Ref. Ares(2019)6615710 - 25/10/2019

DELIVERABLE 3.3

Metrics to validate Eco-Bot engagement actions and proposed measures

adelphi

October, 2019

www.eco-bot.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 767625

D3.3: Metrics to validate Eco-Bot engagement actions and proposed measures
Summary

This deliverable D3.3. provides objective and subjective parameters in a metrics table to validate the effectiveness and acceptance of Eco-Bot as an energy saving tool in the business and residential sector. The basis for the metrics is the Key Performance Indicators (KPIs) as stated in the proposal (Section 1.1 and Section 2.1.). Furthermore, literature research and input from all the project partners fed into the development of additional parameters evaluating the multi-facet performance of Eco-Bot. The present version is an update of the deliverable following the project review in April 2019.

DELIVERABLE NUMBER	WORK PACKAGE
D3.3	WP3
LEAD BENEFICIARY	**DELIVERABLE AUTHORS**
adelphi	Lena Domröse (adelphi) Kathrin Kohl (adelphi) Bianca Llerandi (adelphi) Nanne Zwagermann (adelphi) Aggeliki Giakoumaki (RISA)
QUALITY ASSURANCE	
Reviewer 1: Farah Cheaib Reviewer 2: Claudia Julius	EYPESA SEC
PLANNED DELIVERY DATE	**ACTUAL DELIVERY DATE**
22/03/2019	22/10/2019
DISSEMINATION LEVEL	x PU = Public □ PP = Restricted to other programme participants □ CO = Confidential, only for members of the consortium

Table of contents

List of Tables

List of Acronyms and Abbreviations

B2C: business-to-consumer

B2B2C: business-to-business-to-consumer

B2B: business-to-business

DoW: Description of Work

KPI: Key Performance Indicators

NILM: Non-Intrusive Load Monitoring

WP: Work Package

Executive summary

This deliverable D3.3 provides objective and subjective parameters in a metrics table to validate the effectiveness and acceptance of Eco-Bot as an energy saving tool in the business and residential sector. The basis for the metrics is the Key Performance Indicators (KPIs) as stated in the proposal (Section 1.1 and Section 2.1.). Furthermore, literature research and input from all the project partners fed into the development of additional parameters evaluating the multi-facet performance of Eco-Bot. The present version is an update of the deliverable following the project review in April 2019.

The metrics are composed of five sub evaluation metrics, namely the Behavioural model evaluation metrics, the NILM evaluation metrics, the Chatbot evaluation metrics, the Eco-Bot impact metrics and the Pilot specific metrics. There is a total of 38 parameters, which are presented in overview tables, indicating the respective reference number and giving more information, such as the parameter description, unit of evaluation, measurement method, time frame of evaluation and of presentation of the results, further utilization of results and linkage to the DoWs KPIs and targets, as well as adapted and/or additionally created targets.

1. Introduction

This deliverable presents the metrics developed to evaluate the effectiveness and acceptance of Eco-Bot as an energy saving tool in the business and residential sector. It will set the basis to evaluate theoretical, technical, economic and environmental aspects of the project.

The evaluation that will be performed within the three pilots of WP5 will assess the quality of Eco-Bot, the thoroughness of the theoretical framework, the performance of the technical aspects and also the success in terms of engaging users and achieving energy savings. This evaluation will be invaluable in improving Eco-Bot and create the basis for future projects building on the knowledge gained and lessons learned. It will also feed into the exploitation activities in WP7 and provide helpful information/ input for the communication and dissemination activities toward the end of the project under WP6.

This deliverable is structured as follows: Chapter 2 presents the aim of this deliverable, Chapter 3 the methodology and Chapter 4 the Eco-Bot evaluation metrics. Within Chapter 4 each group of parameters (sub-metric) has its own sub-chapters giving more information on the provenance and relevance of the respective parameters. Finally, Chapter 5 gives an outlook on the further enabled group level evaluations/statistics.

2. Aim of this deliverable

The aim of this deliverable is to define the evaluation metrics, i.e. operationalize the KPIs of Section 1.1. and Section 2.1. in the DoW and decompose the different aspects of Eco-Bot to identify further parameters. This is undertaken to evaluate crucial aspects of the project based on the outcomes of the three pilots:

1) **Residential consumers of Estabanell in Spain (B2C):**

Estabanell is an electricity provider (utility). This use case will demonstrate how delivering personalized information on appliance-level and relevant energy efficiency tips can affect the behaviour of utility customers. Estabenell has set up a separate webpage for eco-bot, where the user can login, register and reach the bot.

2) **Residential consumers of Senercon/co2online in Germany (B2B2C):**

Senercon, together with co2online, reaches residential energy consumers directly via its energy monitoring software. The consumers can register an energy savings account, in order to track their energy use. Eco-bot will be integrated directly into the energy monitoring software iESA.

3) **Facility/Energy managers of Dexma in Spain and the United Kingdom (B2B):**

Dexma is a building energy management system provider in Spain. It is a leading European Software-as-a-Service platform focused on energy efficiency with over 260 active partners (building managers / energy managers). Eco-bot will be integrated directly into the energy monitoring software DEXCELL.

This deliverable is the basis for the evaluation to be performed in WP5 and serves as valuable source of information for WP6 and WP7.

3. Methodology

The aim of this deliverable is to identify relevant parameters to evaluate the effectiveness and acceptance of Eco-Bot as a personalized virtual energy assistant.

The deliverable is founded on three pillars:

1) Operationalization of the KPIs (Section 1.1. and Section 2.1 in the DoW)

2) Input by the consortium: Each partner gave input and feedback on the parameters that correspond to their field of expertise. Accordingly, KAT was mainly involved in the establishment of the parameters that evaluate the behavioural model, USTRAT in the ones that evaluate the NILM and itemized billing services, ERRA, RISA and PLEGMA in the ones that evaluate the chatbot itself and the pilots (DEXMA; EYPESA and SEN) in the ones regarding the impact of the Eco-Bot on an individual and group level.

3) Literature research on chatbot evaluation metrics.

In the beginning, the KPIs (Section 1.1. and Section 2.1. of the proposal) were operationalized in a first draft[1]. Further information on each parameter was provided through additional information, such as a clear definition, unit of measurement and evaluation along with examples/clarifications and targets, if specified in the proposal. This was done to ensure the comprehensibility of the metrics.

Establishment of the five sub evaluation metrics

Eco-Bot builds on the behavioral model that was created, the energy analytics and disaggregation algorithms that were performed and the chat-bot that was designed. Accordingly, three sub evaluation metrics were established to account for the **theoretical and technical framework of Eco-Bot**, namely:

- the Behavioural model evaluation metrics, consisting of 2 Parameters (P1 – P2)
- the NILM evaluation metrics, consisting of five parameters (P3-P7)
- and the Chatbot evaluation metrics, consisting of 15 Parameters

Two sub metrics were created to **assess the impact/ the outcome of Eco-Bot**, namely:

- the Eco-Bot impact evaluation metrics, consisting of 9 Parameters

[1] Please note that KPI_1.1: *The number of policies, strategies, scientific papers and energy efficiency models examined together with the number of identified factors relevant for creating the model of customer behaviour for Eco-Bot purposes* and KPI_2.1: *Number of target groups (segments) successfully identified within the use cases and mapped on the behavioural model that was identified as the most relevant for Eco-Bot*, from Section 1.1. of the DoW were operationalized in the taxonomy related parameters. They will, however, only appear in the annex, since the creation of the taxonomy preceded D3.3. (see D2.3 and D3.2).

