instance, a child is not allowed to control the kitchen or the washing machine. Examples of devices that support smart controlling include television sets, the kitchen, the refrigerator, washing machines, the garden watering system, etc.

- **Communication.** An important everyday activity that is supported by the AmI residence is that of communication. The environment supports voice and video calls, as well as exchange of text and multimedia messages among the home residents, or with other remote counterparts (e.g., family and friends). Calls and messages are available in any of the smart devices of the residence, including smartphones, tablets, laptops, PCs, mirrors and televisions.

- **Daily activities.** The AmI environment aims to facilitate daily activities and needs of its inhabitants, such as cooking and shopping. For instance, the kitchen proposes meals for a healthy and balanced diet, while when cooking it assists the resident by providing a step-by-step recipe. The recipes proposed are based on ingredients that can be found in the smart refrigerator and the smart shelves, while when an ingredient runs out it is automatically added to the shopping list.

- **AAL.** The AmI residence also features Ambient Assisted Living facilities, including for instance health status monitoring, medication monitoring, and psychophysiological monitoring indicating emotional stress and mental workload. When required, the environment suggests activities to the inhabitants, including receiving medication, cooking a specific meal, communicating with a caregiver or a friend, or engaging in a leisure activity.

5.5.2 EVALUATION

The design and evaluation of the AmI residence is a challenging and complex task, to which two UX experts have been engaged. Alessandra and Antonella are both users of the UXAmI framework and its tools and have used them in the past for smaller projects. Alessandra logs in to the UX engineers professional platform (UXAmI OC) and initiates a shared project with Antonella. From now on, they will both be able to have a common view of the project and collaborate, using the OC. Given the nature of the project, almost all the attributes proposed by the UXAmI framework will need to be tested, while several rounds of expert-based reviews and user testing will be required.

As the system evolves Antonella coordinates the expert-based reviews in several time instances of the environment's development lifecycle. First, using the UXAmI Inspector tool she collects and reviews guidelines relevant to the individual systems comprising the environment and employs them in the context of the reviews to be carried out. More specifically, the tool suggests four guideline sets to be taken into account, namely the Heuristic Evaluation Guidelines, the Natural User Interaction Guidelines, Guidelines for the evaluation of Ambient Displays, as well as AAL Heuristic Guidelines. Antonella includes all four sets in her inspection project and proceeds with the review process. The results are given to the developers' team to eliminate the problems located. Consulting the UXAmI Framework, Antonella also organises a cognitive walkthrough evaluation to assess the overall system's learnability. To this end, she

uses the indicative scenarios that have been developed to guide her test procedures, aiming to include all the potential features and characteristics of the AmI residence. The scenarios have been uploaded in the UXAmI OC and Antonella can further enrich them in collaboration with her teammate. Additional expert-based reviews that Antonella organises refer to:

- the embedment of the system
- the validity of system interpretations of acquired sensor data before deciding an adaptation
- the appropriateness of the adaptations of interaction modalities, system output and content for the different usage scenarios
- the appropriateness of recommendations under different usages, as well as if they are adequately explained and if users can express and revise their preferences
- consistency of cross-platform tasks in the various devices, as well as synchronization of the content and the available actions
- the environment's social behaviour in terms of conforming to social etiquette
- the appropriateness of the environment's responses to implicit interactions
- the aesthetics of the individual systems
- the physical accessibility of the individual systems, by exploring if the systems and their controls have appropriate size and positioning and verifying that no ergonomic guidelines are violated
- the physical accessibility of the entire AmI environment for all the potential target users
- the electronic accessibility of each individual system, by taking into account electronic accessibility guidelines
- issues regarding safety and privacy (e.g., user control over the data and the dissemination of information, customisation of the level of control of the AmI environment, safety of the operator, the environment, and the general public).

Having noticed that several of the above attributes could employ their own heuristics checklist (e.g., safety and privacy), Antonella is currently in the process of preparing a relevant publication for a scientific journal. As soon as it will become accepted, she plans to upload the guidelines in UXAmI OC, so that the community members can benefit from this knowledge, as the guidelines will therefore be suggested by UXAmI Inspector for all AmI systems and applications.

Alessandra, on the other hand, is responsible for the user testing experiments. Given the complexity of the project, several user tests have been planned, each targeting specific functionalities of the system. To this end, a small number of users participates in each test, while all tests are administered through the UXAmI Observer tool. Before each test, she creates an experiment in the tool, and adds participants and

tasks. She runs a pilot test, and among others she uses it to identify when each task successfully ends. While Alessandra coordinates each evaluation experiment, her teammate Antonella observes the experiment through the Observer's live view and annotates points of interest. After each participant session, Alessandra reviews the session, makes her own annotations, confirms or rejects the tool's inferences regarding input errors and adaptation rejections and keeps notes in an evaluation report. Once the experiment is complete, Alessandra views the insights offered by the tool aiming to draw conclusions regarding the functionalities that have been used for each system, users' awareness of the interaction vocabulary, distractions caused to users, appropriateness of adaptation decisions, effectiveness, efficiency and learnability, as well as users' emotions. Employing the UXAmI Observer's Insights View, she creates a detailed report for each system, identifying the major UX problems that she has found.

Once all systems are improved and a final prototype of the entire AmI system has been created, a more extended experiment in a simulated AmI residence takes place, involving a larger number of users participating in multi-user set ups. Following the UXAmI framework, Alessandra also creates three questionnaires, one to be handed before the experiment, one shortly after each user session, and one after the last interaction of a participant with the environment. Moreover, she plans to employ all the automated and semi-automated measurements advocated by the framework and instantiated in UXAmI Observer. During the final experiment preparation phase Alessandra creates a new experiment in UXAmI Observer, and adds participants. This long-term experiment will not feature specific tasks, therefore she does not define any in the experiment screen. She reviews the results on a per session and aggregated basis, while she employs the various insights views (overall, per system, per application) to enhance her understanding of which exact systems and behaviours cause problems to users. In summary, she reviews the following parameters per user, per application, per system, as well as for the entire AmI environment and all the involved users:

- Functionality that has been used
- Percentage of input modalities used
- Erroneous user inputs
- Percentage of erroneous user inputs per input modality
- Deviations from the primary task
- Adaptations introduced
- Adaptations overridden by the users
- Behaviour after an adaptation has been introduced
- Behaviour after switching device during a specific activity
- Appropriateness of system recommendations

- Effectiveness (e.g., number of input errors and interaction errors)
- Efficiency (e.g., help requests and time spent on errors)
- User errors (input and interaction) and help requests over time
- Users' implicit interactions and system behaviour regarding these
- Users' detected emotion POIs
- Usability problems due to concurrent usage by more than one users
- Any electronic or physical accessibility problems
- Detailed usage statistics (which systems and applications are used and how often)

Furthermore, she acquires users' responses to questionnaires regarding:

- System embedment
- Appropriateness of recommendations
- User satisfaction
- Aesthetics and Fun
- Perceived ease of use and perceived usefulness
- Expected outcomes
- Trust in the system
- Other system attributes, such as trialability, relative advantage, end-user support, observability and image
- Users' personal attributes, such as age, gender, self-efficacy etc.

Following each session, she carries out a short interview with the participants addressing among others the points suggested by UXAmI Framework, such as embedment, appropriateness of recommendations and accessibility. Being sure, that through the tools she will acquire plenty of information, she does not plan for extensive questionnaires and interviews, avoiding to overburden the users or resorting to users' subjective views for all the UX aspects. Reviewing all the results and insights available through the UXAmI Observer tool, Alessandra is able to easily identify issues that need to be changed and to prioritize them.

6   SUMMARY, CONCLUSIONS AND FUTURE WORK

This section concludes the current thesis, by providing a summary of the main motivations and of the work that has been carried out (Section 6.1), discussing the contributions (Section 6.2) and presenting directions for future work (Section 6.3).

6.1   SUMMARY

Ambient Intelligence is an emerging field of research and development, constituting a new technological paradigm, integrating sensors and interconnected devices, equipped with computer vision and reasoning mechanisms, and focused on serving in the best possible manner the needs of its inhabitants, through recognizing their needs and (proactively) responding in an unobtrusive way, also exhibiting adaptive behaviour if required. On the other hand, determining what would make a technology acceptable by users was widely recognized as a significant field of research since the seventies. Although the methods employed and the models developed have evolved ever since, acceptance, usability, and qualitative user experience continue to remain at the forefront of HCI research.

The paradigm shift from the typical desktop computer used in an organizational context to interconnected, embedded, portable devices that are used in multiple contexts by more than one users concurrently, has brought about new challenges and increased the complexity of evaluating UX and usability, yet it promises new opportunities as well. One of the challenges faced is the immense number of parameters one would have to take into account in order to explore the multiple facets of UX in such technologically advanced environments. The majority of existing approaches to evaluation mainly resort to estimating UX based on user-reported values, combining them in some cases with observers' notes for a user's interaction. As the number of factors to be explored increases, such approaches become unrealistic and obsolete.

In this context, although several evaluation frameworks have been proposed for usability, UX, adaptive systems, and UbiComp systems, and despite the fact that the notion of AmI has been established since 2001, there is a total lack of frameworks for the evaluation in AmI environments. To this end, this thesis has proposed an inclusive conceptual and methodological framework for the evaluation of user experience in Ambient Intelligence environments. The proposed UXAmI framework takes into account the attributes of an AmI environment, as well as traditional and modern models and evaluation approaches, including technology acceptance models, usability and user experience evaluation methods, methods for the evaluation of adaptive systems, ubiquitous computing systems and Ambient Intelligence systems. UXAmI, taking advantage of the infrastructure that is inherent in AmI environments, alleviates the need

for exhaustive manual recording of observations and lengthy questionnaires addressed to users and advocates objective measures automatically acquired. Furthermore, the framework is accompanied by three tools instantiating the suggested approach and facilitating user-based experiments and expert-based reviews, as well as knowledge exchange through a professional social network for UX experts.

6.2 CONTRIBUTIONS

In the context of the current thesis a systematic review of evaluation practices and methods in the fields of user acceptance, usability, user experience, adaptive systems, ubiquitous computing systems and AmI environments has been carried out. A review of 43 technology acceptance models, and a classification of the involved parameters according to their point of reference, objectiveness, assessment method employed and context has been conducted. Moreover, 42 evaluation frameworks suggested in literature in the aforementioned domains have been reviewed, resulting in:

- A classification of parameters explored by the frameworks towards assessing usability, resulting in a list of 97 attributes organized under 20 main categories.
- A classification of parameters explored by User Experience evaluation frameworks, resulting in 79 attributes organized under 11 categories.
- A classification of parameters suggested by UbiComp evaluation frameworks, resulting in 70 attributes organized under 17 categories.
- A classification of all the evaluation frameworks, according to their type, usage of concrete metrics, evaluation field, and context.

Based on the above, an inclusive conceptual and methodological framework for the evaluation of UX in AmI environments has been proposed, featuring 103 metrics organized in subcategories, which are in turn classified under seven fundamental attributes of AmI environments, namely intuitiveness, unobtrusiveness, adaptivity, usability, appeal and emotions, safety and privacy, and acceptance. UXAmI has proposed 39 novel metrics, mainly motivated by the need to complement expert-based reviews, user-reported metrics, and observers' remarks with objective measurable behaviours. The novel metrics introduced refer to the categories of awareness of application capabilities and of the interaction vocabulary, distractions, appropriateness and impact of adaptations, appropriateness of recommendations, cross-platform usability, multi-user usability, implicit interactions and usage of the AmI environment. An important contribution of the proposed framework is that it minimizes the effort required by the observer in recording detailed notes for users' actions and behaviours, while it increases the objectivity of the metrics through the automated measurements it suggests, a contribution which was acknowledged in the evaluation of the framework that has been carried out with six experts.

The three tools accompanying the framework contribute each towards different needs of the evaluation in AmI environments. UXAmI Observer facilitates user-based experiments by acquiring data from the AmI environment, making inferences and calculating automated metrics regarding users' and the environment's behaviour, and provides valuable different views, as well as statistical and usage information. In a nutshell, the tool supports: detailed analysis of a participant's interaction and aggregated insights from all the participants; distinction in user errors to input and interaction errors; automatically acquired measurements regarding adaptations applied and adaptation rejections, user input errors, detected user emotions and implicit interactions; manual annotation of points of interest and control over the automatically calculated measurements; synchronization of all the data with session videos; powerful visualizations and statistics. Also, an important contribution of the tool is that it does not require instrumentation from the application developers, in contrast to the majority of previous efforts in the field of usability evaluation automation.

UXAmI Inspector supports expert-based evaluations and addresses the challenges associated with these approaches, which are rooted to the increasing volume of guidelines. In summary, the tool facilitates evaluators in setting up an inspection project through a step-by-step wizard for the definition of evaluation targets, automatically retrieves guidelines relevant to the evaluation targets, and supports the inspection process. The main advantage of the tools is that it substantially reduces the time and effort needed to prepare and carry out an expert-based evaluation, while its main contribution is that it is the first tool to automatically suggest high level evaluation guidelines which are appropriate for the evaluation target, without the need for explicit input by the evaluator.

Finally, the proposed social network for UX engineers (UXAmI OC) aims to become a medium for UX professionals to exchange knowledge and promote collaboration and a personal repository, while it interoperates with UXAmI Inspector and is used as a source of guidelines. With the aim to ensure that the content contributed is of high quality, and also that users actively participate in the community, addressing as a result issues related to crowdsourcing approaches, an adaptive reward scheme has been designed for the community. The proposed reward scheme features a system of points and membership levels accompanied by specific privileges for each membership level, employs reputation mechanisms, and embraces rating content by endorsing its helpfulness. The rewarding mechanism is adaptive in awarding points for timely answers, number of followers, and deducting points for violating the terms of the OC. UXAmI Observer and UXAmI Inspector are offered as rewards to higher membership level users, while at the same time users of higher levels act as moderators of the content and the community, as well potential mentors to other community members, a contribution which further enhances their visibility and reputation in the community.

6.3 FUTURE WORK

Future endeavours include a variety of tasks, ranging from addressing minor usability problems to conceptually expanding the framework, including new methodologies, and employing the framework and its tools in long-term actual experiments.

First, in the short term, all the usability issues that have been identified through the heuristic evaluations will be addressed. Furthermore, an electronic version of the framework is planned to be developed, accompanied by specific suggestions of ready to use standardized questionnaires for user-based evaluations. More importantly however, additional functionality will be developed for both UXAmI Observer and Inspector to better address the practical needs of evaluators. In particular, UXAmI Observer will be further enhanced with functionality to analyse standard questionnaires and combine the results with existing insights, as well as statistical analysis capabilities, supporting the extraction of tables and figures and performing comparisons among the various user groups involved in an experiment. UXAmI Inspector will be extended to support concurrent review by multiple inspectors, while both tools will incorporate facilities for creating evaluation reports with pre-filled data. Finally, all tools will become interoperable, allowing the evaluator to select which metrics of the framework will be used in each experiment, to record the various evaluation iterations under one evaluation project and to identify and prioritize problems that need to be addressed, effectively communicating them to the development team.

The evaluation of UXAmI framework has confirmed its important contribution towards UX evaluation in AmI environments, it has also revealed however the need for creating add-ons that can be considered by evaluators according to their current evaluation project and target domain. Future work in the context of the framework will explore the addition of specific metrics targeted at the most common interaction modalities employed in AmI environments (e.g., gestures, speech), and metrics targeted at specific user groups (e.g., children, elderly, users with disabilities) or contexts (e.g., education, health). Future mid-term activities will also explore the inclusion of evaluation goals, formulated around hypotheses that will be explored through specific UXAmI metrics with the support of UXAmI Observer.

Another line of research is to explore whether and how UXAmI can itself produce rules, based on how it "senses" user interaction by combining the data acquired. These rules could be fed back to the AmI environment's reasoning agents to be considered in the context of further adaptations. To that end, two main directions need to be explored, namely how can the evaluator create rules to be considered by the AmI environment based on their insights and conclusions from user-based experiments, as well as how UXAmI can dynamically create such rules based on the automatically acquired data and their potential combinations, revealing specific behavioural patterns of the AmI environment inhabitants.

Finally, long-term pursuits include studying scalability issues of the framework will be studied, exploring how it can be adapted to serve other contexts beyond AmI environments, such as Internet of Things or smart cities. Furthermore, in the long term, the framework and its tools are planned to be employed in actual evaluations of AmI spaces, systems and applications, in order to acquire feedback and insights to further improve them. Last, another important concern which constitutes the focus of future work is to explore how evaluation through simulations can be integrated in the framework and to develop the corresponding tools. Simulations could be used as part of the iterative evaluation process to test system behaviour in various complex situations, when it is too early to involve users, or for reproducing rare or long-term usage conditions of an AmI system or environment that may be difficult or improbable to achieve with real users. Simulations can also be used as a means towards benchmarking, allowing for quantifiable measurements and comparisons of AmI systems and applications.

REFERENCES

Aarts, E. H., & de Ruyter, B. E. (2009). New research perspectives on Ambient Intelligence. JAISE, 1(1), 5-14.

Aarts, E., & Wichert, R. (2009). Ambient intelligence (pp. 244-249). Springer Berlin Heidelberg.

Abascal, J., Arrue, M., Fajardo, I., Garay, N., & Tomás, J. (2004). The use of guidelines to automatically verify Web accessibility. Universal Access in the Information Society, 3(1), 71-79.

Abowd, G. D. (2012, September). What next, ubicomp?: celebrating an intellectual disappearing act. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (pp. 31-40). ACM.

Adikari, S., McDonald, C., & Campbell, J. (2011). A design science framework for designing and assessing user experience. Human-computer interaction. Design and development approaches, 25-34.

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. MIS quarterly, 665-694.

Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behaviour.

Al-Azzawi, A. (2014). Theories of Experience. Experience with Technology: Dynamics of User Experience with Mobile Media Devices, 7-29.

Alexander, J., Cockburn, A., & Lobb, R. (2008). AppMonitor: A tool for recording user actions in unmodified Windows applications. Behavior Research Methods, 40(2), 413-421.

Alfred, V. (2014). Algorithms for finding patterns in strings. Algorithms and Complexity, 1, 255-300.

Allam, A. H., & Dahlan, H. M. (2013). User experience: Challenges and opportunities. Journal of Information Systems Research and Innovation, 3, 28-36.

Alspaugh, S., Ganapathi, A., Hearst, M. A., & Katz, R. (2014). Better logging to improve interactive data analysis tools. In KDD Workshop on Interactive Data Exploration and Analytics (IDEA).

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007, April). What happened to remote usability testing?: an empirical study of three methods. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 1405-1414). ACM.

Assila, A., de Oliveira, K. M., & Ezzedine, H. (2016, June). An environment for integrating subjective and objective usability findings based on measures. In Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on (pp. 1-12). IEEE.

Au, F. T., Baker, S., Warren, I., & Dobbie, G. (2008, January). Automated usability testing framework. In Proceedings of the ninth conference on Australasian user interface-Volume 76 (pp. 55-64). Australian Computer Society, Inc.

Augusto, J. C. (2007). Ambient intelligence: the confluence of ubiquitous/pervasive computing and artificial intelligence. Intelligent Computing Everywhere, 213-234.

Augusto, J. C., & McCullagh, P. (2007). Ambient intelligence: Concepts and applications. Computer Science and Information Systems, 4(1), 1-27.

Augusto, J. C., Nakashima, H., & Aghajan, H. (2010). Ambient intelligence and smart environments: A state of the art. In Handbook of ambient intelligence and smart environments (pp. 3-31). Springer US.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Prentice-Hall, Inc.

Bandyopadhyay, K., & Fraccastoro, K. A. (2007). The effect of culture on user acceptance of information technology. Communications of the Association for Information Systems, 19(1), 23.

Bao, L., Li, J., Xing, Z., Wang, X., Xia, X., & Zhou, B. (2017). Extracting and analyzing time-series HCI data from screen-captured task videos. *Empirical Software Engineering*, *22*(1), 134-174.

Baptista, G., & Oliveira, T. (2015). Understanding m banking: The unified theory of acceptance and use of technology combined with cultural moderators. Computers in Human Behavior, 50, 418-430.

Bardram, J., & Friday, A. (2010). Ubiquitous computing systems. Ubiquitous Computing Fundamentals, 37-94, CRC Press.

Bargas-Avila, J. A., & Hornbæk, K. (2011, May). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2689-2698). ACM.

Bar-Ilan, J., Shoham, S., Idan, A., Miller, Y., & Shachak, A. (2008). Structured versus unstructured tagging: a case study. Online Information Review, 32(5), 635-647.

Barton, J. J., & Vijayaraghavan, V. (2003). UBIWISE, a simulator for ubiquitous computing systems design. Hewlett-Packard Laboratories Palo Alto, â AI HPL-2003-93.

Bateman, S., Gutwin, C., Osgood, N., & McCalla, G. (2009, July). Interactive usability instrumentation. In Proceedings of the 1st ACM SIGCHI symposium on Engineering interactive computing systems (pp. 45-54). ACM.

Beirekdar, A., Vanderdonckt, J., & Noirhomme-Fraiture, M. (2003). KWARESMI–Knowledge-based Web Automated Evaluation Tool with Reconfigurable Guidelines Optimization. In Proc. Of 2nd International Conf. on Universal Access in Human-Computer Interaction UAHCI'2003.

Belous, Y. (2014). Why do we share information online? A study of motivations. International Journal of Arts & Sciences, 7(4), 383.

Benlian, A., & Hess, T. (2011). The signaling role of IT features in influencing trust and participation in online communities. International Journal of Electronic Commerce, 15(4), 7-56.

Berlanga, A. J., Bitter-Rijpkema, M., Brouns, F., Sloep, P. B., & Fetter, S. (2011). Personal profiles: enhancing social interaction in learning networks. International Journal of Web Based Communities, 7(1), 66-82.

Bernard, F. T., Marfisi-Schottman, I., & Habieb-Mammar, H. (2009). AnAmeter: The first steps to evaluating adaptation. arXiv preprint arXiv:0911.3644.

Bevan, N. (2009a). Extending quality in use to provide a framework for usability measurement. In Human Centered Design (pp. 13-22). Springer Berlin Heidelberg.

Bevan, N. (2009b). What is the difference between the purpose of usability and user experience evaluation methods. In Proceedings of the Workshop UXEM (Vol. 9).

Bevan, N., Carter, J., & Harker, S. (2015). ISO 9241-11 Revised: What Have We Learnt About Usability Since 1998?. In Human-Computer Interaction: Design and Evaluation (pp. 143-151). Springer International Publishing.

Bevan, N., & Macleod, M. (1994). Usability measurement in context. Behaviour & information technology, 13(1-2), 132-145.

Bias, R.G. (1994) Pluralistic usability walkthrough: coordinated empathies. In J. Nielsen & R.L. Mack (Eds.), Usability Inspection Methods (pp. 63-76). New York, NY: Wiley and Sons, Inc.

Bibri, S. E. (2015). The human face of Ambient Intelligence: Cognitive, emotional, affective, behavioral and conversational aspects (Vol. 9). Springer.

Bieberstein, N. (2006). Service-oriented architecture compass: business value, planning, and enterprise roadmap. FT Press.

Blythe, M., Robinson, J., & Frohlich, D. (2008, October). Interaction design and the critics: what to make of the weegie. In Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges (pp. 53-62). ACM.

Bouzit, S., Calvary, G., Chêne, D., & Vanderdonckt, J. (2016). Automated Evaluation of Menu by Guidelines Review. In RoCHI (pp. 11-21).

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. Convergence, 14(1), 75-90.

Bretzke, H., & Vassileva, J. (2003). Motivating cooperation on peer to peer networks. User Modeling 2003, 148-148.

Brown, S. A., & Venkatesh, V. (2005). Model of adoption of technology in households: A baseline model test and extension incorporating household life cycle. MIS quarterly, 399-426.

Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2004). Layered evaluation of adaptive learning systems. International Journal of Continuing Engineering Education and Life Long Learning, 14(4-5), 402-421.

Bueno, S., & Salmeron, J. L. (2008). TAM-based success modeling in ERP. Interacting with Computers, 20(6), 515-523.

Bulmer, D., & DiMauro, V. (2009). The New Symbiosis of Professional Networks: Social Media's Impact on Business and Decision-making. Society for New Communications Research.

Byrne, M. D., Wood, S. D., Foley, J. D., Kieras, D. E., & Sukaviriya, P. N. (1994, April). Automating interface evaluation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 232-237). ACM.

Caplan, S. (1990). Using focus group methodology for ergonomic design. Ergonomics, 33(5), 527-533.

Carlos Roca, J., José García, J., & José de la Vega, J. (2009). The importance of perceived trust, security and privacy in online trading systems. Information Management & Computer Security, 17(2), 96-113.

Carter, S., & Mankoff, J. (2005, April). When participants do the capturing: the role of media in diary studies. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 899-908). ACM.

Carter, S., Mankoff, J., & Heer, J. (2007, April). Momento: support for situated ubicomp experimentation. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 125-134). ACM.

Carvalho, R. M., Andrade, R. M., & Oliveira, K. M. (2015, August). Using the GQM method to evaluate calmness in ubiquitous applications. In International Conference on Distributed, Ambient, and Pervasive Interactions (pp. 13-24). Springer, Cham.

Carvalho, R. M., de Castro Andrade, R. M., de Oliveira, K. M., de Sousa Santos, I., & Bezerra, C. I. M. (2017). Quality characteristics and measures for human–computer interaction evaluation in ubiquitous systems. Software Quality Journal, 25(3), 743-795.

Casacuberta, J., Sainz, F., & Madrid, J. (2012, December). Evaluation of an Inclusive Smart Home Technology System. In IWAAL (pp. 316-319).

Castañeda, J. A., Muñoz-Leiva, F., & Luque, T. (2007). Web Acceptance Model (WAM): Moderating effects of user experience. Information & management, 44(4), 384-396.

Chai, S., & Kim, M. (2012). A socio-technical approach to knowledge contribution behavior: An empirical investigation of social networking sites users. International Journal of Information Management, 32(2), 118-126.

Chang, E., & Dillon, T. S. (1997). Automated usability testing. In Human-Computer Interaction INTERACT'97 (pp. 77-84). Springer US.

Chang, H. H., & Chuang, S. S. (2011). Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. Information & management, 48(1), 9-18.

Chang, C. M., Hsu, M. H., Hsu, C. S., & Cheng, H. L. (2014). Examining the role of perceived value in virtual communities continuance: its antecedents and the influence of experience. Behaviour & Information Technology, 33(5), 502-521.

Charfi, S., Trabelsi, A., Ezzedine, H., & Kolski, C. (2011, May). Towards ergonomic guidelines integration within graphical interface controls for the evaluation of the interactive system Application to a transport

network Information Support System. In Logistics (LOGISTIQUA), 2011 4th International Conference on (pp. 76-82). IEEE.

Charfi, S., Trabelsi, A., Ezzedine, H., & Kolski, C. (2013). A User-Oriented Test environment based on User-Interface evaluation graphical controls. IFAC Proceedings Volumes, 46(15), 497-504.

Chen, I. Y. (2007). The factors influencing members' continuance intentions in professional virtual communities—a longitudinal study. Journal of Information Science, 33(4), 451-467.

Chen, T. Y., Chen, C. H., Shih, C. S., & Liu, J. W. S. (2008, October). A simulation environment for development and evaluation of smart devices for the elderly. In Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on (pp. 2285-2290). IEEE.

Chen, C. J., & Hung, S. W. (2010). To give or to receive? Factors influencing members' knowledge sharing and community promotion in professional virtual communities. Information & management, 47(4), 226-236.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2014). How community feedback shapes user behavior. arXiv preprint arXiv:1405.1429.

Cheng, R., & Vassileva, J. (2005, January). User motivation and persuasion strategy for peer-to-peer communities. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 193a-193a). IEEE.

Cheng, R., & Vassileva, J. (2006). Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. User Modeling and User-Adapted Interaction, 16(3), 321-348.

Cheung, C. M., Lee, M. K., & Lee, Z. W. (2013). Understanding the continuance intention of knowledge sharing in online communities of practice through the post-knowledge-sharing evaluation processes. Journal of the Association for Information Science and Technology, 64(7), 1357-1374.

Chiang, H. S. (2013). Continuous usage of social networking sites: The effect of innovation and gratification attributes. Online Information Review, 37(6), 851-871.

Chin, D. (2001). Empirical Evaluation of User Models and User-Adapted Systems. In User Modeling and User-Adapted Interaction 11, 1-2 (March 2001), 181-194.

Chiu, C. M., Hsu, M. H., & Wang, E. T. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. Decision support systems, 42(3), 1872-1888.

Choi, J., Lee, H. J., Sajjad, F., & Lee, H. (2014). The influence of national culture on the attitude towards mobile recommender systems. Technological Forecasting and Social Change, 86, 65-79.

Christensen, L., & Frøkjær, E. (2010, October). Distributed usability evaluation: enabling large-scale usability evaluation with user-controlled instrumentation. In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (pp. 118-127). ACM.

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. MIS quarterly, 189-211.

Compeau, D., Higgins, C. A., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. MIS quarterly, 145-158.

Condori-Fernandez, N. (2017, May). HAPPYNESS: an emotion-aware QoS assurance framework for enhancing user experience. In Proceedings of the 39th International Conference on Software Engineering Companion (pp. 235-237). IEEE Press.

Connelly, K. (2007, September). On developing a technology acceptance model for pervasive computing. In 9th International Conference on Ubiquitous Computing (UBICOMP)-Workshop of Ubiquitous System Evaluation (USE), Springer, Innsbruck, Austria (p. 520).

Consolvo, S., & Walker, M. (2003). Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Computing, (2), 24-31.

Consolvo, S., Arnstein, L., & Franza, B. R. (2002). User study techniques in the design and evaluation of a ubicomp environment. In UbiComp 2002: Ubiquitous Computing (pp. 73-90). Springer Berlin Heidelberg.

Consolvo, S., Harrison, B., Smith, I., Chen, M. Y., Everitt, K., Froehlich, J., & Landay, J. A. (2007). Conducting in situ evaluations for and with ubiquitous computing technologies. International Journal of Human-Computer Interaction, 22(1-2), 103-118.

Constantine, L. L., & Lockwood, L. A. (1999). Software for use: a practical guide to the models and methods of usage-centered design. Pearson Education.

Cook, D., & Das, S. K. (2004). Smart environments: Technology, protocols and applications (Vol. 43). John Wiley & Sons.

Cook, D. J., Augusto, J. C., & Jakkula, V. R. (2009). Ambient intelligence : Technologies, applications, and opportunities. Pervasive and Mobile Computing, 5(4), 277-298.

Costello, B., & Edmonds, E. (2007, August). A study in play, pleasure and interaction design. In Proceedings of the 2007 conference on Designing pleasurable products and interfaces (pp. 76-91). ACM.

Cota, C. X. N., Díaz, A. I. M., & Duque, M. Á. R. (2014, October). Developing a framework to evaluate usability in m-learning systems: Mapping study and proposal. In Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 357-364). ACM.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., ... & Ravaja, N. (2016). The psychophysiology primer: a guide to methods and a broad review with a focus on human–computer interaction. Foundations and Trends® in Human–Computer Interaction, 9(3-4), 151-308.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the Experience-Sampling Method. The Journal of nervous and mental disease, 175(9), 526-536.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993, February). Wizard of Oz studies: why and how. In Proceedings of the 1st international conference on Intelligent user interfaces (pp. 193-200). ACM.

Daniels, J., Fels, S., Kushniruk, A., Lim, J., & Ansermino, J. M. (2007). A framework for evaluating usability of clinical monitoring technology. Journal of clinical monitoring and computing, 21(5), 323-330.

Davis, F. D. (1985). A technology acceptance model for empirically testing new end-user information systems: theory and results (Doctoral dissertation, Massachusetts Institute of Technology, Sloan School of Management). Retrieved from: http://hdl.handle.net/1721.1/15192

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace1. Journal of applied social psychology, 22(14), 1111-1132.

De Carolis, B., Ferilli, S., & Novielli, N. (2012). Recognizing the User Social Attitude in Multimodal Interaction in Smart Environments. In Ambient Intelligence (pp. 240-255). Springer Berlin Heidelberg.

De Moor, K., Ketyko, I., Joseph, W., Deryckere, T., De Marez, L., Martens, L., & Verleye, G. (2010). Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting. Mobile Networks and Applications, 15(3), 378-391.

De Ruyter, B., & Aarts, E. (2010). Experience research: a methodology for developing human-centered interfaces. In Handbook of Ambient Intelligence and Smart Environments (pp. 1039-1067). Springer US.

De Ruyter, B., Van Loenen, E., & Teeven, V. (2007). User centered research in experiencelab. In Ambient Intelligence (pp. 305-313). Springer Berlin Heidelberg.

Den Uyl, M. J., & Van Kuilenburg, H. (2005, August). The FaceReader: Online facial expression recognition. In Proceedings of Measuring Behavior (Vol. 30, pp. 589-590).

Dervin, B. (1998). Sense-making theory and practice: an overview of user interests in knowledge seeking and use. Journal of knowledge management, 2(2), 36-46.

Desmet, P. (2005). Measuring emotion: Development and application of an instrument to measure emotional responses to products. In Funology (pp. 111-123). Springer Netherlands.

Desmet, P., Overbeeke, K., & Tax, S. (2001). Designing products with added emotional value: Development and application of an approach for research through design. The design journal, 4(1), 32-47.

Dias, M. S., Vilar, E., Sousa, F., Vasconcelos, A., Pinto, F. M., Saldanha, N., & Eloy, S. (2015, August). A living labs approach for usability testing of Ambient Assisted Living technologies. In International Conference of Design, User Experience, and Usability (pp. 167-178). Springer, Cham.

Dingli, A., & Mifsud, J. (2011). Useful: A framework to mainstream web site usability through automated evaluation. International Journal of Human Computer Interaction (IJHCI), 2(1), 10.

Dhouib, A., Trabelsi, A., Kolski, C., & Neji, M. (2016, July). A classification and comparison of usability evaluation methods for interactive adaptive systems. In Human System Interactions (HER), 2016 9th International Conference on (pp. 246-251). IEEE.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. Communications of the ACM, 54(4), 86-96.

Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J., & Burgelman, J. C. (2001). Scenarios for ambient intelligence in 2010 (pp. 3-8). Luxembourg: Office for official publications of the European Communities.

Fang, X., Chan, S., Brzezinski, J., & Xu, S. (2005). Moderating effects of task type on wireless technology acceptance. Journal of Management Information Systems, 22(3), 123-157.

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. Organization Science, 22(5), 1224-1239.

Farzan, R., DiMicco, J. M., Millen, D. R., Dugan, C., Geyer, W., & Brownholtz, E. A. (2008, April). Results from deploying a participation incentive mechanism within the enterprise. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 563-572). ACM.

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-services adoption: a perceived risk facets perspective. International journal of human-computer studies, 59(4), 451-474.

Fernandez, A., Insfran, E., & Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study. Information and Software Technology, 53(8), 789-817.

Fernández, M. Á., Peláez, V., López, G., Carus, J. L., & Lobato, V. (2012, December). Multimodal Interfaces for the Smart Home: Findings in the Process from Architectural Design to User Evaluation. In UCAmI (pp. 173-180).