- the Pilot specific evaluation metrics, consisting of 7 Parameters

Choosing an adequate evaluation tool

When completing the work on the metrics, it was crucial to identify the right tool and timeframe for each measurement. In close cooperation with each partner regarding their respective field of expertise, the unit of measurement and evaluation, as well as the tool (how to measure) and time frame were defined. Keeping in mind that the user should not be discouraged by having to answer too many questions at once, we used balanced tools. Questions will be asked via a user survey (more detailed explanation will be given in D5.1), feedback mechanisms of the bot itself and other means of evaluation, such as data analysis of the backend. This practical task resulted in the final refinement of the metrics.

4. The Eco-Bot evaluation metrics

This chapter presents each of the five sub metrics and the respective parameters. More information on the provenance and relevance of the parameters is given. Each parameter is presented in an overview table to enhance intelligibility and readability of the Evaluation metrics. The structure of the overview table is always the same. The header names the parameter and attributes a reference number to it (e.g. P1), the parameter is described in the first row and the second row indicates the unit, in which the parameter will be evaluated. The third row outlines the measurement method and the fourth row specifies the time of evaluation, followed by the fifth row indicating the time of presentation of the evaluation results. Rows six and seven demonstrate the relevance of the parameter by giving insights into the further utilization of the results and indicating the linkage to the DoW's KPIs and targets. The last row indicates adapted and/or additionally created targets.

4.1 The Behavioural model evaluation metrics

This section presents the metrics that will evaluate the behavioural model (as described in D3.2.) and the usefulness of the tailored recommendations given by Eco-Bot. There are a few KPIs that relate to the thoroughness of the preparation of the model, which were operationalized in parameters and can be found in the annex. Given that the model and the recommendations were already established in the work steps preceding D3.3 (see D2.3 and D3.2.), the focus of this deliverable lies on the evaluation of the recommendations and the accuracy of the classification model (i.e. correct allocation of users to the segments).

The first parameter P1, assesses the usefulness of the recommendations.

Table 1: P1 - Recommendations usefulness

P1 - Recommendations usefulness	
Parameter Description	This indicator expresses the percentage of recommendations found useful by the users.
Unit	%
Measurement Method	After receiving a recommendation, each user will be asked by the bot to evaluate its usefulness. (Recommendations that users have already implemented before they obtained it via the bot and are commented with: "I have already implemented it"; are evaluated as useful, as well.) If the user indicates a recommendation as not useful, he/she will be asked by the bot to identify the reason through the selection of one of the following options: • "I used a similar solution"; • "I do not like it"; • "It limits my comfort"; • "Not applicable". Those responses will be recorded and used to improve (optimize) the recommendations during the assessment/evaluation phase. The following scale will be adopted: – Under 30% of recommendations found useful: prepared recommendations are unsuitable and thus the entire recommendation set should be changed. – 31-60% of recommendations found useful: The set of recommendations should be significantly changed e.g. by removing unnecessary recommendations, proposing new solutions and better adapting them to individual segments. – 61-80% of recommendations found useful: Selected recommendations from the set will be rephrased and/or more specified. – Above 80% of recommendations found useful: The recommendations remain unchanged.
Time of evaluation	It will be monitored on an ongoing basis, starting from the small-scale validation phase (M26-M28) and until the completion of the pilot phase (M40). The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation (M28), b) in the middle of the pilots (M34), and c) at the end of the pilots (M40). The preliminary evaluation that will take place after the completion of the 3-month small-scale validation will allow a first refinement of the recommendations if found necessary, and according to the evaluation results that will be available after the first 6 months of the demonstration activities of WP5 (M34), a more comprehensive assessment will take place, so as to improve recommendations – if needed – for the second half of the demonstration activities. The overall assessment will take place after the completion of the pilots (M40).
Presentation of evaluation results	The results of the small-scale validation and of the first 6 months of the pilot, the following analysis and re-modelling will be documented in D3.4 (M43). The final evaluation results will be presented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	As outlined above, according to the percentage of recommendations found useful by the users, the recommendations per segment will be improved and further refined (WP3), and the Eco-Bot system will be updated accordingly (WP4) so as to provide the users with the refined recommendations during the second half of the demonstration activities (WP5). Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.

	Moreover it will be useful for similar future projects to build on the experience.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	Eco-Bot users found more than 60% of recommendations useful

The second parameter assess the accuracy of the statistical model. The statistical classification model classifies users into five main segments.

Table 2: P2 - Accuracy of the classification model

P2 – Accuracy of the classification model	
Parameter Description	This indicator evaluates the quality of the classification model. It measures the percentage of correct predictions made by the classification model. The term "correct prediction" means that it is identical to the actual segment the user belongs to.
Unit	%
Measurement Method	Users will be asked to respond to an extended survey, which will also include the 10 initial questions that were selected for the classification model. Based on the results of the survey, a set of data will be created, and each user will be assigned to his/her actual behavioural type (identified by the answers to segmenting questions), which will be treated as the ground truth. The accuracy of the classification model will be evaluated by comparing the user's behavioural type obtained from the classification model with the ground truth and it is defined as the percentage of the correct classification model predictions, i.e. the number of correct predictions (segment allocations of users) made by the classification model divided by the total number of predictions made. Users will be asked to respond to the extended survey again after the completion of the first pilot phase, thus allowing the overall assessment of the final classification model at the end of the project. It should be noted that the above described measurement plan, apart from the evaluation of the model's accuracy, enables also the identification of potential transition of users to different segments in the course of the project. *An extensive and more detailed explanation of the measurement method can be found in D3.2. Chapter 7.*
Time of evaluation	Evaluation of the classification model will take place after the first half of the validation activities of WP5 is completed (M34), so as to revise the classification model – if needed – and to release an updated Eco-Bot version for the remaining part of the pilot phase. The overall assessment of the final classification model will take place after the completion of the pilot phase (M40-43).
Presentation of evaluation results	The first evaluation results (after the first 6 months of the pilot) will be documented in D3.4 (M43) and the final evaluation results will be presented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	If the model turns out to be weak (i.e. less than 80% classification accuracy) it will be revised (WP3). Then it is also likely that the questions, which the model will indicate as the most important, will have to be adapted, and the Eco-Bot system will

P2 – Accuracy of the classification model	
	be updated accordingly (WP4) so as to incorporate the revised classification model for the remaining part of the demonstration activities (WP5). Exploitation (WP7)/Dissemination (WP6): The results will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot. Moreover it will be useful for similar future projects to build on the experience.
Linkage to DoW's KPIs and targets	none
Adapted or additionally created targets	The classification error should not exceed 20%, or equivalently, the classification agreement between the true behavioural type (determined using clustering methods) and the segments to which the users have been assigned by the classification model is not less than 80%.

4.2 The NILM evaluation metrics

This section presents the metrics that will evaluate the NILM and effectiveness and performance of the itemized billing service. This is part of the theoretical and technical framework of Eco-Bot. Evaluation is used to ensure feedback reliability and to maintain consumer trust in NILM results when it regards reasonable estimation of their appliance consumption estimation. The NILM evaluation metric is composed of five parameters. The presented parameters all constitute operationalization's of the KPIs in the DoW. The following overview tables present each of the parameters. More detailed information and the evaluation can be found in D4.2.

Table 3: P3 – Accuracy and limitations of state-of-the-art NILM algorithms at low smart-meter data sampling rates on real datasets at scale

P3 – Accuracy and limitations of state-of-the-art NILM algorithms at low smart-meter data sampling rates on real datasets at scale	
Parameter Description	This parameter measures, qualitatively, the effectiveness and performance of the disaggregation algorithms, and represents the outcome of the literature review in terms of accuracy and implementation to evaluate applicability of state-of-the-art appliance specific disaggregation models to Eco-Bot pilots
Unit	String (qualitative recommendations)
Measurement Method	A peer-reviewed literature research focusing on NILM approaches suggested for low to very low rate disaggregation will identify detailed recommendations for NILM solutions relevant for the Eco-Bot pilots.
Time of evaluation	Task 4.2
Presentation of evaluation results	D4.2

WPs that the results will feed into and further utilization of results	The results of this study will impact selection or new development of the NILM algorithm suitable for the pilots in WP4. Furthermore, the outcomes will indirectly influence the implementation of NILM algorithms for demonstration phases in WP5.
Linkage to DoW's KPIs and targets	From Section 1.1. of the DoW KPI_4.1: Accuracy and limitations of state-of-the-art Non-Intrusive Appliance Load Monitoring (NILM) algorithms at low smart-meter data sampling rates on real datasets at scale
Adapted or additionally created targets	None

Table 4: P4 – Practicality, scalability and near real-time suite of NILM algorithms that yield accurate appliance-specific disaggregation, with little to no training, robust to appliance heterogeneity

P4 – Practicality, scalability and near real-time suite of NILM algorithms that yield accurate appliance-specific disaggregation, with little to no training, robust to appliance heterogeneity	
Parameter Description	This parameter measures the specific implementation aspects of the selected NILM algorithms, i.e., ability of the algorithm to be run at scale on multiple buildings with no a priori training on that particular building's smart meter data, and ability to be run in near real-time for quick response via the bot.
Unit	String (qualitative recommendations)
Measurement Method	The choice of residential and commercial NILM algorithms, following outcomes of KPI_4.1, are evaluated for scalability and ability to provide a result in near real-time using information provided by all three Eco-Bot pilots and historical representative pilot data (if available) and publicly available datasets
Time of evaluation	Task 4.2
Presentation of evaluation results	D4.2
WPs that the results will feed into and further utilization of results	The result of this study will directly influence the implementation of NILM algorithms (WP4) for demonstration phases in WP5.
Linkage to DoW's KPIs and targets	From Section 1.1. of the DoW: KPI_4.2: Practicality, scalability and near real-time suite of NILM algorithms that yield accurate appliance-specific disaggregation, with little to no training, robust to appliance heterogeneity
Adapted or additionally created targets	None

Table 5: P5 – Accuracy in estimated energy consumption

P5 – Accuracy in estimated energy consumption	
Parameter Description	This parameter measures the accuracy in estimated energy consumption compared to the actual one (referred to as consumption accuracy).
Unit	%
Measurement Method	By means of a statistical estimation accuracy metric the consumption accuracy is measured by comparing estimated appliance-level electricity consumption with the true consumption (obtained via submetering).
Time of evaluation	Given that consumption accuracy can be calculated only if submetering is available, for the residential pilots – where submetering is not available – the consumption accuracy results will be produced – in the context of Task 4.2 – using publicly available datasets. For the commercial pilot, submetering data is available in historical data and during field testing, therefore consumption accuracy will be evaluated not only in the context of Task 4.2, but also in the small-scale validation of Task 4.5 and in the validation activities of WP5. Specifically, this parameter will be evaluated (where submetering is available) in four different phases: a) during Task 4.2, b) at the end of the small-scale validation, c) in the middle of the pilots, and d) at the end of the pilots.
Presentation of evaluation results	The evaluation results that will be produced in the context of Task 4.2 will be documented in D4.2, the small-scale validation results will be documented in D4.5 (M28) and the pilot results will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	The result of this study will determine the accuracy of the NILM algorithm for particular appliances and pilots and lessons learned will influence the Updated Front-/Back-End Platform ready for the demonstration activities V1 in D4.5 and similarly Updated Front-/Back-End Platform ready for the demonstration activities V2 in D4.6 to be used in WP5.
Linkage to DoW's KPIs and targets	From Section 1.1. of the DoW: KPI_4.4: Quality of new energy feedback and itemized billing practices based on energy disaggregation and integrating user feedback into NILM algorithm design Target: More than 80% accuracy in estimating energy disaggregated consumption patterns
Adapted or additionally created targets	None. Targets as specified in the DoW (see above)

Table 6: P6 – Learning performance of NILM algorithms with regard to appliance signatures from user feedback

P6 – Learning performance of NILM algorithms with regard to appliance signatures from user feedback	
Parameter Description	This parameter measures the accuracy in appliance detection (referred to as classification accuracy)
Unit	%
Measurement Method	By means of a statistical estimation accuracy metric, appliance detection or classification accuracy will be obtained by asking participants, via the bot, when

	they used an appliance and comparing when NILM estimated they used the appliance.
Time of evaluation	Given that classification accuracy can be calculated only if submetering or other means of ground truth (e.g. time diary) are available, for the residential pilots – where submetering is not available – the classification accuracy results will be produced during the small-scale validation and the pilot phase by comparing the users' feedback in the time diary with the NILM-estimated use of appliances. In Task 4.2, submetering is available for estimating classification accuracy for residential pilots using public datasets. For the commercial pilot, submetering data is available in historical data and during field testing, therefore classification accuracy will be evaluated not only in the context of Task 4.2, but also in the small-scale validation of Task 4.5 and in the validation activities of WP5. Specifically, this parameter will be evaluated (where submetering or time diary – i.e. an alternative means of ground truth – is available) in four different phases: a) during Task 4.2, b) at the end of the small-scale validation (M28), c) in the middle of the pilots (M34), and d) at the end of the pilots (M40).
Presentation of evaluation results	The evaluation results that will be produced in the context of Task 4.2 will be documented in D4.2, the small-scale validation results will be documented in D4.5 (M28) and the pilot results will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	The result of this study will determine the accuracy of the NILM algorithm for particular appliances and pilots and lessons learned will influence the Updated Front-/Back-End Platform ready for the demonstration activities V1 in D4.5 and similarly Updated Front-/Back-End Platform ready for the demonstration activities V2 in D4.6 to be used in WP5.
Linkage to DoW's KPIs and targets	From Section 1.1. of the DoW: KPI_4.3: Learning performance of NILM algorithms with regard to appliance signatures from user feedback.
Adapted or additionally created targets	None, since it is a subjective measurement for residential pilots where submetering is not available and is based on the participant's ability to recall exactly all instances of all appliances that were run. It is a non-subjective parameter for commercial buildings, where submetering is available.