Ferre, X., Juristo, N., & Moreno, A. M. (2005). Framework for integrating usability practices into the software process. In Product focused software process improvement (pp. 202-215). Springer Berlin Heidelberg.

Fetscherin, M., & Lattemann, C. (2008). User acceptance of virtual worlds. Journal of electronic commerce research, 9(3), 231.

Fields, B., Amaldi, P., Wong, W., & Gill, S. (2007). In Use, In Situ: Extending Field Research Methods. International Journal of Human-Computer Interaction 22 (1-2): 1-6

Findlater, L., & Gajos, K. Z. (2009). Design space and evaluation challenges of adaptive graphical user interfaces. AI Magazine, 30(4), 68.

Følstad, A., Law, E., & Hornbæk, K. (2012, May). Analysis in practical usability evaluation: a survey study. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2127-2136). ACM.

Freiberg, M., & Baumeister, J. (2008). A survey on usability evaluation techniques and an analysis of their actual application. Report No. 450. Institute of Computer Science, University of Wurzburg, Germany.

Freyne, J., Jacovi, M., Guy, I., & Geyer, W. (2009, October). Increasing engagement through early recommender intervention. In Proceedings of the third ACM conference on Recommender systems (pp. 85-92). ACM.

Friedewald, M., Vildjiounaite, E., Punie, Y., & Wright, D. (2007). Privacy, identity and security in ambient intelligence: a scenario analysis. Telematics and Informatics, 24(1), 15-29.

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007, June). MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In Proceedings of the 5th international conference on Mobile systems, applications and services (pp. 57-70). ACM.

Fuchkina, E., Fischer, P. T., Tien, C. N., von der Heide, A., & Hornecker, E. (2016, September). In-situ versus simulation-based experience evaluation of media augmented urban environments. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (pp. 65-68). ACM.

Fugelstad, P., Dwyer, P., Filson Moses, J., Kim, J., Mannino, C. A., Terveen, L., & Snyder, M. (2012, February). What makes users rate (share, tag, edit...)?: predicting patterns of participation in online communities. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 969-978). ACM.

Gaggioli, A. (2005). Optimal experience in ambient intelligence. In G. Riva, F. Vatalaro, F. Davide, M. Alcañiz (Eds.) Ambient intelligence, 35-43, IOS Press.

Geerts, D., De Moor, K., Ketyko, I., Jacobs, A., Van den Bergh, J., Joseph, W., ... & De Marez, L. (2010, June). Linking an integrated framework with appropriate methods for measuring QoE. In Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on (pp. 158-163). IEEE.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. MIS quarterly, 27(1), 51-90.

Gena, C. (2005). Methods and techniques for the evaluation of user-adaptive systems. The knowledge engineering review, 20(01), 1-37.

Goes, P. B., Lin, M., & Au Yeung, C. M. (2014). "Popularity effect" in user-generated content: evidence from online product reviews. Information Systems Research, 25(2), 222-238.

Grammenos, D., Akoumianakis, D., & Stephanidis, C. (2000). Integrated support for working with guidelines: the Sherlock guideline management system. Interacting with Computers, 12(3), 281-311.

Grill, T., Polacek, O., & Tscheligi, M. (2012, December). Conwiz: A tool supporting contextual wizard of oz simulation. In Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (p. 21). ACM.

Gulati, A., & Dubey, S. K. (2012). Critical analysis on usability evaluation techniques. International Journal of Engineering Science and Technology, 4(3), 990-997.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. ACM Sigkdd Explorations Newsletter, 12(1), 58-72.

Guy, I. (2015). Social recommender systems. In Recommender Systems Handbook (pp. 511-543). Springer US.

Guy, M., & Tonkin, E. (2006). Tidying up tags. D-lib Magazine, 12(1), 1082-9873.

Ha, S., & Stoel, L. (2009). Consumer e-shopping acceptance: Antecedents in a technology acceptance model. Journal of Business Research, 62(5), 565-571.

Halbrügge, M., Quade, M., Engelbrecht, K. P., Möller, S., & Albayrak, S. (2016, September). Predicting user error for ambient systems by integrating model-based UI development and cognitive modeling. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 1028-1039). ACM.

Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. Computers in human behavior, 71, 469-478.

Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. Journal of the Association for Information Science and Technology, 67(9), 2047-2059.

Hanington, B., & Martin, B. (2012). Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions. Rockport Publishers.

Harpur, P., & De Villiers, M. R. (2015). MUUX-E, a framework of criteria for evaluating the usability, user experience and educational features of m-learning environments. South African Computer Journal, 56, 1-21.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. International journal of human-computer interaction, 13(4), 373-410.

Hashemi, M., & Herbert, J. (2016, June). A Service-Oriented User Interaction Analysis Framework Supporting Adaptive Applications. In Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual (Vol. 2, pp. 546-551). IEEE.

Hashizume, A., & Kurosu, M. (2016). "UX Graph Tool" for Evaluating the User Satisfaction. International Journal of Computer Science Issues (IJCSI), 13(5), 86.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. Human-Computer Interaction, 19(4), 319-349.

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. Behaviour & information technology, 25(2), 91-97.

Hassenzahl, M., & Wessler, R. (2000). Capturing design space from a user perspective: The repertory grid technique revisited. International Journal of Human-Computer Interaction, 12(3-4), 441-459.

Heilbrunn, B., Herzig, P., & Schill, A. (2014, December). Tools for gamification analytics: A survey. In Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on (pp. 603-608). IEEE.

Helfrich, B., & Landay, J. A. (1999). QUIP: quantitative user interface profiling. Unpublished manuscript. Available at http://home. Earthlink. Net/~ bhelfrich/quip/index. Html.

Hellweger, S., & Wang, X. (2015). What is User Experience Really: towards a UX Conceptual Framework. arXiv preprint arXiv:1503.01850.

Henninger, S. (2000). A methodology and tools for applying context-specific usability guidelines to interface design. Interacting with computers, 12(3), 225-243.

Herder, E. (2003). Utility-based evaluation of adaptive systems. In S. Weibelzahl and A. Paramythis (eds.) Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems (at the 9th International Conference on User Modeling, UM2003), Pittsburgh, pp. 25–30

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the almere model. International journal of social robotics, 2(4), 361-375.

Hernández, B., Jiménez, J., & Martín, M. J. (2010). Customer behavior in electronic commerce: The moderating effect of e-purchasing experience. Journal of business research, 63(9), 964-971.

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction, 13(4), 421-443.

Heo, J., Ham, D. H., Park, S., Song, C., & Yoon, W. C. (2009). A framework for evaluating the usability of mobile phones based on multi-level, hierarchical model of usability factors. Interacting with Computers, 21(4), 263-275.

Hilbert, D. M., & Redmiles, D. F. (2000). Extracting usability information from user interface events. ACM Computing Surveys (CSUR), 32(4), 384-421.

Holden, R. J., & Karsh, B. T. (2010). The technology acceptance model: its past and its future in health care. Journal of biomedical informatics, 43(1), 159-172.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. International journal of human-computer studies, 64(2), 79-102.

Hornbæk, K., & Law, E. L. C. (2007, April). Meta-analysis of correlations among usability measures. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 617-626). ACM.

Hossain, M. M., & Prybutok, V. R. (2008). Consumer acceptance of RFID technology: An exploratory study. IEEE transactions on engineering management, 55(2), 316-328.

Howard, T. (2009). Design to thrive: Creating social networks and online communities that last, pp. 29-41. Morgan Kaufmann.

Howarth, J., Smith-Jackson, T., & Hartson, R. (2009). Supporting novice usability practitioners with usability engineering tools. International Journal of Human-Computer Studies, 67(6), 533-549.

Howe, J. (2006). The rise of crowdsourcing. Wired magazine, 14(6), 1-4.

Hsu, C. L., & Lu, H. P. (2004). Why do people play on-line games? An extended TAM with social influences and flow experience. Information & management, 41(7), 853-868.

Hu, P. J. H., Al-Gahtani, S. S., & Hu, H. F. (2014). Arabian Workers' Acceptance of Computer Technology: A Model Comparison Perspective. Journal of Global Information Management (JGIM), 22(2), 1-22.

Hussain, A., & Kutar, M. (2009). Usability metric framework for mobile phone application. PGNet, ISBN, 2099, 978-1.

Igbaria, M., Guimaraes, T., & Davis, G. B. (1995). Testing the determinants of microcomputer usage via a structural equation model. Journal of management information systems, 87-114.

Igbaria, M., Zinatelli, N., Cragg, P., & Cavaye, A. L. (1997). Personal computing acceptance factors in small firms: a structural equation model. MIS quarterly, 279-305.

Im, I., Hong, S., & Kang, M. S. (2011). An international comparison of technology adoption: Testing the UTAUT model. Information & management, 48(1), 1-8.

Interaction Design Foundation (2017). Shadowing in User Research – Do You See What They See? Retrieved from https://www.interaction-design.org/literature/article/shadowing-in-user-research-do-you-see-what-they-see

International Organization for Standardization, 9241-210:2010: Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems, 2010.

International Organization for Standardization, ISO/IEC 9241-11: Ergonomic Requirements for Office work with Visual Display Terminals (VDTs) – Part 11: Guidance on Usability, 1998.

Intille, S. S., Larson, K., Beaudin, J. S., Nawyn, J., Tapia, E. M., & Kaushik, P. (2005, April). A living laboratory for the design and evaluation of ubiquitous computing technologies. In CHI'05 extended abstracts on Human factors in computing systems (pp. 1941-1944). ACM.

Intille, S., Larson, K., Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., & Rockinson, R. (2006). Using a live-in laboratory for ubiquitous computing research. Pervasive Computing, 349-365.

Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., & Bao, L. (2003, April). A context-aware experience sampling tool. In CHI'03 extended abstracts on Human factors in computing systems (pp. 972-973). ACM.

Irshad, S., & Rambli, D. R. A. (2015, December). Preliminary user experience framework for designing mobile augmented reality technologies. In Interactive Digital Media (ICIDM), 2015 4th International Conference on (pp. 1-4). IEEE.

ISO/IEC TR 9126-2:2003 (2003). Software engineering – Product quality – Part 2: External metrics.

Ivory, M. Y. (2013). Automated Web Site Evaluation: Researchers' and Practioners' Perspectives (Vol. 4). Springer Science & Business Media.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. ACM Computing Surveys (CSUR), 33(4), 470-516.

Ivory, M. Y., & Hearst, M. A. (2002). Improving web site design. IEEE Internet Computing, 6(2), 56-63.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991, April). User interface evaluation in the real world: a comparison of four techniques. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 119-124). ACM.

Jin, B. S., Ji, Y. G., Choi, K., & Cho, G. (2009). Development of a usability evaluation framework with quality function deployment: from customer sensibility to product design. Human Factors and Ergonomics in Manufacturing & Service Industries, 19(2), 177-194.

Jin, X. L., Zhou, Z., Lee, M. K., & Cheung, C. M. (2013). Why users keep answering questions in online question answering communities: A theoretical and empirical investigation. International Journal of Information Management, 33(1), 93-104.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. Science, 306(5702), 1776-1780.

Karagiannidis, C., & Sampson, D. G. (2000, January). Layered evaluation of adaptive applications and services. In Adaptive Hypermedia and Adaptive Web-Based Systems (pp. 343-346). Springer Berlin Heidelberg.

Karahanna, E., & Limayem, M. (2000). E-mail and v-mail usage: Generalizing across technologies. Journal of organizational computing and electronic commerce, 10(1), 49-66.

Karahanna, E., & Straub, D. W. (1999). The psychological origins of perceived usefulness and ease-of-use. Information & management, 35(4), 237-250.

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. B. (2009, April). User experience over time: an initial framework. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 729-738). ACM.

Kashfi, P., Nilsson, A., & Feldt, R. (2016). Integrating User eXperience Practices into Software Development Processes: Implications of Subjectivity and Emergent Nature of UX. arXiv preprint arXiv:1605.03783.

Kaushik, A. (2007). Web analytics: An hour a day (W/Cd). John Wiley & Sons.

Kellar, M., Hawkey, K., Inkpen, K. M., & Watters, C. (2008). Challenges of capturing natural Web-based user behaviors. Intl. Journal of Human–Computer Interaction, 24(4), 385-409.

Keyson, D. V. (2010, January). Context triggered experience sampling. In Ambient Intelligence and Future Trends-International Symposium on Ambient Intelligence (ISAmI 2010) (pp. 193-196). Springer Berlin Heidelberg.

Khan, M., Sulaiman, S., Said, A. M., & Tahir, M. (2011, March). Development of usability evaluation framework for haptic systems. In VR Innovation (ISVRI), 2011 IEEE International Symposium on (pp. 343-344). IEEE.

Kieras, D. (2009). Model-based evaluation. In A. Sears, J. Jacko (Eds.) Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications (pp. 1192-1208). Lawrence Erlbaum Associates, Taylor & Francis Group, New York.

Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B., Pagulayan, R. J., & Wixon, D. (2008, April). Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 443-452). ACM.

Kim, S. H., Kim, S. W., & Park, H. M. (2003). Usability challenges in ubicomp environment. Proc. Of Int'l. Ergonomics Assoc.

Kim, T. G., Lee, J. H., & Law, R. (2008). An empirical examination of the acceptance behaviour of hotel front office systems: An extended technology acceptance model. Tourism management, 29(3), 500-513.

Kiu, C. C., & Tsui, E. (2011). TaxoFolk: A hybrid taxonomy–folksonomy structure for knowledge classification and navigation. Expert Systems with Applications, 38(5), 6049-6058.

Kjeldskov, J., & Skov, M. B. (2014, September). Was it worth the hassle?: ten years of mobile HCI research discussions on lab and field evaluations. In Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (pp. 43-52). ACM.

Kraut, R. E., & Resnick, P. (2011). Encouraging contribution to online communities. Building successful online communities: Evidence-based social design, 21-76.

Kroski, E. (2005). The hive mind: Folksonomies and user-based tagging. InfoTangle Blog, December.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., & Sinnelä, A. (2011). UX Curve: A method for evaluating long-term user experience. Interacting with Computers, 23(5), 473-483.

Kuo, K. M., Liu, C. F., & Ma, C. C. (2013). An investigation of the effect of nurses' technology readiness on the acceptance of mobile electronic medical record systems. BMC medical informatics and decision making, 13(1), 88.

Kurosu, M. (2015, August). Usability, quality in use and the model of quality characteristics. In International Conference on Human-Computer Interaction (pp. 227-237). Springer, Cham.

Kurosu, M., Hashizume, A., Ueno, Y., Tomida, T., & Suzuki, H. (2016, July). UX Graph and ERM as Tools for Measuring Kansei Experience. In International Conference on Human-Computer Interaction (pp. 331-339). Springer International Publishing.

Lachner, F., Naegelein, P., Kowalski, R., Spann, M., & Butz, A. (2016, October). Quantified UX: Towards a Common Organizational Understanding of User Experience. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction (p. 56). ACM.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings. NIMH Center for the Study of Emotion and Attention, 39-58.

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. New Directions for Methodology of Social & Behavioral Science.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. International journal of human-computer studies, 60(3), 269-298.

Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009, April). Understanding, scoping and defining user experience: a survey approach. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 719-728). ACM.

Law, E., Roto, V., Vermeeren, A. P., Kort, J., & Hassenzahl, M. (2008, April). Towards a shared definition of user experience. In CHI'08 extended abstracts on Human factors in computing systems (pp. 2395-2398). ACM.

Lecerof, A., & Paternò, F. (1998). Automatic support for usability evaluation. IEEE Transactions on Software Engineering, 24(10), 863-888.

Lee, M. C. (2009). Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit. Electronic commerce research and applications, 8(3), 130-141.

Lee, Y., Kozar, K. A., & Larsen, K. R. (2003). The technology acceptance model: Past, present, and future. Communications of the Association for information systems, 12(1), 50.

Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. Information & management, 40(3), 191-204.

Leichtenstern, K., & André, E. (2010, June). MoPeDT: features and evaluation of a user-centred prototyping tool. In Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems (pp. 93-102). ACM.

Leichtenstern, K., André, E., & Rehm, M. (2010, October). Using the hybrid simulation for early user evaluations of pervasive interactions. In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (pp. 315-324). ACM.

Leimeister, J. M. (2010). Collective intelligence. Business & Information Systems Engineering, 2(4), 245-248.

Li, D. C. (2011). Online social network acceptance: a social perspective. Internet Research, 21(5), 562-580.

Li, Y., & Landay, J. A. (2008). Into the wild: low-cost ubicomp prototype testing. Computer, (6), 94-97.

Liao, C., To, P. L., & Hsu, F. C. (2013). Exploring knowledge sharing in virtual communities. Online Information Review, 37(6), 891-909. Lin, H. F. (2007). The role of online and offline features in sustaining virtual communities: an empirical study. Internet Research, 17(2), 119-138.

Lin, H. F. (2008). Determinants of successful virtual communities: Contributions from system characteristics and social factors. Information & Management, 45(8), 522-527.

Lin, H., Fan, W., & Chau, P. Y. (2014). Determinants of users' continuance of social networking sites: A self-regulation perspective. Information & Management, 51(5), 595-603.

Lin, K. M. (2011). E-Learning continuance intention: Moderating effects of user e-learning experience. Computers & Education, 56(2), 515-526.

Lin, K. Y., & Lu, H. P. (2011). Why people use social networking sites: An empirical study integrating network externalities and motivation theory. Computers in human behaviour, 27(3), 1152-1161.

Lin, M. J. J., Hung, S. W., & Chen, C. J. (2009). Fostering the determinants of knowledge sharing in professional virtual communities. Computers in Human Behavior, 25(4), 929-939.