Table 7: P7 – Quality of new energy feedback and itemized billing practices based on energy disaggregation and integrating user feedback into NILM algorithm design

P7 – Quality of new energy feedback and itemized billing practices based on energy disaggregation and integrating user feedback into NILM algorithm design	
Parameter Description	This parameter presents the assessment of suitability of time-diary and gathering user-feedback to replace submetering for the purposes of validation of classification accuracy: that is, suitability of user feedback for estimating detection/classification accuracy vs. submetering which is not practical at scale (i.e. since the reliability of participant's recall of when they ran an appliance is uncertain, this is also evaluated based on how often they respond and how accurate their response is).
Unit	String (qualitative recommendations) or % consumption accuracy (if submetering is available, i.e., in the case of commercial pilots)

Measurement Method	Classification accuracy and consumption accuracy
Time of evaluation	Small-scale pilots
Presentation of evaluation results	D4.5.
WPs that the results will feed into and further utilization of results	Detailed recommendations for NILM algorithm implementation and quality of user feedback for evaluation of NILM performance that will directly influence the Updated Front-/Back-End Platform ready for the demonstration activities V2 in D4.6
Linkage to DoW's KPIs and targets	From Section 1.1. of the DoW: KPI_4.4: Quality of new energy feedback and itemized billing practices based on energy disaggregation and integrating user feedback into NILM algorithm design Target: More than 80% accuracy in estimating energy disaggregated consumption patterns
Adapted or additionally created targets	None. Targets as specified in the DoW (see above).

4.3 The Chatbot evaluation metrics

This section presents the metrics that will be measured in order to evaluate the chatbot. It should be noted that although a growing body of literature has suggested a number of approaches for the evaluation of chatbots, evaluation standards in this field have not yet been established and it is still quite difficult to find approaches that have been widely accepted and adopted. Therefore, this chapter is more extensive, presenting a technical literature discussion and more details on the provenance and relevance of the chosen parameters.

Several existing chatbot evaluation frameworks are based on standards originally defined for use in software or web applications, generalising the metrics tested in different types of systems and overlooking the increasingly complex developments in the field of chatbots (Kaleem et al. 2016). Indicatively, such frameworks include the Software Usability Measurement Inventory (SUMI)[2], which is strongly focused on graphical user interface, and System Usability Scale (SUS)[3], which is a simple ten-item Likert scale focusing on the perceived ease of use and learnability of using a system. The SUS has been used for the evaluation of conversational systems (Hoque et al. 2013; DeVault et al. 2014), however certain items included in the questionnaire do not fit well in the chatbot domain and, additionally, it does not address sufficiently other aspects that are of importance to chatbots, e.g. user experience.

There is a general consensus that usability is of critical importance in chatbots; no matter how important functionality is, without usability the system will not be given the chance to demonstrate functionality (Kaleem et al. 2016). Therefore, besides quality and efficiency metrics, subjective tests must also take place in order to be able to assess the impact of the

[2] http://sumi.uxp.ie/
[3] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

chatbot's capabilities on user satisfaction and to get a valuable insight on both its shortcomings and advantages.

Literature suggests performing evaluation of chatbots by distributing a questionnaire to the users so as to capture their subjective assessment of the chatbot, as well as to examine the user-chatbot conversations and to combine the two evaluation methodologies (Silvervarg & Jönsson 2011). Rauschenberger et al. (2013) propose a framework to measure user experience and software quality in interactive software applications through the User Evaluation Questionnaire (UEQ)[4]. They state that the evaluation of interactive software quality falls into two distinct categories, i.e. pragmatic quality and hedonic quality. Pragmatic quality refers to task-oriented quality like task completion effectiveness and efficiency, while hedonic quality refers to non-task related aspects like aesthetic impressions and user stimulation.

Subjective aspects like user satisfaction are usually evaluated by using questionnaires with Likert-type scale questions, while objective metrics can be measured through records and logs of the user's conversations with the chatbot. These objective metrics can be captured in two different phases: a) during the course of a lab usability testing, where the user is asked to perform a set of predefined tasks, enabling the performance evaluation of the chatbot in terms of effectiveness, efficiency and satisfaction in a specified context of use, and b) during user-chatbot interaction in real conditions, gaining insight into utility and usability aspects that enables system fine-tuning and optimisation. There is a general consensus in the literature that the effectiveness of a chatbot system should be evaluated through a combination of subjective and objective measures, which ensures that not only the chatbot's functionality is assessed, but also the usability from the user's perspective is captured (Kaleem et al. 2016; Rauschenberger et al. 2013).

As mentioned above, though, there is a lack of chatbot evaluation standards, so there is no widely-adopted framework that can be followed to derive the individual metrics that should be tested in order to perform chatbot evaluation. Radziwill & Benton (2017) reviewed chatbot evaluation literature and presented a summary of quality attributes falling under efficiency, effectiveness and satisfaction categories. Other researchers followed different approaches and emphasised on chatbot-specific quality attributes. Indicatively, Jain et al. (2018) focused on conversational intelligence, which refers to the quality of the conversation over and beyond mere functionality; it focuses on the chatbot's understanding of the input text and how it handles failures to understanding user's text. A chatbot should ideally have "human-like" conversational capabilities, including intra-session context preservation, understanding of negative statements, admitting failure, and ability to ask questions to engage the user in a meaningful conversation, along with helping the user complete the task. Kuligowska (2015) also focused on the above attributes, indicating the importance of examining the conversational abilities, language skills and context sensitiveness of a chatbot. According to Kuligowska, the biggest challenge in designing a chatbot architecture is inventing the mechanism of dialogue context detection, which makes the chatbot keeping pace with a constantly changing topic of conversation. Language skills and context sensitiveness require

[4] https://www.ueq-online.org/

the integration of conversational behaviours, such as giving feedback, taking turn, and repairing of dialogue.

The parameters that will be assessed in order to evaluate Eco-Bot were defined by combining different approaches and attributes suggested in chatbot evaluation literature. Our evaluation framework includes both subjective and objective measures. Users will be asked after the completion of the pilot to fill in a user experience questionnaire, which will enable the assessment of both user satisfaction and the usability from the user's perspective. This questionnaire will also be given to test users that will participate in the lab usability testing before Eco-Bot's release to end-users, and will be also filled in by the users that will test Eco-Bot during the small-scale validation phase. Besides the subjective feedback, Eco-Bot will be evaluated in terms of objective metrics that will enable the assessment of its performance, users' engagement and retention, as well as usage patterns, thus facilitating system refinement and optimisation. The following subsections present in detail the metrics that were selected to be measured so as to evaluate Eco-Bot.