Liu, Y., Li, H., & Carlsson, C. (2010). Factors driving the adoption of m-learning: An empirical study. Computers & Education, 55(3), 1211-1219.

Lou, J., Fang, Y., Lim, K. H., & Peng, J. Z. (2013). Contributing high quantity and quality knowledge to online Q&A communities. Journal of the Association for Information Science and Technology, 64(2), 356-371.

Lu, J., Yao, J. E., & Yu, C. S. (2005). Personal innovativeness, social influences and adoption of wireless Internet services via mobile technology. The Journal of Strategic Information Systems, 14(3), 245-268.

Lui, S. M., Lang, K. R., & Kwok, S. H. (2002, January). Participation incentive mechanisms in peer-to-peer subscription systems. In System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on (pp. 3925-3931). IEEE.

Ma, X., Yan, B., Chen, G., Zhang, C., Huang, K., Drury, J., & Wang, L. (2013). Design and implementation of a toolkit for usability testing of mobile apps. Mobile Networks and Applications, 18(1), 81-97.

MacDonald, C. M., & Atwood, M. E. (2013, April). Changing perspectives on evaluation in HCI: past, present, and future. In CHI'13 Extended Abstracts on Human Factors in Computing Systems (pp. 1969-1978). ACM.

Macleod, M., Bowden, R., Bevan, N., & Curson, I. (1997). The MUSiC performance measurement method. Behaviour & Information Technology, 16(4-5), 279-293.

Macleod, M., & Rengger, R. (1993). The development of DRUM: A software tool for video-assisted usability evaluation. People and Computers, 293-293.

Madden, T. J., Ellen, P. S., & Ajzen, I. (1992). A comparison of the theory of planned behaviour and the theory of reasoned action. Personality and social psychology Bulletin, 18(1), 3-9.

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In 11th australasian conference on information systems (Vol. 53, pp. 6-8).

Magoulas, G. D., Chen, S. Y., & Papanikolaou, K. A. (2003, July). Integrating layered and heuristic evaluation for adaptive learning environments. In Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003, Pittsburgh (pp. 5-14).

Maia, C. L. B., & Furtado, E. S. (2016a). A systematic review about user experience evaluation. In International Conference of Design, User Experience, and Usability (pp. 445-455). Springer International Publishing.

Maia, C. L. B., & Furtado, E. S. (2016b). A study about psychophysiological measures in user experience monitoring and evaluation. In Proceedings of the 15th Brazilian Symposium on Human Factors in Computer Systems (p. 7). ACM.

Mäkelä, K., Salonen, E. P., Turunen, M., Hakulinen, J., & Raisamo, R. (2001, December). Conducting a Wizard of Oz experiment on a ubiquitous computing system doorman. In Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue (pp. 115-119).

Mallat, N., Rossi, M., Tuunainen, V. K., & Öörni, A. (2009). The impact of use context on mobile services acceptance: The case of mobile ticketing. Information & management, 46(3), 190-195.

Malý, I., Curín, J., Kleindienst, J., & Slavík, P. (2008, July). Creation and visualization of user behaviour in ambient intelligent environment. In Information Visualisation, 2008. IV'08. 12th International Conference (pp. 497-502). IEEE.

Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. Behaviour & information technology, 25(2), 141-158.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., & Ames, M. (2003, April). Heuristic evaluation of ambient displays. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 169-176). ACM.

Marti, P., & Iacono, I. (2016, September). Anticipated, momentary, episodic, remembered: the many facets of User eXperience. In Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on (pp. 1647-1655). IEEE.

Martins, A. I., Queirós, A., Silva, A. G., & Rocha, N. P. (2014). Usability evaluation methods: A systematic review. Human Factors in Software Development and Design, 250.

Martins, A. I., Queirós, A., Silva, A. G., & Rocha, N. P. (2016, December). Usability evaluation of ambient assisted living systems using a multi-method approach. In Proceedings of the 7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (pp. 261-268). ACM.

Masip, L., Martinie, C., Winckler, M., Palanque, P., Granollers, T., & Oliva, M. (2012, October). A design process for exhibiting design choices and trade-offs in (potentially) conflicting user interface guidelines. In International Conference on Human-Centred Software Engineering (pp. 53-71). Springer, Berlin, Heidelberg.

McCarthy, J. C., Wright, P. C., & Meekison, L. (2005). A Practitioner-Centered Assessment of a User-Experience Framework. IJTHI, 1(2), 1-23.

McCoy, S., Galletta, D. F., & King, W. R. (2007). Applying TAM across cultures: the need for caution. European Journal of Information Systems, 16(1), 81-90.

Medlock, M. C., Wixon, D., Terrano, M., Romero, R., & Fulton, B. (2002). Using the RITE method to improve products: A definition and a case study. Usability Professionals Association, 51.

Mekler, E. D., Brühlmann, F., Opwis, K., & Tuch, A. N. (2013, October). Do points, levels and leaderboards harm intrinsic motivation?: an empirical analysis of common gamification elements. In Proceedings of the First International Conference on gameful design, research, and applications (pp. 66-73). ACM.

Melas, C. D., Zampetakis, L. A., Dimopoulou, A., & Moustakis, V. (2011). Modeling the acceptance of clinical information systems among hospital medical staff: an extended TAM model. Journal of biomedical informatics, 44(4), 553-564.

Metsis, V., Le, Z., Lei, Y., & Makedon, F. (2008, July). Towards an evaluation framework for assistive environments. In Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments (p. 12). ACM.

Miki, H. (2014, June). User experience evaluation framework for human-centered design. In International Conference on Human Interface and the Management of Information (pp. 602-612). Springer, Cham.

Moeller, S., Engelbrecht, K. P., Hillmann, S., & Ehrenbrink, P. (2014, September). New ITG Guideline for the Usability Evaluation of Smart Home Environments. In Speech Communication; 11. ITG Symposium; Proceedings of (pp. 1-4). VDE.

Mohammadyari, S., & Singh, H. (2015). Understanding the effect of e-learning on individual performance: The role of digital literacy. Computers & Education, 82, 11-25.

Möller, S., Engelbrecht, K. P., Kühnel, C., Wechsung, I., & Weiss, B. (2009). Evaluation of multimodal interfaces for ambient intelligence. Human-Centric Interfaces for Ambient Intelligence, 347-370.

Moon, J. W., & Kim, Y. G. (2001). Extending the TAM for a World-Wide-Web context. Information & management, 38(4), 217-230.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. Information systems research, 2(3), 192-222.

Moores, T. T. (2012). Towards an integrated model of IT acceptance in healthcare. Decision Support Systems, 53(3), 507-516.

Morae (2017). Morae 3 Tutorial. Retrieved from: https://www.techsmith.com/tutorial-morae-configure-remote.html

Mourouzis, A., Antona, M., Boutsakis, E., & Stephanidis, C. (2006). A user-orientation evaluation framework: Assessing accessibility throughout the user experience lifecycle. In Computers Helping People with Special Needs (pp. 421-428). Springer Berlin Heidelberg.

Muhi, K., Szőke, G., Fülöp, L. J., Ferenc, R., & Berger, Á. (2013). A Semi-automatic Usability Evaluation Framework. In Computational Science and Its Applications–ICCSA 2013 (pp. 529-542). Springer Berlin Heidelberg.

Muller, M. J., Haslwanter, J. H., & Dayton, T. (1997). Participatory practices in the software lifecycle. Handbook of human-computer interaction, 2, 255-297.

Mulwa, C., Lawless, S., Sharp, M., & Wade, V. (2011). The evaluation of adaptive and personalised information retrieval systems: a review. International Journal of Knowledge and Web Intelligence, 2(2-3), 138-156.

Mun, Y. Y., Jackson, J. D., Park, J. S., & Probst, J. C. (2006). Understanding information technology acceptance by individual professionals: Toward an integrative view. Information & Management, 43(3), 350-363.

Neely, S., Stevenson, G., Kray, C., Mulder, I., Connelly, K., & Siek, K. A. (2008). Evaluating pervasive and ubiquitous systems. IEEE Pervasive Computing, 7(3).

Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J., & Stenild, S. (2006, October). It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field. In Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles (pp. 272-280). ACM.

Nielsen, J. (1992). The usability engineering life cycle. Computer, 25(3), 12-22.

Nielsen, J. (1994a). Heuristic evaluation. Usability inspection methods, 17(1), 25-62.

Nielsen, J. (1994b). Usability engineering. Elsevier.

Nielsen, J. (1994c). Usability inspection methods. In Conference companion on Human factors in computing systems (pp. 413-414). ACM.

Nielsen, J. (2001). Success Rate: The Simplest Usability Metric. Retrieved from: https://www.nngroup.com/articles/success-rate-the-simplest-usability-metric/

Nielsen, J. (2010). Interviewing Users. Retrieved from: http://www.nngroup.com/articles/interviewing-users/

Nielsen, J. (2012). Thinking aloud: The #1 Usability Tool. Retrieved from: http://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

Nielsen, L. (2012). Personas-user focused design (Vol. 15). Springer Science & Business Media.

Nishikawa, H., Yamamoto, S., Tamai, M., Nishigaki, K., Kitani, T., Shibata, N., Yasumoto, K., & Ito, M. (2006). UbiREAL: realistic smartspace simulator for systematic testing. In UbiComp 2006: Ubiquitous Computing (pp. 459-476). Springer Berlin Heidelberg.

Norman, K. L., & Panizzi, E. (2005). Levels of automation and user participation in usability testing. Interacting with computers, 18(2), 246-264.

Nov, O., Naaman, M., & Ye, C. (2010). Analysis of participation in an online photo-sharing community: A multidimensional perspective. Journal of the Association for Information Science and Technology, 61(3), 555-566.

Ntoa, S., Antona, M., & Stephanidis, C. (2017). Towards Technology Acceptance Assessment in Ambient Intelligence Environments. In the Proceedings of the Seventh International Conference on Ambient Computing, Applications, Services and Technologies (AMBIENT 2017), 12 - 16 November, Barcelona, Spain.

Ntoa, S., Margetis, G., & Stephanidis, C. (2017). Rewarding Contributions in a User-Generated Video Online Community. 1st International Workshop on Converging User-Generated Material with Professional Video User Experiences in the context of ACM TVX 2017, 14-16 June, Hilversum, The Netherlands. Retrieved from: https://zenodo.org/record/834200#.WcUVn7IjFaR

Nysveen, H., Pedersen, P. E., & Thorbjørnsen, H. (2005). Intentions to use mobile services: Antecedents and cross-service comparisons. Journal of the academy of marketing science, 33(3), 330-346.

Olsson, T. (2014, April). Layers of user expectations of future technologies: an early framework. In Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems (pp. 1957-1962). ACM.

O'Neill, E., Klepal, M., Lewis, D., O'Donnell, T., O'Sullivan, D., & Pesch, D. (2005, February). A testbed for evaluating human interaction with ubiquitous computing environments. In Testbeds and Research Infrastructures for the Development of Networks and Communities, 2005. Tridentcom 2005. First International Conference on (pp. 60-69). IEEE.

Ozkul, T. A. R. I. K., & El Zarka, A. T. (2013). A new model for measuring Human-Smart machine interface quality. Recent Advances in Electrical Engineering Series, 12, 191-196.

Ozok, A. A. (2009). Survey design and implementation in HCI. In A. Sears, J. Jacko (Eds.) Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications (pp. 1151-1170). Lawrence Erlbaum Associates, Taylor & Francis Group, New York.

Paganelli, L., & Paternò, F. (2002, January). Intelligent analysis of user interactions with web applications. In Proceedings of the 7th international conference on Intelligent user interfaces (pp. 111-118). ACM.

Pai, F. Y., & Yeh, T. M. (2014). The effects of information sharing and interactivity on the intention to use social networking websites. Quality & Quantity, 48(4), 2191-2207.

Paramythis, A., Totter, A., & Stephanidis, C. (2001, July). A modular approach to the evaluation of adaptive user interfaces. In Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001 (pp. 9-24).

Paramythis, A., Weibelzahl, S., & Masthoff, J. (2010). Layered evaluation of interactive adaptive systems: framework and formative methods. User Modeling and User-Adapted Interaction, 20(5), 383-453.

Park, J., Han, S. H., Kim, H. K., Oh, S., & Moon, H. (2013). Modeling user experience: A case study on a mobile device. International Journal of Industrial Ergonomics, 43(2), 187-196.

Park, S. Y. (2009). An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. Journal of Educational Technology & Society, 12(3), 150.

Parush, A. (2000). A database approach to building and using online human computer interaction guidelines. Tools for Working with Guidelines, 77-84.

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. International journal of electronic commerce, 7(3), 101-134.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. Journal of personality and social psychology, 89(3), 277.

Paz, F., & Pow-Sang, J. A. (2014, December). Current trends in usability evaluation methods: A systematic review. In Advanced Software Engineering and Its Applications (ASEA), 2014 7th International Conference on (pp. 11-15). IEEE.

Pereira, C., Teixeira, A., & e Silva M. P. (2014, May). Live evaluation within ambient assisted living scenarios. In Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments (p. 14). ACM.

Pi, S. M., Chou, C. H., & Liao, H. L. (2013). A study of Facebook Groups members' knowledge sharing. Computers in Human Behavior, 29(5), 1971-1979.

Pituch, K. A., & Lee, Y. K. (2006). The influence of system characteristics on e-learning use. Computers & Education, 47(2), 222-244.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. Encyclopedia of human computer interaction, 1, 211-219.

Poppe, R., Rienks, R., & van Dijk, B. (2007). Evaluating the future of HCI: challenges for the evaluation of emerging applications. Artifical Intelligence for Human Computing, 234-250

Porter, C. E., & Donthu, N. (2006). Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics. Journal of business research, 59(9), 999-1007.

Portet, F., Vacher, M., Golanski, C., Roux, C., & Meillon, B. (2013). Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Personal and Ubiquitous Computing, 17(1), 127-144.

Poslad, S. (2011). Ubiquitous computing: smart devices, environments and interactions. John Wiley & Sons.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. AIS Transactions on Human-Computer Interaction, 1(1), 13-32.

Pribeanu, C. (2009). A usability assistant for the heuristic evaluation of interactive systems. Studies in Informatics and Control, 18(4), 355-362.

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164). ACM.

Raake, A., & Egger, S. (2014). Quality and quality of experience. In Quality of experience (pp. 11-33). Springer International Publishing.

Rauterberg, M. (1993). AMME: an automatic mental model evaluation to analyse user behaviour traced in a finite, discrete state space. Ergonomics, 36(11), 1369-1380.

Read, J. C., & MacFarlane, S. (2006, June). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In Proceedings of the 2006 conference on Interaction design and children (pp. 81-88). ACM.

Reitberger, W., Obermair, C., Ploderer, B., Meschtscherjakov, A., & Tscheligi, M. (2007). Enhancing the shopping experience with ambient displays: a field study in a retail store. In Ambient intelligence (pp. 314-331). Springer Berlin Heidelberg.

Remagnino, P., & Foresti, G. L. (2005). Ambient intelligence : A new multidisciplinary paradigm. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 35(1), 1-6.

Roca, J. C., Chiu, C. M., & Martínez, F. J. (2006). Understanding e-learning continuance intention: An extension of the Technology Acceptance Model. International Journal of human-computer studies, 64(8), 683-696.

Rogers Everett, M. (1995). Diffusion of innovations. New York.

Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R. E., ... & Toscos, T. (2007, September). Why it's worth the hassle: The value of in-situ studies when designing ubicomp. In International Conference on Ubiquitous Computing (pp. 336-353). Springer Berlin Heidelberg

Roto, V., Obrist, M., & Väänänen-Vainio-Mattila, K. (2009, August). User experience evaluation methods in academic and industrial contexts. In Interact 2009 conference, User Experience Evaluation Methods in Product Development (UXEM'09), Uppsala, Sweden.

Rukshan, A., & Baravalle, A. (2011). A quantitative approach to usability evaluation of web sites.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: a single-item scale of pleasure and arousal. Journal of personality and social psychology, 57(3), 493.

Ryu, H., Hong, G. Y., & James, H. (2006). Quality assessment technique for ubiquitous software and middleware.

Saadé, R., & Bahli, B. (2005). The impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: an extension of the technology acceptance model. Information & management, 42(2), 317-327.

Sadri, F. (2011). Ambient intelligence: A survey. ACM Computing Surveys (CSUR), 43(4), 36.

Sanders, E. B. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. Co-design, 4(1), 5-18.

Santos, R. M., de Oliveira, K. M., Andrade, R. M., Santos, I. S., & Lima, E. R. (2013). A quality model for human-computer interaction evaluation in ubiquitous systems. In Human Computer Interaction (pp. 63-70). Springer International Publishing.

Savioja, P., Liinasuo, M., & Koskinen, H. (2014). User experience: does it matter in complex systems?. Cognition, technology & work, 16(4), 429-449.

Schade A. (2013). Competitive Usability Evaluations: Learning from Your Competition. Retrieved from: https://www.nngroup.com/articles/competitive-usability-evaluations/

Schatz, R., Hoßfeld, T., Janowski, L., & Egger, S. (2013). From packets to people: quality of experience as a new measurement challenge. In Data traffic monitoring and analysis (pp. 219-263). Springer Berlin Heidelberg.

Schlatter, T., & Levinson, D. (2013). Visual usability: Principles and practices for designing digital applications. Newnes.

Schmidt, A. (2000). Implicit human computer interaction through context. Personal technologies, 4(2), 191-199.

Schmidt, A. (2003). Ubiquitous computing-computing in context (Doctoral dissertation, Lancaster University).

Scholtz, J., & Consolvo, S. (2004). Toward a framework for evaluating ubiquitous computing applications. Pervasive Computing, IEEE, 3(2), 82-88.

Schroer, J., & Hertel, G. (2009). Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. Media Psychology, 12(1), 96-120.

Schulze, K., & Krömker, H. (2010, August). A framework to measure user experience of interactive online products. In Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research (p. 14). ACM.

Sears, A. (1995, December). AIDE: A step toward metric-based interface development tools. In Proceedings of the 8th annual ACM symposium on User interface and software technology (pp. 101-110). ACM.

Sears, A. (1997). Heuristic walkthroughs: Finding the problems without the noise. International Journal of Human-Computer Interaction, 9(3), 213-234.

Seffah, A., Kececi, N., & Donyae, M. (2001). QUIM: a framework for quantifying usability metrics in software quality models. In Quality Software, 2001. Proceedings. Second Asia-Pacific Conference on (pp. 311-318). IEEE.

Seffah, A., Donyaee, M., Kline, R. B., & Padda, H. K. (2006). Usability measurement and metrics: A consolidated model. Software Quality Journal, 14(2), 159-178.

Seo, D. W., Kim, H., Kim, J. S., & Lee, J. Y. (2016). Hybrid reality-based user experience and evaluation of a context-aware smart home. Computers in Industry, 76, 11-23.

Sipos, R., Ghosh, A., & Joachims, T. (2014, April). Was this review helpful to you?: it depends! Context and voting patterns in online content. In Proceedings of the 23rd international conference on World wide web (pp. 337-348). ACM.

Shin, D. H. (2010). Ubiquitous computing acceptance model: end user concern about security, privacy and risk. International Journal of Mobile Communications, 8(2), 169-186.

Shneiderman, B. (2017): The Paradigm Birth of HCI and User Experience Design: Starting a Discipline and Launching an Industry. Retrieved from: https://medium.com/@ben.shneiderman/the-paradigm-birth-of-hci-and-user-experience-design-starting-a-discipline-and-launching-an-c3929e5a3efe

Singh, P., Ha, H. N., Kuang, Z., Olivier, P., Kray, C., Blythe, P., & James, P. (2006, September). Immersive video as a rapid prototyping and evaluation tool for mobile and ambient applications. In Proceedings of the 8th conference on Human-computer interaction with mobile devices and services (pp. 264-264). ACM.

Sommaruga, L., Rota, P., & Catenazzi, N. (2011). "Tagsonomy": Easy Access to Web Sites through a Combination of Taxonomy and Folksonomy. Advances in Intelligent Web Mastering–3, 61-71.

Sommerville, I., & Dewsbury, G. (2007). Dependable domestic systems design: A socio-technical approach. Interacting with Computers, 19(4), 438-456.

Song, W., Tjondronegoro, D., & Docherty, M. (2012). Understanding user experience of mobile video: framework, measurement, and optimization. In Mobile Multimedia-User and Technology Perspectives. InTech.

Spencer, D. (2009). Card sorting: Designing usable categories. Rosenfeld Media.

Spiteri, L. F. (2007). The structure and form of folksonomy tags: The road to the public library catalog. Information technology and libraries, 26(3), 13.

Spool, J. (2004). The KJ-Technique: A Group Process for Establishing Priorities. https://articles.uie.com/kj_technique/

Srite, M., & Karahanna, E. (2006). The role of espoused national cultural values in technology acceptance. MIS quarterly, 679-704.

Ståhl, A., Höök, K., Svensson, M., Taylor, A. S., & Combetto, M. (2009). Experiencing the affective diary. Personal and Ubiquitous Computing, 13(5), 365-378.

Stanton, N. A. (2006). Hierarchical task analysis: Developments, applications, and extensions. Applied ergonomics, 37(1), 55-79.

Stephanidis, C. (2012). Human factors in ambient intelligence environments. Handbook of Human Factors and Ergonomics, 4th edn. John Wiley and Sons, USA (to appear, 2012).

Stephanidis, C., Paramythis, A., Akoumianakis, D., & Sfyrakis, M. (1998, October). Self-adapting web-based systems: Towards universal accessibility. In 4th Workshop on User Interface For All, Stockholm, Sweden.

Strasser, E., Weiss, A., Grill, T., Osswald, S., & Tscheligi, M. (2012). Combining implicit and explicit methods for the evaluation of an ambient persuasive factory display. In Ambient Intelligence (pp. 113-128). Springer Berlin Heidelberg.

Stutzman, F., Capra, R., & Thompson, J. (2011). Factors mediating disclosure in social network sites. Computers in Human Behavior, 27(1), 590-598.

Šumak, B., HeričKo, M., & PušNik, M. (2011). A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types. Computers in Human Behavior, 27(6), 2067-2077.

Sun, Y., Fang, Y., & Lim, K. H. (2012). Understanding sustained participation in transactional virtual communities. Decision Support Systems, 53(1), 12-22.

Sweeney, M., Maguire, M., & Shackel, B. (1993). Evaluating user-computer interaction: a framework. International journal of man-machine studies, 38(4), 689-711.

Tausczik, Y. R., Dabbish, L. A., & Kraut, R. E. (2014, February). Building loyalty to online communities through bond and identity-based attachment to sub-groups. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 146-157). ACM.

Taylor, S., & Todd, P. (1995). Decomposition and crossover effects in the theory of planned behaviour: A study of consumer adoption intentions. International journal of research in marketing, 12(2), 137-155.

Tedjamulia, S. J., Dean, D. L., Olsen, D. R., & Albrecht, C. C. (2005, January). Motivating content contributions to online communities: Toward a more comprehensive theory. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 193b-193b). IEEE.

Thayer, A., & Dugan, T. E. (2009, July). Achieving design enlightenment: Defining a new user experience measurement framework. In Professional Communication Conference, 2009. IPCC 2009. IEEE International (pp. 1-10). IEEE.

Thompson, R. L., Higgins, C. A., & Howell, J. M. (1991). Personal computing: toward a conceptual model of utilization. MIS quarterly, 125-143.

Tiedtke, T., Märtin, C., & Gerth, N. (2002, June). AWUSA–A tool for automated website usability analysis. In proceedings of 9th International Workshop on Design, Specification and Verification DSV-IS.

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings. Engineering Management, IEEE Transactions on, (1), 28-45.

Tran, C. D., Ezzedine, H., & Kolski, C. (2013). EISEval, a generic reconfigurable environment for evaluating agent-based interactive systems. International Journal of Human-Computer Studies, 71(6), 725-761.

Triandis, H. C. (1979). Values, attitudes, and interpersonal behaviour. In Nebraska symposium on motivation. University of Nebraska Press.

Tung, F. C., Chang, S. C., & Chou, C. M. (2008). An extension of trust and TAM model with IDT in the adoption of the electronic logistics information system in HIS in the medical industry. International journal of medical informatics, 77(5), 324-335.

Turner, P. (2008). Towards an account of intuitiveness. Behaviour & Information Technology, 27(6), 475-482.

Väänänen-Vainio-Mattila, K., Olsson, T., & Häkkilä, J. (2015, September). Towards deeper understanding of user experience with ubiquitous computing systems: systematic literature review and design framework. In Human-Computer Interaction (pp. 384-401). Springer, Cham.

Van Helvert, J., & Wagner, C. (2016). From scenarios to "free-play": Evaluating the user's experience of ambient technologies in the home. In Next Generation Intelligent Environments (pp. 325-344). Springer International Publishing.

Van Velsen, L., Van Der Geest, T., Klaassen, R., & Steehouder, M. (2008). User-centered evaluation of adaptive and adaptable systems: a literature review. The knowledge engineering review, 23(3), 261-281.

Vanderdonckt, J. (1999). Development milestones towards a tool for working with guidelines.

Vanderdonckt, J. (2001). A small knowledge-based system for selecting interaction styles. In *Tools for Working with Guidelines* (pp. 247-262). Springer, London., J. (Vanderdonckt, J. (2001). A small knowledge-based system for selecting interaction styles. In *Tools for Working with Guidelines* (pp. 247-262). Springer, London.). A small knowledge-based system for selecting interaction styles. In Tools for Working with Guidelines (pp. 247-262). Springer, London.

Vassileva, J. (2012). Motivating participation in social computing applications: a user modeling perspective. User Modeling and User-Adapted Interaction, 22(1), 177-201.

Vastenburg, M. H., & Herrera, N. R. (2010, January). Adaptive experience sampling: addressing the dynamic nature of in-situ user studies. In Ambient Intelligence and Future Trends-International Symposium on Ambient Intelligence (ISAmI 2010) (pp. 197-200). Springer Berlin Heidelberg.

Vastenburg, M. H., & Herrera, N. A. R. (2011). Experience tags: enriching sensor data in an awareness display for family caregivers. In Ambient Intelligence (pp. 285-289). Springer Berlin Heidelberg.

Vavoula, G. N., & Sharples, M. (2009). Meeting the challenges in evaluating mobile learning: a 3-level evaluation framework.

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. Decision sciences, 39(2), 273-315.

Venkatesh, V., & Davis, F. D. (1996). A model of the antecedents of perceived ease of use: Development and test. Decision sciences, 27(3), 451-481.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management science, 46(2), 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS quarterly, 425-478.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS quarterly, 36(1), 157-178.

Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010, October). User experience evaluation methods: current state and development needs. In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (pp. 521-530). ACM.

Vora, P. R., & Helander, M. G. (1995). A teaching method as an alternative to the concurrent think-aloud method for usability testing. Advances in Human Factors/Ergonomics, 20, 375-380.

Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring the hedonic and utilitarian dimensions of consumer attitude. Journal of marketing research, 40(3), 310-320.

Wäljas, M., Segerståhl, K., Väänänen-Vainio-Mattila, K., & Oinas-Kukkonen, H. (2010, September). Cross-platform service user experience: a field study and an initial framework. In Proceedings of the 12th international conference on Human computer interaction with mobile devices and services (pp. 219-228). ACM.

Wang, H., Chung, J. E., Park, N., McLaughlin, M. L., & Fulk, J. (2012). Understanding online community participation: A technology acceptance perspective. Communication Research, 39(6), 781-801.

Want, R. (2010). An introduction to ubiquitous computing. Ubiquitous Computing Fundamentals, 1-36, CRC Press.

Wasko, M. M., Teigland, R., & Faraj, S. (2009). The provision of online public goods: Examining social structure in an electronic network of practice. Decision Support Systems, 47(3), 254-265.

Ward, R. D., & Marsden, P. H. (2003). Physiological responses to different web page designs. International Journal of Human-Computer Studies, 59(1), 199-212.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of personality and social psychology, 54(6), 1063.

Wechsung, I., & De Moor, K. (2014). Quality of experience versus user experience. In Quality of Experience (pp. 35-54). Springer International Publishing.

Weibelzahl, S. (2002). Evaluation of Adaptive Systems (Doctoral dissertation). University of Freiburg, Germany.

Weibelzahl, S., & Lauer, C. U. (2001). Framework for the evaluation of adaptive CBR-systems. In Proceedings of the 9th German Workshop on Case-Based Reasoning (pp. 254-263).

Weiser, M. (1995). The computer for the 21st century. Scientific American, 272(3), 78-89.

Weiss, A., Bernhaupt, R., Lankes, M., & Tscheligi, M. (2009, April). The USUS evaluation framework for human-robot interaction. In AISB2009: proceedings of the symposium on new frontiers in human-robot interaction (Vol. 4, pp. 11-26).

Werbach, K., & Hunter, D. (2015). The gamification toolkit: dynamics, mechanics, and components for the win. Wharton Digital Press.

Wharton, C., Rieman, J., Lewis, C. & Polson, P. (1994). The Cognitive Walkthrough: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), Usability inspections methods (pp. 105-140). New York: Wiley.

Wilson, C. (2013). Perspective-Based UI Inspection. In User Interface Inspection Methods: A User-Centered Design Method, pp. 50-63. Newnes.

Wu, H. Y., Liao, W. H., Chen, P. M., Hsu, C. Y., & Li, T. Y. (2014). Logging and Analyzing Long-Term Mobile User Behavior. International Journal of Electronic Commerce Studies, 5(2), 161.

Wu, O., Chen, Y., Li, B., & Hu, W. (2011, February). Evaluating the visual quality of web pages using a computational aesthetic approach. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 337-346). ACM.

Wu, W., Arefin, A., Rivas, R., Nahrstedt, K., Sheppard, R., & Yang, Z. (2009, October). Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. In Proceedings of the 17th ACM international conference on Multimedia (pp. 481-490). ACM.

Wu, L., & Chen, J. L. (2005). An extension of trust and TAM model with TPB in the initial adoption of on-line tax: an empirical study. International Journal of Human-Computer Studies, 62(6), 784-808.

Yang, H. I., Chen, C., Abdulrazak, B., & Helal, S. (2010). A framework for evaluating pervasive systems. International Journal of Pervasive Computing and Communications, 6(4), 432-481.

Yarbrough, A. K., & Smith, T. B. (2007). Technology acceptance among physicians: a new take on TAM. Medical Care Research and Review, 64(6), 650-672.

Yu, J., Ha, I., Choi, M., & Rho, J. (2005). Extending the TAM for a t-commerce. Information & management, 42(7), 965-976.

Yusoff, A., Crowder, R., & Gilbert, L. (2010, March). Validation of serious games attributes using the technology acceptance model. In Games and Virtual Worlds for Serious Applications (VS-GAMES), 2010 Second International Conference on (pp. 45-51). IEEE.

Zachhuber, D., Grill, T., Polacek, O., & Tscheligi, M. (2012). Contextual Wizard of Oz. In Ambient Intelligence (pp. 224-239). Springer Berlin Heidelberg.

Zarmpou, T., Saprikis, V., Markos, A., & Vlachopoulou, M. (2012). Modeling users' acceptance of mobile services. Electronic Commerce Research, 12(2), 225-248.

Zettlemoyer, L. S., St Amant, R., & Dulberg, M. S. (1998, December). IBOTS: Agent control through the user interface. In Proceedings of the 4th international conference on Intelligent user interfaces (pp. 31-37). ACM.

Zhao, L., Lu, Y., Wang, B., Chau, P. Y., & Zhang, L. (2012). Cultivating the sense of belonging and motivating user participation in virtual communities: A social capital perspective. International Journal of Information Management, 32(6), 574-588.

Zhang, J., & Walji, M. F. (2011). TURF: Toward a unified framework of HER usability. Journal of biomedical informatics, 44(6), 1056-1067.

Zhang, Y., Fang, Y., Wei, K. K., & Chen, H. (2010). Exploring the role of psychological safety in promoting the intention to continue sharing knowledge in virtual communities. International Journal of Information Management, 30(5), 425-436.

Zhang, Z. (2010). Feeling the sense of community in social networking usage. IEEE Transactions on Engineering Management, 57(2), 225-239.

Zhang, Z., Basili, V., & Shneiderman, B. (1999). Perspective-based usability inspection: An empirical validation of efficacy. Empirical Software Engineering, 4(1), 43-69.

Zhao, K., Stylianou, A. C., & Zheng, Y. (2013). Predicting users' continuance intention in virtual communities: The dual intention-formation processes. Decision Support Systems, 55(4), 903-910.

Zhao, L., Detlor, B., & Connelly, C. E. (2016). Sharing Knowledge in Social Q&A Sites: The Unintended Consequences of Extrinsic Motivation. Journal of Management Information Systems, 33(1), 70-100.

Zhao, Y., & Zhu, Q. (2014). Evaluation on crowdsourcing research: Current status and future direction. Information Systems Frontiers, 16(3), 417-434.

Zimmermann, P. G. (2008). Beyond Usability: Measuring Aspects of User Experience (Doctoral dissertation, ETH).

Zülch, G., & Stowasser, S. (2000). Usability evaluation of user interfaces with the computer-aided evaluation tool PROKUS. MMI-Interaktiv, 3, 1-17.

APPENDIX A – LIST OF PUBLICATIONS

Papers in Proceedings of International Conferences (full-paper reviewed)

- Ntoa, S., Antona, M., & Stephanidis, C. (2017). Towards Technology Acceptance Assessment in Ambient Intelligence Environments. In the Proceedings of the Seventh International Conference on Ambient Computing, Applications, Services and Technologies (AMBIENT 2017), 12 - 16 November, Barcelona, Spain.

- Evans, M., Margetis, G., Ntoa, S., & Weerakkody, R. (2017). Converging User-Generated Material with Professional Video User Experiences. In the *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video, (ACM TVX 2017)*, 14-16 June, Hilversum, The Netherlands (pp. 135-137). New York: ACM.

Papers in Proceedings of international workshops (full-paper reviewed)

- Ntoa, S., Margetis, G., & Stephanidis, C. (2017). Rewarding Contributions in a User-Generated Video Online Community. 1st International Workshop on Converging User-Generated Material with Professional Video User Experiences in the context of ACM TVX 2017, 14-16 June, Hilversum, The Netherlands. [On-line]. Available at: https://zenodo.org/record/834200#.WcUVn7IjFaR

APPENDIX B – UXAMI TERMS AND DEFINITIONS

Accuracy of input (sensor) data perceived by the system: Check how accurate are the raw data received by the system (e.g., data from sensors).

Validity of system interpretations: Check how valid is the meaning given by the system to the collected raw data.

Input error: an error referring to incorrect usage of input modalities.

Interaction error: an error referring to incorrect usage of the user interface (e.g., selecting an inappropriate menu item for the task at hand)

Cross-platform task: a task that is carried out in more than one devices.

Correctness of system's conflict resolution: Assessment of how correct was the decision taken by the system to resolve a conflict of interests / demands between two or more users.

Social etiquette: code of behaviour that delineates expectations for social behaviour according to contemporary conventional norms within a society, social class, or group.

Implicit interaction: an action performed by the user that is not primarily aimed to interact with a computerised system but which such a system understands as input.