4.3.1 The User experience related parameters

User experience is an important aspect in the evaluation of conversational interfaces; however, there is still no consensus in the literature on the definition and evaluation of user experience (Law et al. 2009; Bargas-Avila & Hornbæk 2011). Particularly in the conversational systems literature, the concept of user experience and its relationship with usability is controversial; user experience has been variably defined as: usability, an aspect going beyond usability, an aspect of usability, user satisfaction, and a combination of ease of use, overall feeling and user satisfaction (Kocaballi et al. 2018). Despite the large deviations in the definition, the main tendency nowadays is to consider user experience as a design and evaluation factor that captures usability as part of it and is closely related with user satisfaction.

There are several questionnaires that have been suggested in the literature for the evaluation of conversational interfaces, including – among others – the AttrakDiff, the Subjective Assessment of Speech System Interfaces (SASSI), the Speech User Interface Service Quality (SUISQ), the Mean Opinion Scale (MOS), and the Paradigm for Dialogue Evaluation System (PARADISE). Kocaballi et al. (2018) examined in detail a number of questionnaires including the above and concluded that there is no questionnaire providing sufficient coverage across all user experience dimensions. They recommended, if feasible, to use a combination of multiple questionnaires, in order to obtain a more complete assessment that covers a larger range of user experience dimensions rather than focusing on a particular one.

Research has shown that in order to be able to measure the user experience and satisfaction, it is necessary to consider both pragmatic and hedonic aspects (Rauschenberger et al. 2013). Pragmatic quality refers to the perceived ability to support the achievement of "do-goals", to provide an effective, efficient and easy to use means to perform a task; it focuses on the utility and usability aspects. Hedonic quality, on the other hand, refers to the product's perceived ability to support the achievement of "be-goals" and to support stimulation (Hassenzahl, 2008). Considering both pragmatic and hedonic quality aspects is the underlying idea of the User Experience Questionnaire (UEQ), in which user experience is

conceived as the overall impression of a user when he/she interacts with a product (Laugwitz et al. 2008). The UEQ contains 6 scales with 26 items and evaluates attractiveness, pragmatic quality aspects (perspicuity, efficiency, dependability) and hedonic quality aspects (stimulation, novelty). A number of studies have used the UEQ for evaluation of chatbots (Stefanidi et al. 2019, Valtolina et al. 2018).

The questionnaire that the Eco-Bot users will be asked to fill in after the completion of the pilot phase is based partially on certain attributes evaluated by the UEQ, which are of relevance to Eco-Bot. Our questionnaire examines also additional parameters that are more focused on the chatbot, so as to enable valuable insight into the way the user perceives Eco-Bot's chat interface, usefulness, and conversational intelligence. The parameters that have been selected to be evaluated in this context, have been defined by combining a number of metrics proposed in the chatbot performance and usability aspects literature (Jain et al. 2018; Kuligowska 2015; Radziwill & Benton 2017). The questionnaire is completed by three parameters that aim to capture the generic experience of the user after testing Eco-Bot. The generic user experience has no specific focus on any user experience aspects; it refers to a general impression or sentiment of the overall use or experience with the system (Bargas-Avila & Hornbæk 2011).

Details on the approach to be followed for evaluating the user experience with Eco-Bot, as well as the parameters addressed by the questionnaire, are given in the following tables.

Table 8: P8 - User experience

P8 – User experience	
Parameter Description	User experience involves both pragmatic and hedonic quality aspects; pragmatic quality aspects are typical usability aspects, i.e. efficiency, perspicuity, dependability, while hedonic ones are related to the way the user perceives the system in terms of stimulation and novelty. User experience involves also the perceived quality from the user's perspective as regards the chat interface and conversational intelligence of the chatbot, as well as indicates user acceptability and satisfaction. Feedback on user experience will enable the qualitative assessment of the Eco-Bot system based on subjective evaluation from the users.
Unit	%
Measurement Method	A user experience questionnaire will be given to the users after the completion of the pilot; this questionnaire will also be filled in by the participants of the small-scale validation at the end of this phase. Users will be asked to give feedback on perceived quality in parameters related to efficiency, perspicuity, dependability, stimulation, novelty, chat interface, conversational intelligence, and generic user experience.
Time of evaluation	Three phases are foreseen: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	D4.5 and D5.5
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for evaluating user experience with Eco-Bot in terms of usability and user satisfaction, and will enable system evaluation, refinement and optimisation. Dissemination (WP6): It will also be useful for the dissemination activities of Eco-Bot.

P8 – User experience	
Linkage to DoW's KPIsand targets	From Section 1.1. of the DoW: KPI_3.1: User Acceptability indicators of the chat-bot delivered. Target: More than 85% user acceptability of the proposed ICT tools for the users to be engaged based on the market segmentation.
Adapted or additionally created targets	Based on the DoW, the target for user experience and satisfaction would be: Average value of the users' scores in all questions of the 5-point Likert scale > 4.25 It should be noted that there are no available data in the literature w.r.t the acceptability of chatbots in the energy domain. Literature suggests that due to the large deviations recorded in different domains, the comparison of chatbots of different markets is not meaningful, and instead proposes that the targets should be that a) the newer versions of a chatbot are enhanced in terms of the monitored parameters and b) the user prefers using the chatbot instead of the existing channels it aims to replace (Discover.Bot 2019; Debecker 2017; Yao 2017; Merritt 2017).

The following table presents which parameters and how they will be evaluated through the user experience questionnaire:

Table 9: User Experience Questionnaire

	PRAGMATIC QUALITY ASPECTS	
Efficiency	• Is it possible to use Eco-Bot quickly and efficiently?	3 pairs of adjectives (slow/fast, inefficient/efficient, impractical/practical) measured on the 5-point Likert-type scale
Perspicuity	• Is it easy to get familiar with Eco-Bot? Is it easy to learn how to use it?	3 pairs of adjectives (clear/confusing, difficult to learn/easy to learn, complicated/easy) measured on the 5-point Likert-type scale
Dependability	• Do you feel that Eco-Bot is reliable and trustworthy? Is the interaction secure and predictable?	3 pairs of adjectives (unpredictable/predictable, not secure/secure, unreliable/reliable) measured on the 5-point Likert-type scale
	HEDONIC QUALITY ASPECTS	
Stimulation	• Is it interesting and exciting to use Eco-Bot? Do you feel motivated to use it further?	3 pairs of adjectives (boring/exciting, not interesting/interesting, demotivating/ motivating) measured on the 5-point Likert-type scale
Novelty	• Is Eco-Bot creative and inventive?	2 pairs of adjectives (dull/creative, uninventive/inventive) measured on the 5-point Likert-type scale
	CHATBOT-SPECIFIC ASPECTS	
Chat Interface	• How do you rate Eco-Bot's visual look?	5-point Likert-type scale of 1 (Very Poor) to 5 (Very Good)