Systems to which a pervasive application has been deployed: pervasiveness refers to the capability of an application to run in multiple systems (e.g., tablet, smartphone, large screen display).

Trialability: the degree to which an innovation may be experimented with before adoption.

Relative advantage: the degree to which an innovation is perceived as being better than its precursor / competitors.

Self-efficacy: an individual's convictions about his or her abilities to mobilize motivation, cognitive resources and courses of action needed to successfully execute a specific task within a given context (e.g., computer self-efficacy is defined as an individual judgement of one's capability to use a computer)

Personal innovativeness: the individual's willingness to try out any new technology

Subjective norm: a person's perception that most people who are important to him think he should or should not perform the behaviour in question.

Voluntariness: the extent to which potential adopters perceive the adoption decision to be non-mandatory.

End-user support: specialized instruction, guidance, coaching and consulting.

Visibility: the degree to which the innovation is visible in context of use (e.g., the organization).

Observability: the degree to which aspects of an innovation may be conveyed to others.

Image: the degree to which use of an innovation is perceived to enhance one's status in one's social system.

APPENDIX C – UXAMI METRICS PER EVALUATION METHOD

Classification of metrics per method and per UX evaluation phase: metrics that should be assessed through expert-based reviews (Table 23), questionnaire-based metrics for user-based experiments to be acquired before the experiment (Table 24), observation metrics that can be automatically acquired with the help of the AmI environment during the experiment (Table 25), observation metrics regarding the experiment that need to be marked by the evaluator and receive automation support for calculations through tools (Table 26), metrics that should be pursued through questionnaires (Table 27) or interviews (Table 28) shortly after the system usage in a user-based experiment, as well as metrics that should be acquired long time after the system usage through questionnaires (Table 29).

Table 30. UXAmI metrics to be assessed through expert-based reviews

UNOBTRUSIVENESS	
Embedment	The system and its components are appropriately embedded in the surrounding architecture
ADAPTABILITY AND ADAPTIVITY	
Interpretations	Validity of system interpretations
Appropriateness of adaptation	Interaction modalities are appropriately adapted according to the user profile and context of use
	System output is appropriately adapted according to the user profile and context of use
	Content is appropriately adapted according to the user profile and context of use
Appropriateness of recommendations	The system adequately explains any recommendations
	The system provides an adequate way for users to express and revise their preferences
	Recommendations are appropriate for the specific user and context of use
USABILITY	
Conformance with guidelines	The user interfaces of the systems comprising the AmI environment conform to relevant guidelines
Learnability	Users can easily understand and use the system
Accessibility	The system conforms to accessibility guidelines

Physical UI	The systems of the AmI environment are electronically accessible
	The AmI environment is physically accessible
	The system does not violate any ergonomic guidelines
	The size and position of the system is appropriate for its manipulation by the target user groups
Cross-platform usability	Consistency among the user interfaces of the individual systems
	Content is appropriately synchronised for cross-platform tasks
	Available actions are appropriately synchronised for cross-platform tasks
Multi-user usability	Social etiquette is followed by the system
Implicit interactions	Appropriateness of system responses to implicit interactions

APPEAL AND EMOTIONS	
Aesthetics	The systems follow principles of aesthetic design

SAFETY AND PRIVACY	
User control	User has control over the data collected
	User has control over the dissemination of information
	The user can customise the level of control that the AmI environment has: high (acts on behalf of the person), medium (gives advice), low (executes a person's commands)
Privacy	Availability of the user's information to other users of the system or third parties
	Availability of explanations to a user about the potential use of recorded data
	Expressiveness of the security (privacy) policy
Safety	The AmI environment is safe for its operators
	The AmI environment is safe in terms of public health
	The AmI environment does not cause environmental harm
	The AmI environment will not cause harm to commercial property, operations or reputation in the intended contexts of use

Table 31. UXAmI metrics to be assessed through questionnaires before a user-based study

TECHNOLOGY ACCEPTANCE AND ADOPTION	
User attributes	Self-efficacy
	Computer attitude
	Age
	Gender
	Personal innovativeness
Social influences	Subjective norm
	Voluntariness
Facilitating conditions	Visibility
Expected outcomes	Perceived benefit
	Long-term consequences of use
Trust	User trust towards the system

Table 32. UXAmI metrics automatically measured during a user-based study

INTUITIVENESS	
Awareness of application capabilities	Functionalities that have been used for each system
	Undiscovered functionalities of each system
Awareness of the interaction vocabulary	Percentage of input modalities used
	Erroneous user inputs (inputs that have not been recognized by the system) for each supported input modality
	Percentage of erroneous user inputs per input modality
ADAPTABILITY AND ADAPTIVITY	
Adaptation impact	Number of erroneous user inputs (i.e., incorrect use of input commands) once an adaptation has been applied
Appropriateness of recommendations	Percentage of accepted system recommendations
USABILITY	

Effectiveness	Number of input errors
	Number of system failures
Efficiency	Task time
Learnability	Number of interaction errors over time
	Number of input errors over time
	Number of help requests over time
Cross-platform usability	After switching device: number of interaction errors until task completion
	Help requests after switching devices
	Cross-platform task time compared to the task time when the task is carried out in a single device (per device)
Multi-user usability	Number of collisions with activities of others
	Percentage of conflicts resolved by the system
Implicit interactions	Implicit interactions carried out by the user
	Number of implicit interactions carried out by the user
	Percentages of implicit interactions per implicit interaction type
Usage	Global interaction heat map: number of usages per hour on a daily, weekly and monthly basis for the entire AmI environment
	Systems' interaction heat map: number of usages for each system in the AmI environment per hour on a daily, weekly and monthly basis
	Applications' interaction heat map: number of usages for each application in the AmI environment per hour on a daily, weekly and monthly basis
	Time duration of users' interaction with the entire AmI environment
	Time duration of users' interaction with each system of the AmI environment
	Time duration of users' interaction with each application of the AmI environment
	Analysis (percentage) of applications' used per system (for systems with more than one applications)
	Percentage of systems to which a pervasive application has been deployed, per application
APPEAL AND EMOTIONS	

Actionable emotions	Detection of users' emotional strain through physiological measures, such as heart rate, skin resistance, blood volume pressure, gradient of the skin resistance and speed of the aggregated changes in the all variables' incoming data.

Table 33. UXAmI metrics that should be measured during a user-based study and can have automation support by tools

UNOBTRUSIVENESS	
Distraction	Number of times that the user has deviated from the primary task
	Time elapsed from a task deviation until the user returns to the primary task
ADAPTABILITY AND ADAPTIVITY	
Input (sensor) data	Accuracy of input (sensor) data perceived by the system
Interpretations	Validity of system interpretations
Appropriateness of adaptation	Interaction modalities are appropriately adapted according to the user profile and context of use
	System output is appropriately adapted according to the user profile and context of use
	Content is appropriately adapted according to the user profile and context of use
	Adaptations that have been manually overridden by the user
Adaptation impact	Number of erroneous user interactions once an adaptation has been applied
	Percentage of adaptations that have been manually overridden by the user
Appropriateness of recommendations	The system adequately explains any recommendations
	Recommendations are appropriate for the specific user and context of use
	Recommendations that have not been accepted by the user
USABILITY	
Effectiveness	Task success
	Number of interaction errors
Efficiency	Number of help requests
	Time spent on errors

User satisfaction	Percent of favourable user comments / unfavourable user comments
	Number of times that users express frustration
	Number of times that users express clear joy
Cross-platform usability	After switching device: time spent to continue the task from where it was left
	Cross-platform task success compared to the task success when the task is carried out in a single device (per device)
Multi-user usability	Correctness of system's conflict resolution
	Percentage of conflicts resolved by the user(s)
Implicit interactions	Appropriateness of system responses to implicit interactions

Table 34. UXAmI metrics to be assessed through questionnaires shortly after a user-based study

UNOBTRUSIVENESS	
Embedment	The system and its components are appropriately embedded in the surrounding architecture
ADAPTABILITY AND ADAPTIVITY	
Appropriateness of recommendations	User satisfaction by system recommendations (appropriateness, helpfulness / accuracy)
USABILITY	
User satisfaction	Users believe that the system is pleasant to use
APPEAL AND EMOTIONS	
Aesthetics	The AmI environment and its systems are aesthetically pleasing for the user
Fun	Interacting with the AmI environment is fun
Actionable emotions	Users' affective reaction to the system
TECHNOLOGY ACCEPTANCE AND ADOPTION	
System attributes	Perceived usefulness
	Perceived ease of use
	Trialability
	Relative advantage

	Cost (installation, maintenance)
Facilitating conditions	End-user support
Expected outcomes	Perceived benefit
	Long-term consequences of use
	Observability
	Image
Trust	User trust towards the system

Table 35. UXAmI metrics to be assessed through interviews shortly after a user-based study

UNOBTRUSIVENESS	
Embedment	The system and its components are appropriately embedded in the surrounding architecture
ADAPTABILITY AND ADAPTIVITY	
Appropriateness of recommendations	User satisfaction by system recommendations (appropriateness, helpfulness / accuracy)
USABILITY	
Accessibility	The AmI environment is physically accessible

Table 36. UXAmI metrics to be assessed through questionnaires long after a user-based study

USABILITY	
User satisfaction	Users believe that the system is pleasant to use
APPEAL AND EMOTIONS	
Aesthetics	The AmI environment and its systems are aesthetically pleasing for the user
Fun	Interacting with the AmI environment is fun
Actionable emotions	Users' affective reaction to the system
TECHNOLOGY ACCEPTANCE AND ADOPTION	
System attributes	Perceived usefulness

	Perceived ease of use	
	Relative advantage	
Facilitating conditions	End-user support	
Expected outcomes	Perceived benefit	
	Long-term consequences of use	
	Observability	
	Image	
Trust	User trust towards the system	

APPENDIX D – UXAMI CONCEPTS AND METRICS REFERENCES

This section lists all the metrics and concepts of the UXAmI framework, referencing to frameworks or theories from which the metrics were derived, and indicating which are the novel metrics introduced by UXAmI.

Table 37. UXAmI framework and metrics with indication of novel metrics and references to related sources

INTUITIVENESS		
Awareness of application capabilities **Scholtz & Consolvo (2004)**	Functionalities that have been used for each system	Novel metric in UXAmI
	Undiscovered functionalities of each system	Novel metric in UXAmI
Awareness of the interaction vocabulary **Scholtz & Consolvo (2004)**	Percentage of input modalities used	Novel metric in UXAmI
	Erroneous user inputs (inputs that have not been recognized by the system) for each supported input modality	Novel metric in UXAmI
	Percentage of erroneous user inputs per input modality	Novel metric in UXAmI
UNOBTRUSIVENESS		
Distraction **Scholtz & Consolvo (2004)**	Number of times that the user has deviated from the primary task	Novel metric in UXAmI
	Time elapsed from a task deviation until the user returns to the primary task	Scholtz & Consolvo (2004)
Embedment **Connelly (2007)**	The system and its components are appropriately embedded in the surrounding architecture	Carvalho et al. (2017)
ADAPTABILITY AND ADAPTIVITY		
Input (sensor) data	Accuracy of input (sensor) data perceived by the system	Paramythis et al. (2010)
Interpretations	Validity of system interpretations	Paramythis et al. (2010)
Appropriateness of adaptation	Interaction modalities are appropriately adapted according to the user profile and context of use	Novel metric in UXAmI

Paramythis et al. (2010)	System output is appropriately adapted according to the user profile and context of use	Novel metric in UXAmI
	Content is appropriately adapted according to the user profile and context of use	Novel metric in UXAmI
	Adaptations that have been manually overridden by the user	Novel metric in UXAmI
Adaptation impact	Number of erroneous user inputs (i.e., incorrect use of input commands) once an adaptation has been applied	Novel metric in UXAmI
	Number of erroneous user interactions once an adaptation has been applied	Novel metric in UXAmI
	Percentage of adaptations that have been manually overridden by the user	Novel metric in UXAmI
Appropriateness of recommendations **Pu et al. (2011)**	The system adequately explains any recommendations	Pu et al. (2011)
	The system provides an adequate way for users to express and revise their preferences	Pu et al. (2011)
	Recommendations are appropriate for the specific user and context of use	Novel metric in UXAmI
	Recommendations that have not been accepted by the user	Novel metric in UXAmI
	Percentage of accepted system recommendations	Novel metric in UXAmI
	User satisfaction by system recommendations (appropriateness, helpfulness / accuracy)	Pu et al. (2011)
USABILITY		
Conformance with guidelines	The user interfaces of the systems comprising the AmI environment conform to relevant guidelines	Nielsen (1994a)
Effectiveness **Nielsen (1994b)**	Task success	Hussain & Kutar (2009)
	Number of interaction errors	Novel metric in UXAmI
	Number of input errors	Novel metric in UXAmI

	Number of system failures	Carvalho et al. (2015)
Efficiency **Nielsen (1994b)**	Task time	Nielsen (1994b), Hussain & Kutar (2009)
	Number of help requests	Nielsen (1994b), Carvalho et al. (2017)
	Time spent on errors	Nielsen (1994b), Carvalho et al. (2017)
Learnability **Nielsen (1994b)**	Users can easily understand and use the system	Wharton (1994)
	Number of interaction errors over time	Novel metric in UXAmI
	Number of input errors over time	Novel metric in UXAmI
	Number of help requests over time	Novel metric in UXAmI
Accessibility **Bevan (2009a)**	The system conforms to accessibility guidelines	Bevan (2009a)
	The systems of the AmI environment are electronically accessible	Mourouzis et al. (2006) Bevan (2009a)
	The AmI environment is physically accessible	Mourouzis et al. (2006)
Physical UI	The system does not violate any ergonomic guidelines	Heo et al. (2009)
	The size and position of the system is appropriate for its manipulation by the target user groups	Heo et al. (2009)
User satisfaction **Nielsen (1994b)**	Users believe that the system is pleasant to use	Nielsen (1994b)
	Percent of favourable user comments / unfavourable user comments	Nielsen (1994b)
	Number of times that users express frustration	Nielsen (1994b)
	Number of times that users express clear joy	Nielsen (1994b)
Cross-platform usability **Wäljas et al. (2010)**	After switching device: time spent to continue the task from where it was left	Novel metric in UXAmI
	After switching device: number of interaction errors until task completion	Novel metric in UXAmI
	Consistency among the user interfaces of the individual systems	Wäljas et al. (2010)

	Content is appropriately synchronised for cross-platform tasks	Wäljas et al. (2010)
	Available actions are appropriately synchronised for cross-platform tasks	Wäljas et al. (2010)
	Help requests after switching devices	Novel metric in UXAmI
	Cross-platform task success compared to the task success when the task is carried out in a single device (per device)	Novel metric in UXAmI
	Cross-platform task time compared to the task time when the task is carried out in a single device (per device)	Novel metric in UXAmI
Multi-user usability	Number of collisions with activities of others	Scholtz & Consolvo (2004)
	Correctness of system's conflict resolution	Novel metric in UXAmI
	Percentage of conflicts resolved by the system	Scholtz & Consolvo (2004)
	Percentage of conflicts resolved by the user(s)	Novel metric in UXAmI
	Social etiquette is followed by the system	Aarts & de Ruyter (2009)
Implicit interactions	Implicit interactions carried out by the user	Novel metric in UXAmI
	Number of implicit interactions carried out by the user	Novel metric in UXAmI
	Percentages of implicit interactions per implicit interaction type	Novel metric in UXAmI
	Appropriateness of system responses to implicit interactions	Novel metric in UXAmI
Usage	Global interaction heat map: number of usages per hour on a daily, weekly and monthly basis for the entire AmI environment	Novel metric in UXAmI
	Systems' interaction heat map: number of usages for each system in the AmI environment per hour on a daily, weekly and monthly basis	Novel metric in UXAmI

	Applications' interaction heat map: number of usages for each application in the AmI environment per hour on a daily, weekly and monthly basis	Novel metric in UXAmI
	Time duration of users' interaction with the entire AmI environment	Novel metric in UXAmI
	Time duration of users' interaction with each system of the AmI environment	Novel metric in UXAmI
	Time duration of users' interaction with each application of the AmI environment	Novel metric in UXAmI
	Analysis (percentage) of applications used per system (for systems with more than one applications)	Novel metric in UXAmI
	Percentage of systems to which a pervasive application has been deployed, per application	Novel metric in UXAmI
APPEAL AND EMOTIONS		
Aesthetics	The systems follow principles of aesthetic design	Nielsen (1994a)
	The AmI environment and its systems are aesthetically pleasing for the user	Adikari et al. (2011)
Fun	Interacting with the AmI environment is fun	Harpur & De Villiers (2015)
Actionable emotions	Detection of users' emotional strain through physiological measures, such as heart rate, skin resistance, blood volume pressure, gradient of the skin resistance and speed of the aggregated changes in the all variables' incoming data.	Cowley et al. (2016)
	Users' affective reaction to the system	Harpur & De Villiers (2015)
SAFETY AND PRIVACY		
User control	User has control over the data collected	Metsis et al. (2008)
	User has control over the dissemination of information	Scholtz & Consolvo (2004)
	The user can customise the level of control that the AmI environment has: high (acts on behalf of the	Novel metric in UXAmI

	person), medium (gives advice), low (executes a person's commands)	
Privacy	Availability of the user's information to other users of the system or third parties	Carvalho et al. (2017)
	Availability of explanations to a user about the potential use of recorded data	Carvalho et al. (2017)
	Expressiveness of the security (privacy) policy	Carvalho et al. (2017)
Safety	The AmI environment is safe for its operators	Bevan (2009a)
	The AmI environment is safe in terms of public health	Bevan (2009a)
	The AmI environment does not cause environmental harm	Bevan (2009a)
	The AmI environment will not cause harm to commercial property, operations or reputation in the intended contexts of use	Bevan (2009a)
TECHNOLOGY ACCEPTANCE AND ADOPTION		
System attributes	Perceived usefulness	Venkatesh & Davis (2000)
	Perceived ease of use	Venkatesh & Davis (2000)
	Trialability	Rogers (1995)
	Relative advantage	Rogers (1995)
	Cost (installation, maintenance)	Brown & Venkatesh (2005)
User attributes	Self-efficacy	Brown & Venkatesh (2005)
	Computer attitude	Heerink et al. (2010)
	Age	Venkatesh et al. (2012)
	Gender	Venkatesh et al. (2012)
	Personal innovativeness	Zarmpou et al. (2012)
Social influences	Subjective norm	Brown & Venkatesh (2005)
	Voluntariness	Venkatesh et al. (2003)
Facilitating conditions	End-user support	Moores (2012)
	Visibility	Mourouzis et al. (2006)