	• Does the user interface look organised?	1 pair of adjectives (cluttered/organised) measured on the 5-point Likert-type scale
Usefulness	• A chatbot in the energy efficiency domain is useful. • Eco-Bot outperforms its existing alternatives (website, app, customer service line, search engines) by offering diverse and/or enhanced functionalities. • Eco-Bot covers the use cases I am interested in	5-point Likert-type scale (Strongly Disagree – Somewhat Disagree – Neither Agree nor Disagree – Somewhat Agree – Strongly Agree)
	• Which are the Eco-Bot features you are most interested in? (multiple choices possible) ○ Energy saving recommendations ○ Information on total energy consumption ○ Information on appliance level consumption ○ Comparison of consumption for different periods ○ Customer-service information (contract, guidelines on how to change contact details etc.) ○ Monitoring of energy efficiency goals ○ High/low consumption days ○ Alerts on high consumption ○ Monitoring of energy saving events ○ Other (please specify) …	Multiple choices possible and free text.
	• What other functionalities would you like Eco-Bot to offer?	Open-ended question (free text)
Conversational Intelligence	• Eco-Bot understands the input text. • Eco-Bot interprets commands accurately. • Eco-Bot executes requested tasks. • Eco-Bot is able to retain conversational context (maintain themed discussion) and follow up on a query.	5-point Likert-type scale (Strongly Disagree – Somewhat Disagree – Neither Agree nor Disagree – Somewhat Agree – Strongly Agree)
	GENERIC USER EXPERIENCE	
Overall Satisfaction	How would you rate your overall satisfaction with Eco-Bot on a scale of 1 (Very Poor) to 5 (Very Good)?	5-point Likert-type scale of 1 (Very Poor) to 5 (Very Good)
Future Use	How likely is it that you will use Eco-Bot in the future?	5-point Likert-type scale of 1 (Not at all likely) to 5 (Extremely likely)
Net Promoter Scale (NPS)	How likely is it that you would recommend Eco-Bot to a friend or colleague?	11-point Likert-type scale of 0 (Not at all likely) to 10 (Extremely likely)

4.3.2 Engagement and retention related parameters

According to the literature (Yao 2017; Merritt 2017), retention and engagement naturally vary with bot category and, although many bots claim to provide utility to users beyond traditional websites and apps, only a few actually convince users to stay engaged long enough. According to Ilker Koksal of BotAnalytics[5], nearly 40% of users stop talking to a bot after the first message and another 25% quit using it after the second one. Additionally, daily and monthly retention numbers are not impressive either, with daily rates lying between 1-2% and monthly rates hardly exceeding 7%. Studies further support the above findings, by presenting a significant decrease in the retention rate after the first month for chatbots across popular platforms like Facebook and Slack (Merritt 2017). In general, just as is the case with mobile apps, bots that naturally tie into a daily or weekly ritual see the best retention and engagement numbers.

It is suggested in the literature (Debecker 2017; Yao 2017; Merritt 2017) that it does not make sense to measure engagement by comparing the results with other chatbots in the market, not only because it is difficult to find published reliable numbers for comparison, but mainly because the engagement greatly varies with the category. Instead, it is considered as making a lot more sense to benchmark a chatbot's engagement against the other channels that it aims to replace. It should be noted that to the best of our knowledge, there are no relevant retention data available in the energy efficiency domain.

The following tables present the parameters that will be monitored in order to assess the engagement and retention of the users. These metrics were derived by examining and combining different parameters that are suggested in papers and articles in the field (Neff 2019; AIMultiple 2019; Radziwill & Benton 2017; Newlands 2017; Discover.Bot 2019; ChatbotPack 2019; Lee 2018; Merritt 2017; Yao 2017; Debecker 2017).

Table 10: P10 - Total users

P10 – Total users	
Parameter Description	This indicator defines the total number of Eco-Bot users during the 12-month period of the demonstration activities of WP5.
Unit	Number
Measurement Method	The total number of users registered to the Eco-Bot system will be retrieved from the Eco-Bot database.
Time of evaluation	First month of the pilot (February 2020 – M29)
Presentation of evaluation results	D5.2 – D5.3 – D5.4 and also in D5.5
WPs that the results will feed into and further utilization of results	System Development (WP4): It will be used as a reference for the engagement, retention and chatbot metrics.

[5] https://botanalytics.co/

P10 – Total users	
	Dissemination (WP6) / Exploitation (WP7): It will also be used for exploitation purposes of the Eco-Bot and will be useful for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	There is no direct linkage to DoW's KPIs, however, the number of users / buildings to be recruited for testing Eco-Bot per pilot are, according to the DoW, as follows: EYPESA: 99 users (EYPESA aims to have 66 bot users, the rest of the users are recruited so that they give consent that their data is used for comparison reasons). SEC: 150 users DEXMA: 20-30 buildings
Adapted or additionally created targets	EYPESA: 99 pilot participants, 66 Eco-Bot users (33 users with 1-hour smart meters and 33 users with 1-min smart meters and a control group that will not use Eco-Bot of 33 users) SEC: 150 users (50 users with 10-sec smart meters and 100 users without smart meters) DEXMA: 7 users (Facility Managers) that will handle a minimum of 20 buildings (hotels, supermarkets and restaurants) *please note that these aren't actually targets but rather the final number of registered participants in the pilots

Table 11: P11 - Active users

P11 – Active users	
Parameter Description	Active users are defined as the people who log in to Eco-Bot and read a message in a specific time frame (week, month, etc.).
Unit	Number
Measurement Method	The number of active users for different time frames will be extracted from the users' logins.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation (M28), b) in the middle of the pilots (M34), and c) at the end of the pilots (M40).
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 (M28) and the pilot results (both 6-month and 12-month) will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	Users that logged in Eco-Bot at least once every two months and at least 12 times in total during the 12-month pilot period > 80% of the total users.

Table 12: P12 - Engaged users

P12 – Engaged users	
Parameter Description	Engaged users are defined as the people who interact with Eco-Bot, i.e. send a message (either making an inquiry or responding to a message from Eco-Bot), in a specific time frame (week, month, etc.). It should be noted that we selected to use this term instead of "interactive users", as "engaged users" is the term widely used in the chatbot literature for the above definition. Notwithstanding, this indicator should not be confused with the engaged users who are – in Eco-Bot's context – those users that shift towards more energy efficiency behaviour through the use of Eco-Bot (despite the direct linkage between the two, as users should be engaged with the system so as for them to be motivated to engage to energy efficient behaviour). The users that changed their behaviour thanks to Eco-Bot will be identified through other metrics (see P26).
Unit	Number
Measurement Method	The number of engaged users for different time frames will be extracted by analysing the conversations.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 (M28) and the pilot results (both 6-month and 12-month) will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	Users that interacted with Eco-Bot at least once every two months and at least 12 times in total during the 12-month pilot period > 80% of the total users.

Table 13: P13 - Retention rate

P13 – Retention rate	
Parameter Description	This indicator shows the percentage of users that returned to Eco-Bot within a certain time period.
Unit	Percentage
Measurement Method	The percentage of users who returned to Eco-Bot within different time periods will be calculated based on the analysis of the users' logins. Daily, weekly and monthly rates will be examined.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the

P13 – Retention rate	
	end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 (M28) and the pilot results (both 6-month and 12-month) will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	No specific target value is defined. The aim of the indicator is to monitor the average frequency with which users return to Eco-Bot and how this frequency changes with progress of time.