Expected outcomes	Perceived benefit	Venkatesh et al. (2003)
	Long-term consequences of use	Brown & Venkatesh (2005)
	Observability	Venkatesh & Davis (2000)
	Image	Brown & Venkatesh (2005)
Trust	User trust towards the system	Zarmpou et al. (2012);

APPENDIX E – LOW-FIDELITY PROTOTYPES OF THE UXAMI OBSERVER

This section includes all the low-fidelity prototypes that have been created for the UXAmI Observer tool.
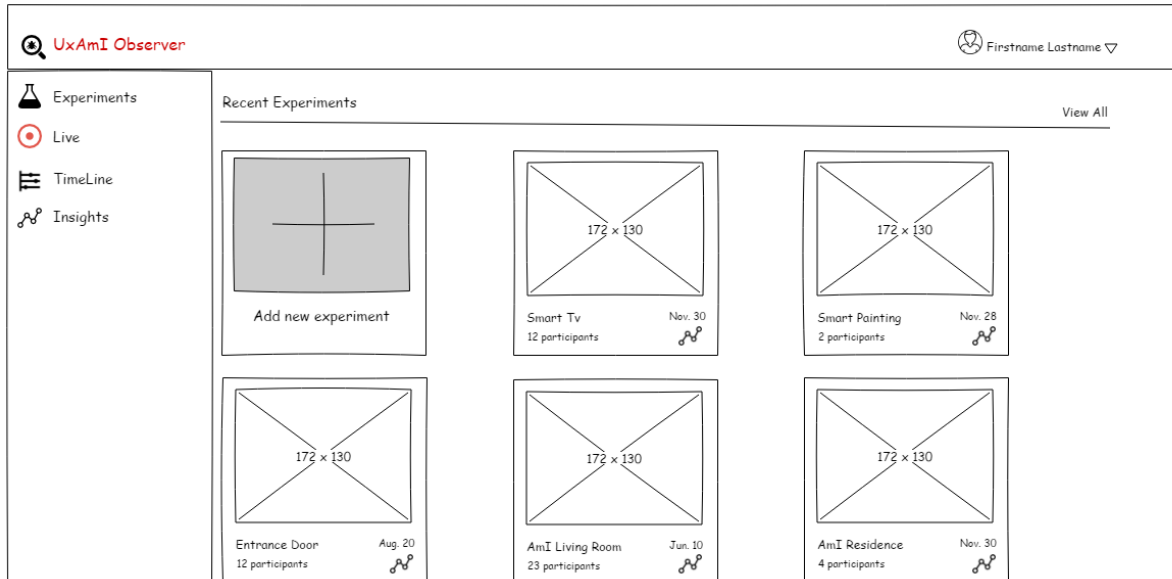
Figure 100. UXAmI Observer Lo-Fi: Home Screen

Figure 101. UXAmI Observer Lo-Fi: Create a new experiment (step 1)

Figure 102. UXAmI Observer Lo-Fi: Create a new experiment (step 2)

Figure 103. UXAmI Observer Lo-Fi: Create a new experiment (step 3)

Figure 104. UXAmI Observer Lo-Fi: Create a new experiment (step 4)
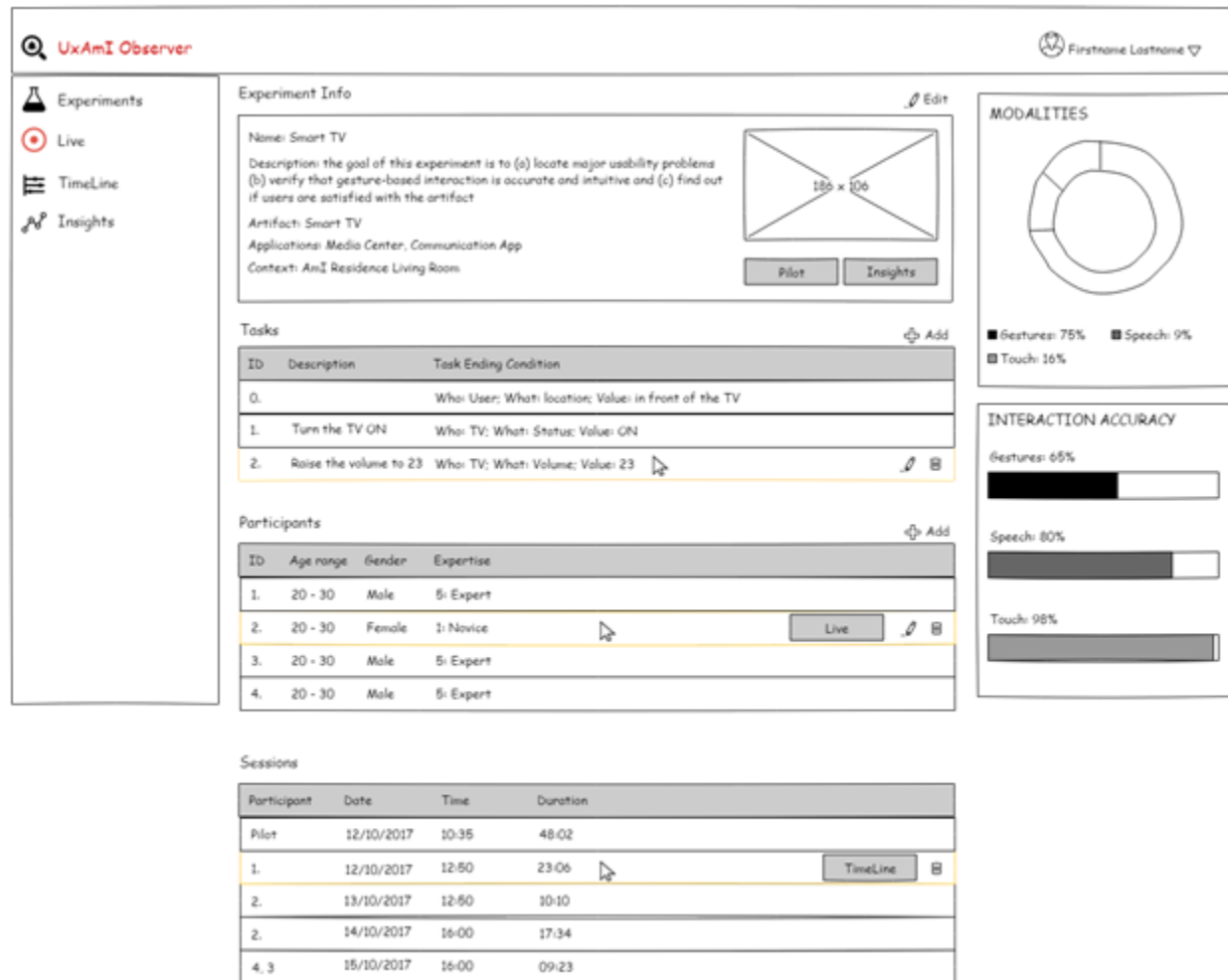
Figure 105. UXAmI Observer Lo-Fi: Experiment details
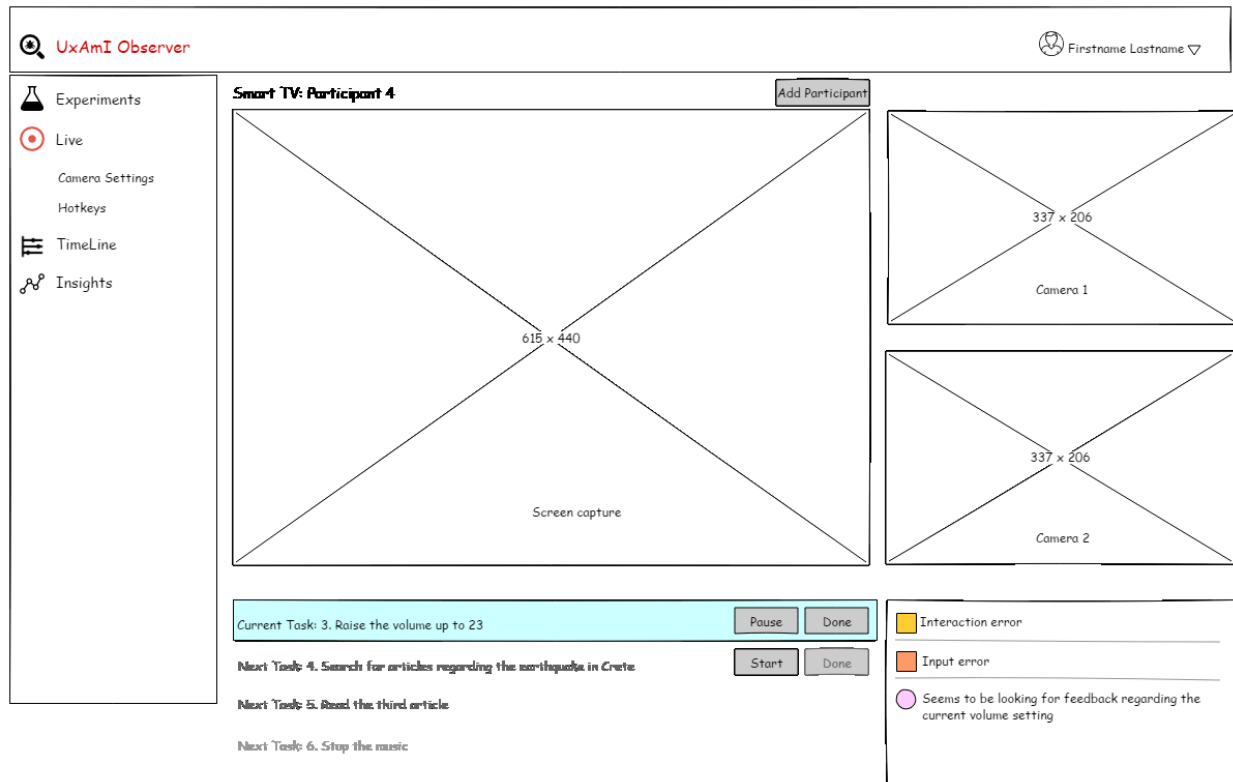
Figure 106. UXAmI Observer Lo-Fi: Live experiment view
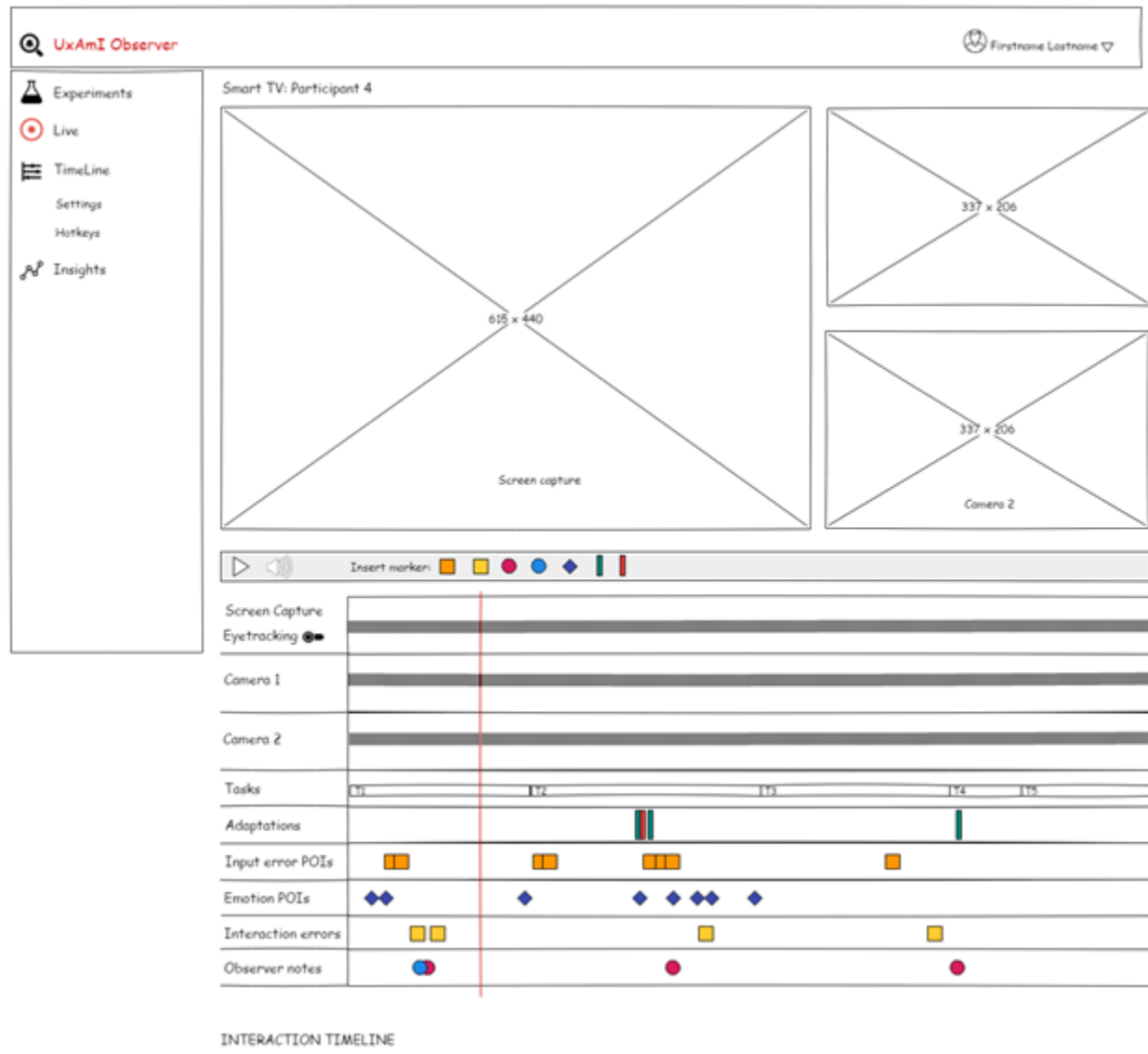
Figure 107. UXAmI Observer Lo-Fi: Timeline view (part A)
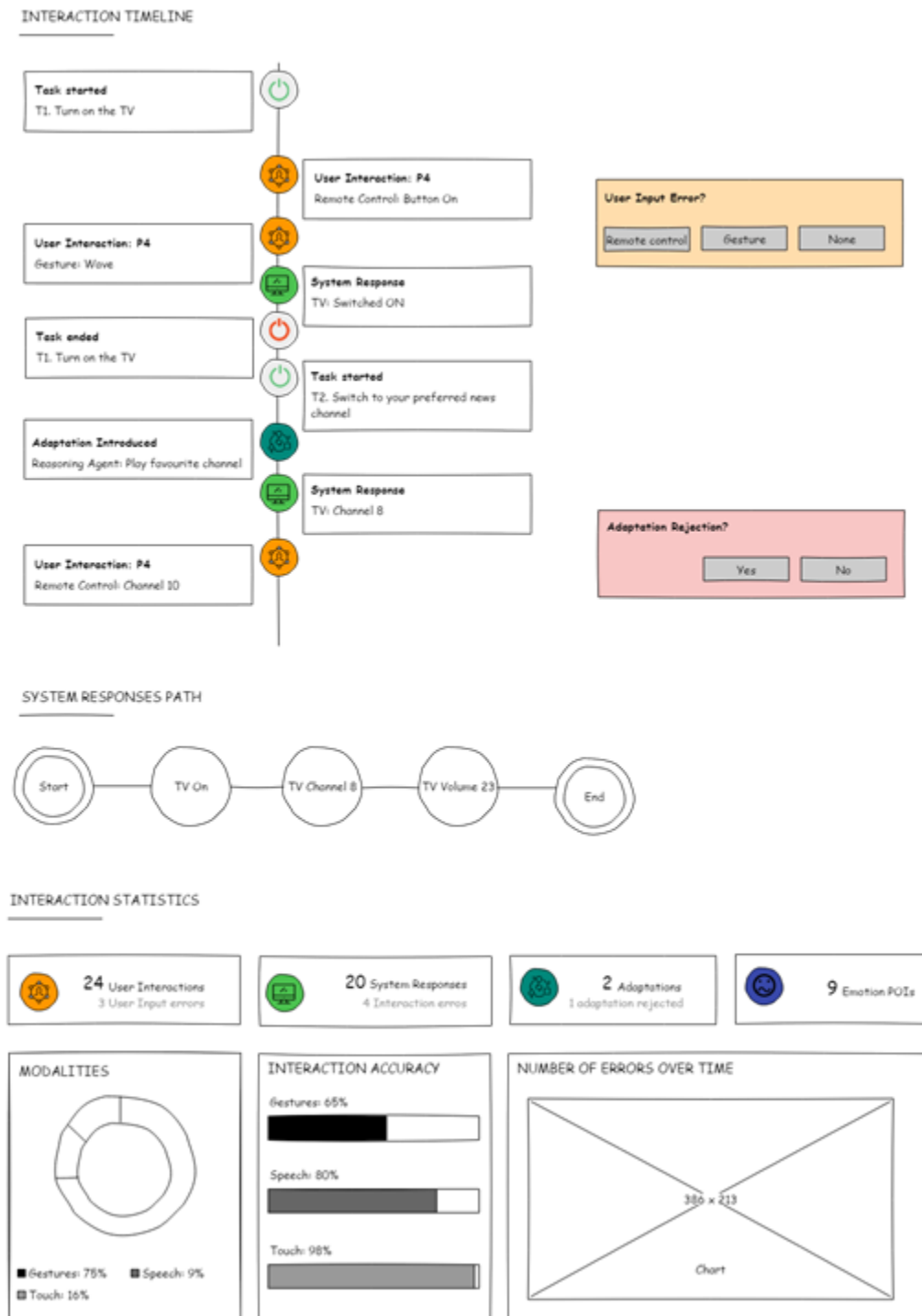
Figure 108. UXAmI Observer Lo-Fi: Timeline view (part B)
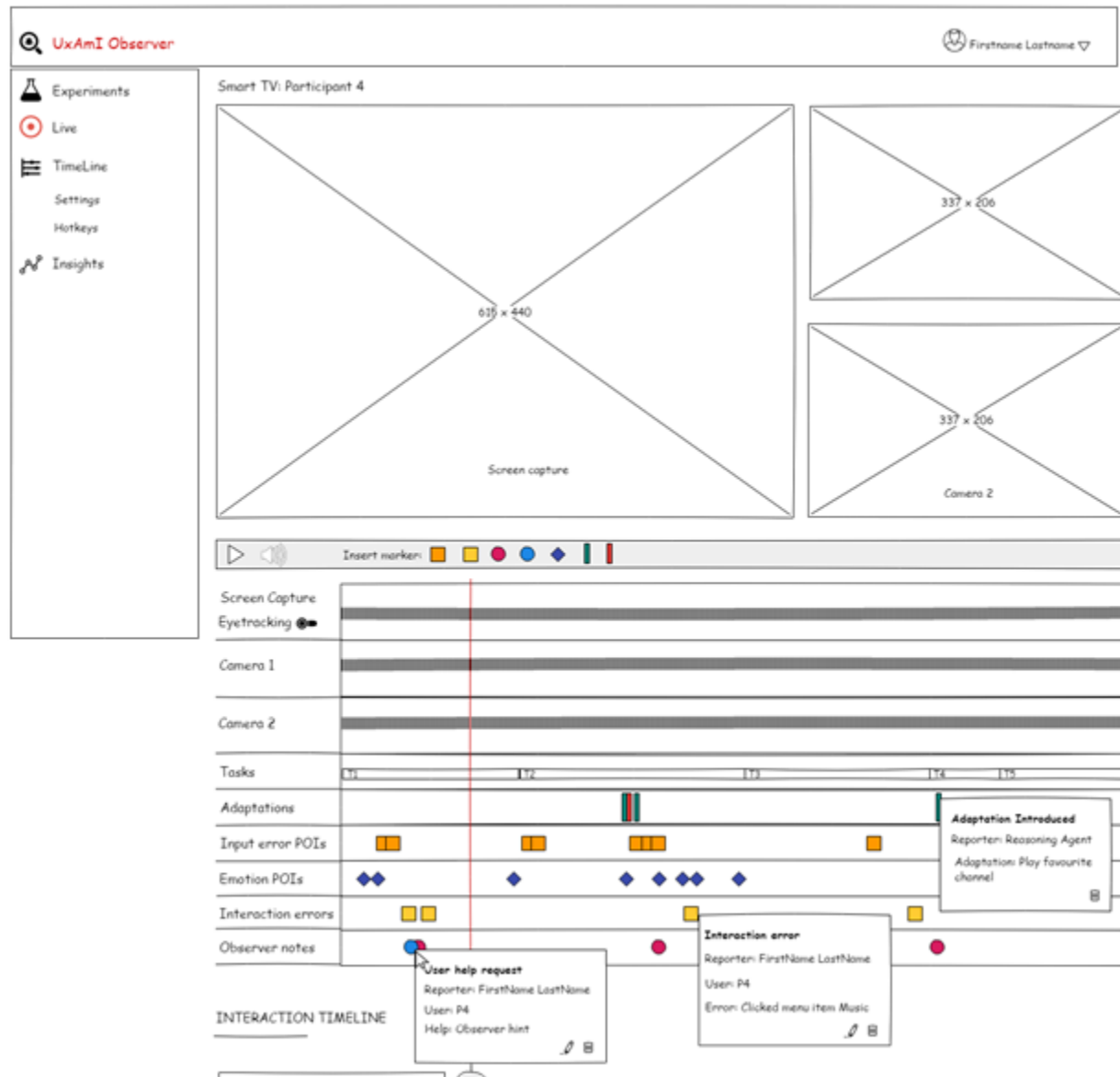
Figure 109 UXAmI Observer Lo-Fi: Timeline view – popups for Points of Interest
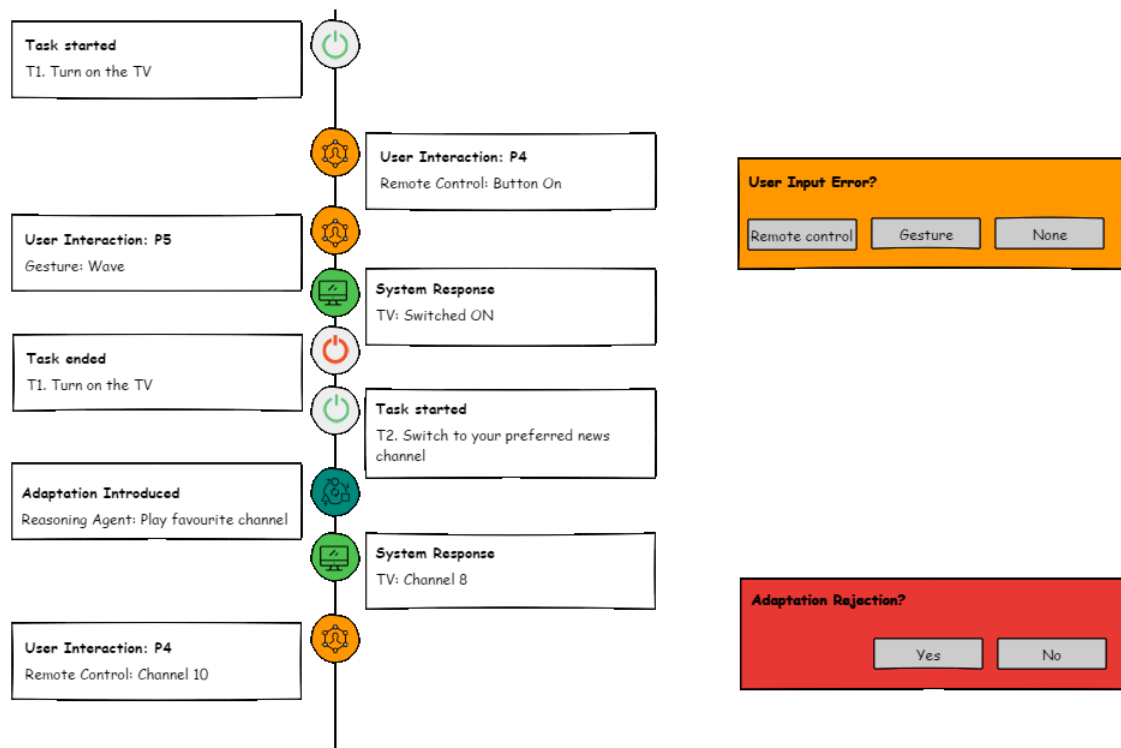
Figure 110. UXAmI Observer Lo-Fi: Timeline view – Interaction Timeline for multiple users