Table 14: P14 - Sessions per day

P14 – Sessions per day	
Parameter Description	This indicator shows the average number of total sessions per day.
Unit	Number
Measurement Method	The number of total sessions per day will be monitored on a daily basis through the users' logins and logouts.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 (M28) and the pilot results (both 6-month and 12-month) will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs	None
Adapted or additionally created targets	No specific target value is defined.

Table 15: P15 - Sessions per user

P15 – Sessions per user	
Parameter Description	This indicator shows the average number of total sessions per user.

P15 – Sessions per user	
Unit	Number
Measurement Method	The average number of total sessions per user will be calculated based on the logins and logouts of all users for the whole demonstration period.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5.
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	No specific target value is defined.

Table 16: P16 - Time per session

P16 – Time per session	
Parameter Description	This indicator shows the average duration of the sessions.
Unit	Number
Measurement Method	The average duration of the sessions will be measured by dividing the whole duration of all sessions by the number of total sessions.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	This is a metric useful for usage results analysis, but there is no specific target related to it. This metric gives a basic idea of usage and it is useful to measure it, however the average session duration depends on a number of parameters, as e.g. the

P16 – Time per session	
	complexity of the inquiry, the fact that if the user is satisfied, he/she may start a second topic etc. As a general rule of thumb, customer service related inquiries should not take long to be answered, while, on the other hand, user-chatbot interactions aiming to engage the user in certain routines (e.g. towards a more energy efficiency behaviour) should be reflected in both high retention rates and sessions of longer duration. This metric can give useful insight into engagement and usability by being analysed in conjunction with the user's intents.

Table 17: P17 - Messages per session

P17 – Messages per session	
Parameter Description	This indicator shows the average number of messages exchanged between the user and the chatbot per session.
Unit	Number
Measurement Method	The average number of messages per session will be derived through analysis of the conversations.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs targets	None
Adapted or additionally created targets	This is a metric useful for usage results analysis, but there is no specific target related to it.

Table 18: P18 - Bot messages per session

P18 – Bot messages per session	
Parameter Description	This indicator shows the average number of messages sent by the chatbot in one session.
Unit	Number
Measurement Method	The average number of bot messages per session will be derived through analysis of the conversations.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the

P18 – Bot messages per session	
	end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 (M28) and the pilot results (both 6-month and 12-month) will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	This is a metric useful for usage results analysis, but there is no specific target related to it.

Table 19: P19 – In messages per session

P19 – In messages per session	
Parameter Description	This indicator shows the average number of messages sent by the user in one session.
Unit	Number
Measurement Method	The average number of in messages per session will be derived through analysis of the conversations.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 (M28) and the pilot results (both 6-month and 12-month) will be documented in D5.5 (M43).
WPs that the results will feed into and further utilization of results	System Development (WP4): this indicator is useful for monitoring user engagement and retention and, consequently, for system evaluation and fine-tuning. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	This is a metric useful for engagement and usage results analysis, but there is no specific target related to it. As a general rule of thumb, if this number is particularly low, the usefulness of the chatbot is questionable, as it might indicate that users prefer alternative ways to address their inquiries, and/or that the chatbot failed to engage them (AIMultipe, 2019).

4.3.3 Chatbot usability related parameters

The following tables present the parameters that will be monitored in order to gain insight into the usability of the chatbot, assessing its efficiency and effectiveness performance as well as its usage, thus enabling system refinement and optimisation. These metrics were derived by examining parameters suggested in papers and articles in the field (Neff 2019; AIMultiple 2019; Radziwill & Benton 2017; Newlands 2017; Discover.Bot 2019; ChatbotPack 2019; Lee 2018).

Table 20: P20 – Topics Popularity Index

P20 – Topics Popularity Index	
Parameter Description	This parameter captures the overall trend in users' preferred inquiries. It highlights the most popular questions asked by the users, thus giving valuable insight into the usability of the bot and enabling refinement, customisation as well as potential identification of new use cases (features) to support.
Unit	%
Measurement Method	The inquiries in which the users are interested will be derived through analysis of the conversations. The extracted intents will be ranked based on their popularity. The popularity index (PI) of each intent will be calculated as follows: PI = (number of occurrences of the intent in all conversations) / (total number of inquiries in all conversations)
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5.
WPs that the results will feed into and further utilization of results	System Development (WP4): this parameter is useful for identifying the most popular inquiries and examine the way the bot treats them, enabling refinement and response effectiveness optimisation. Exploitation (WP7)/Dissemination (WP6): Users' preferred inquiries will give valuable feedback in terms of exploitation and will also be exploited in the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	This is a metric useful for usage results analysis, but there is no specific target related to it.

Table 21: P21 – Fall Back Rate (FBR)

P21 – Fall Back Rate (FBR)	
Parameter Description	Fall Back Rate is the percentage of times the chatbot fails to complete a task.
Unit	%

P21 – Fall Back Rate (FBR)	
Measurement Method	Analysis of the conversations and examination of the unhandled/fallback intents will take place so as to identify the cases that a specific task (inquiry) from the list of predefined use cases (offered services) could not be completed. FBR will be calculated as follows: FBR = (number of times the chatbot failed to complete a task) / (total number of inquiries) * 100%
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5.
WPs that the results will feed into and further utilization of results	System Development (WP4): this parameter is useful for system refinement and optimization. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	This is a metric useful for system refinement and optimization, but there is no specific target related to it.

Table 22: P22 – Confusion triggers

P22 – Confusion triggers	
Parameter Description	This indicator defines the number of times that the chatbot fails to understand a message or misinterprets a message. It is a helpful parameter for identifying how and where the chatbot needs to be improved.
Unit	Number
Measurement Method	Analysis of the conversations and examination of the unhandled/fallback intents will take place so as to identify the confusion triggers. The confusion triggers will be categorised per type: a) the chatbot cannot understand a message/inquiry; b) the chatbot cannot understand the user's response; c) the user sends an inquiry that is beyond Eco-Bot's remit; and d) the chatbot misinterprets a message.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5.
WPs that the results will feed into and further utilization of results	System Development (WP4): this parameter is useful for system refinement and response effectiveness optimization. Dissemination (WP6): it will also be used in the dissemination results.

P22 – Confusion triggers	
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	This is a metric useful for system refinement and response effectiveness optimization, but there is no specific target related to it.

Table 23: P23 – Task completion time

P23 – Task completion time	
Parameter Description	This indicator shows the time that is needed in order for Eco-Bot to complete a task successfully.
Unit	Number
Measurement Method	Analysis of the conversations and examination of the intents will take place so as to calculate the time it takes for Eco-Bot to complete successfully a task (inquiry) from the list of predefined use cases (offered services). As the time depends also on the complexity of the inquiry, the task completion time will be monitored not only as an average value taking into account all different completed tasks, but also the average time needed for each of the tasks (intents) will be assessed.
Time of evaluation	This parameter will be monitored throughout the small-scale validation phase and the pilot. The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.
Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5.
WPs that the results will feed into and further utilization of results	System Development (WP4): this parameter is useful for system refinement and optimization. Dissemination (WP6): it will also be used in the dissemination results.
Linkage to DoW's KPIs and targets	None
Adapted or additionally created targets	This is a metric useful for system refinement and optimization, but there is no specific target related to it.