INTERACTION STATISTICS

Overall

24 User Interactions / 3 User Input errors	20 System Responses / 4 Interaction erros	2 Adaptations / 1 adaptation rejected	9 Emotion POIs

Participant 4

20 User Interactions / 3 User Input errors	20 System Responses / 4 Interaction erros	2 Adaptations / 1 adaptation rejected	6 Emotion POIs

Participant 5

4 User Interactions / 0 User Input errors	20 System Responses / 0 Interaction erros	2 Adaptations / 0 adaptation rejected	3 Emotion POIs

Overall

MODALITIES
- Gestures: 75%
- Speech: 9%
- Touch: 16%

INTERACTION ACCURACY
- Gestures: 65%
- Speech: 80%
- Touch: 98%

NUMBER OF ERRORS OVER TIME
386 x 213
Chart

Participant 4

MODALITIES
- Gestures: 75%
- Speech: 25%

INTERACTION ACCURACY
- Gestures: 85%
- Speech: 100%

NUMBER OF ERRORS OVER TIME
386 x 213
Chart

Participant 5

MODALITIES
- Touch: 100%

INTERACTION ACCURACY
- Touch: 100%

NUMBER OF ERRORS OVER TIME
386 x 213
Chart

Figure 111. UXAmI Observer Lo-Fi: Timeline view – Interaction statistics for multiple users

Figure 112 UXAmI Observer Lo-Fi: Insights view for long term usage (no task-based scenario) – Part A
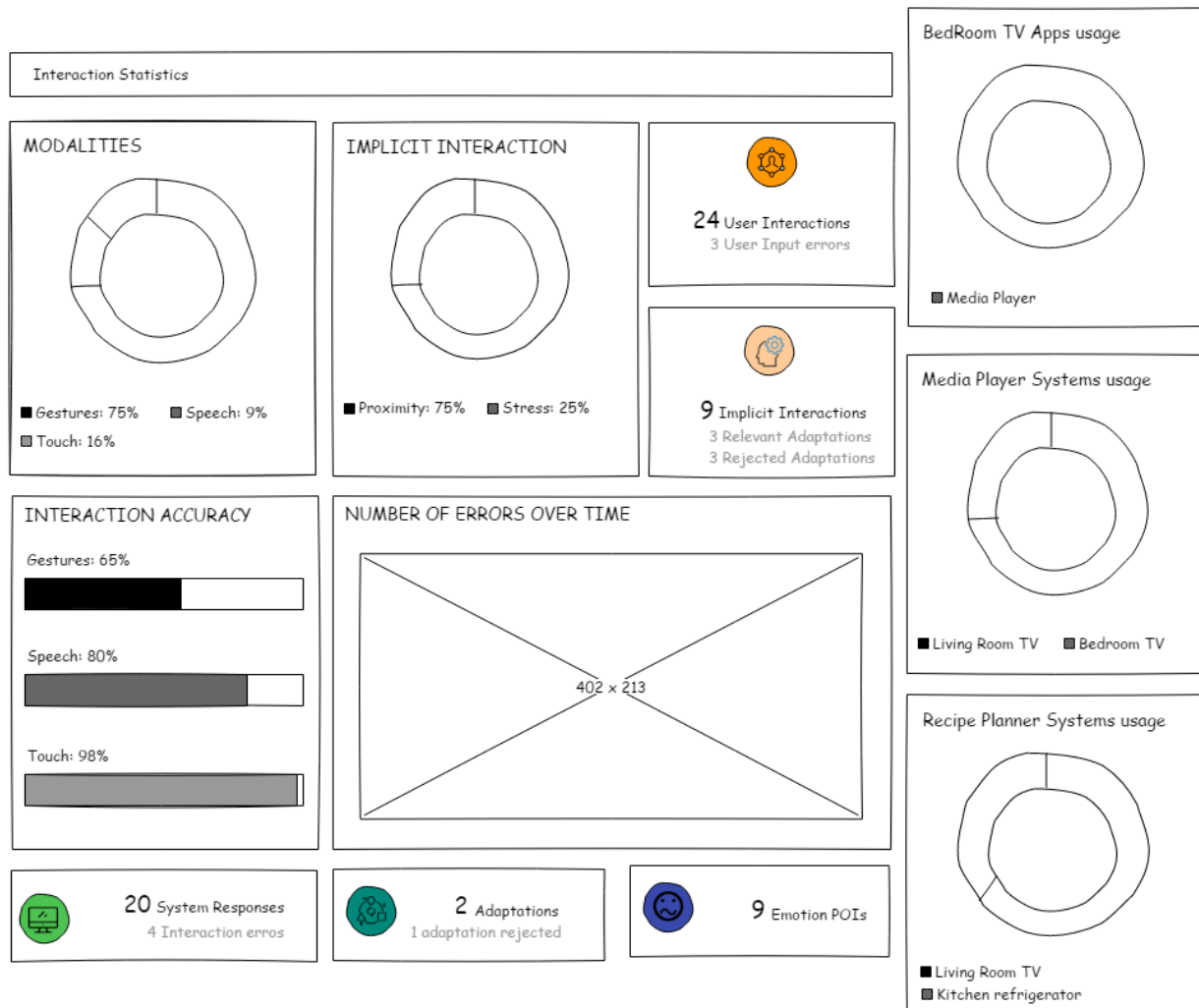
Figure 113. UXAmI Observer Lo-Fi: Insights view for long term usage (no task-based scenario) – Part B
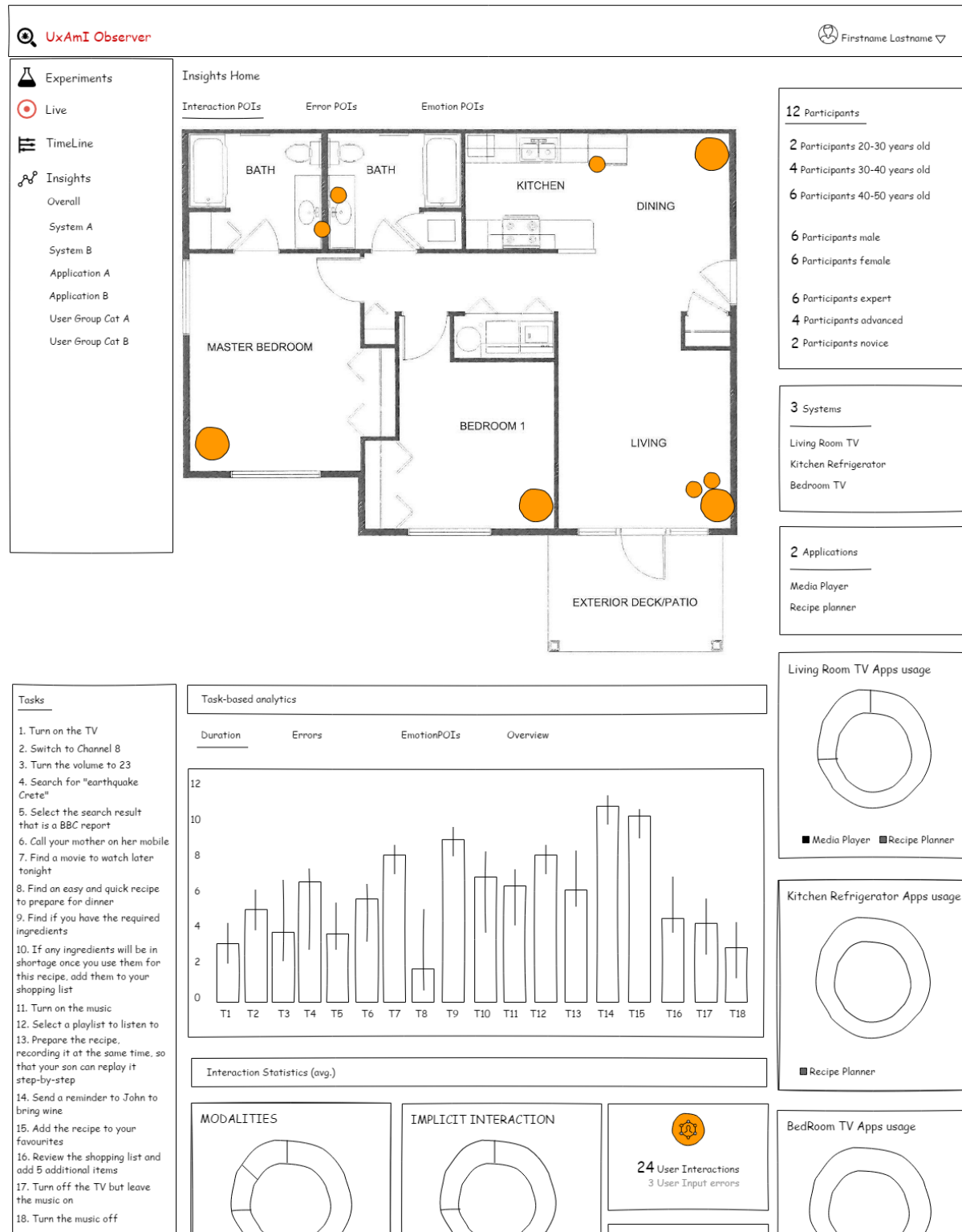
Figure 114. UXAmI Observer Lo-Fi: part of the insights view for task-based scenario – task duration

Figure 115. UXAmI Observer Lo-Fi: part of the insights view for task-based scenario – analytics overview

APPENDIX F – PAPER PROTOTYPES OF THE UXAMI INSPECTOR

This section includes the paper prototypes that have been developed as initial designs of the three-step process for organizing a new inspection in the UXAmI Inspector tool.

Figure 116. UXAmI Inspector paper prototype: Create new inspection paper prototype step 1

Figure 117. UXAmI Inspector paper prototype: Create new inspection paper prototype of step 2a (following the space-oriented approach)

Figure 118. UXAmI Inspector paper prototype: Create new inspection paper prototype of step 2a (following the pervasive application approach)

Figure 119. UXAmI Inspector paper prototype: Create new inspection paper prototype of step 3