4.3.4 Lab usability testing

This section presents a brief overview of the lab usability test that will take place in the context of Task 4.4 in order to test and evaluate the system before its release for validation by end-users in the forthcoming small-scale validation and pilot phases.

Lab usability tests measure a user's ability to complete tasks; in a typical usability test, a user is asked to perform a task or set of tasks using the tool in question. This enables a first evaluation of the tool in terms of effectiveness, efficiency and satisfaction in a specified context of use, and potential refinement/fine-tuning, before its first release for testing in real conditions.

During the lab usability testing that will take place in Task 4.4, test users will be asked to test all the use cases (features) supported by Eco-Bot based on specific scenarios. The whole procedure and the relevant metrics will be described in detail in D4.4. Indicatively, Eco-Bot will be evaluated in terms of a) efficiency, b) dependability, c) conversational intelligence, and d) user experience, as follows:

Table 24: Lab Usability Testing

	LAB USABILITY TESTING	
Efficiency	• Ability to respond fast and efficiently	• Check the Task Completion Time • Check overall system responsiveness in case of multiple concurrent requests
Dependability	• Availability (readiness for correct service) • Reliability (continuity for correct service)	• Check Fall Back Rate and Confusion Triggers • Check overall system responsiveness in case of multiple concurrent requests
Conversational Intelligence	• Eco-Bot understands the input text. • Eco-Bot interprets commands accurately. • Eco-Bot executes requested tasks. • Eco-Bot is able to retain conversational context (maintain themed discussion) and follow up on a query.	• Run tests based on specific scenarios (e.g. "I want to know my total consumption for yesterday" and after a while "What about two days ago?") • Check Fall Back Rate and Confusion Triggers
User Experience	• Evaluation of the system in terms of user experience	• Test users will be asked to respond to the user experience questionnaire after the completion of the usability test.

4.4 The Eco-Bot impact evaluation metrics

This section presents the metrics that will evaluate the impact of Eco-Bot. The impact concerns the energy saving actions, the environmental or green impact, the economic impact and the awareness for the rebound effect.

4.4.1 Energy saving actions related parameters

The following parameters all constitute operationalizations of the KPIs of the DoW. An additional parameter was created to distinguish between the residential application of Eco-Bot and the commercial one. These parameters measure the behavioral changes and investments of Eco-Bot users. Each of the parameters has a set target, either already provided by the DoW or created based on possible/realistic scenarios. Two of the targets have been adapted to the context of the pilot, and one target was additionally created for the more optimistic expectations with regards to the behavioral change and investments that commercial buildings will be affected by, thanks to Eco-Bot.

Table 25: P25 - Total increase of energy savings by participating users

P25 - Total increase of energy savings by participating users	
Parameter Description	This indicator describes the percentage increase of energy savings achieved by the Eco-Bot users after they used Eco-Bot. It compares the energy consumption of the users before and after using Eco-Bot.
Unit	%
Measurement Method	The total energy consumption of the users during the 12 months of the pilot phase will be compared to their total energy consumption of the corresponding reference period before the pilot phase (baseline), so as to calculate the total increase of energy savings in % achieved during the demonstration and validation phase. The comparison of the energy consumptions before and during the use of Eco-Bot and the resulting calculation of the energy savings will be performed by each of the three pilots, which have at their disposal the previous energy consumptions of their users.
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	It is linked to the following two KPIs of Section 1.1 of the DoW: KPI_5.1: Overall decrease of energy consumption of participating user in per cent KPI_5.2: Total increase of carried out energy saving actions by participating users Target for both KPIs: energy savings of approx. 20%.
Adapted or additionally created targets	Adapted target for the duration of the project: 15% Note: This target was adapted. This adapted target refers to the energy savings to be achieved by the end of the pilot. 20% is the expected target after the end of the Eco-Bot project.

Table 26: P26 - Users that made a change to save energy (behavioral change or investments)

P26 - Users that made a change to save energy (behavioural change or investments)	
Parameter Description	This indicator describes how many users (in %) have made at least one change to save energy after the use of Eco-Bot. This indicator takes into account changes that were recommended by Eco-Bot (behavioural change or investments) as well as other energy saving events performed by the user after they started using Eco-Bot.
Unit	%
Measurement Method	The number of users that entered themselves an energy saving measure that they implemented or that confirmed that they implemented an energy saving

	measure suggested by Eco-Bot are counted and put in relation to the total number of Eco-Bot users.
Time of evaluation	This parameter will be evaluated in two phases, i.e. in the middle of the pilots (M34) and at the end of the pilots (M40).
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	From Section 2.1 Expected impacts KPI_3.2: Quantification of energy savings in sample due to behaviour change, pre- and post- Eco-Bot feedback and interaction Target: More than 30% of users will have changed their behaviour towards energy efficiency in the selected measurable sample of the first phase
Adapted or additionally created targets	As specified in the DoW (see above) Resulting thus in a total of users of 64 users during the project, composed of: EYPESA: 19 Users changed behaviour SEN: 45 Users changed behaviour

Table 27: P27 – Consumers making monetary investments to save energy

P27 – Consumers making monetary investments to save energy	
Parameter Description	This parameter measures how many Eco-Bot users in % have made monetary investments to save energy (e.g. purchase of energy efficient products, insulation, etc.).
Unit	%
Measurement Method	The number of users who registered at least one investment/followed at least one recommendation of Eco-Bot regarding investments is put into relation with the total amount of Eco-Bot users.
Time of evaluation	This parameter will be evaluated in two phases, i.e. in the middle of the pilots (M34) and at the end of the pilots (M40).
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	From Section 2.1 Expected impacts Section 2.1: "KPI_1.2: Measuring the decisions of consumers with regards to energy efficiency (e.g. purchase of higher energy efficiency products, use of renewable energy sources, and tendency for "prosumer" attitude).

	Target: More than 15% of the total users have been routed towards more energy efficient behaviour"
Adapted or additionally created targets	Adapted target: 10% of users made energy efficiency investments Note this target has been adapted, because in the span of the pilot the investments will most likely be limited to the purchase of higher efficiency products. It is rather unlikely that in the short course of the pilot someone actually becomes a prosumer, e.g. by having solar panels installed. However, users might be routed towards more energy efficient behavior and thus gain interest or even plan to become prosumers or to switch energy plans or providers to use more renewable energy sources. This intent will be measured by P32.

Table 28: P28 – Commercial buildings (facilities) that were affected by a change to save energy (behavioral change or investments)

P28 – Commercial buildings (facilities) that were affected by a change to save energy (behavioural change or investments)	
Parameter Description	This indicator describes how many commercial buildings (facilities) (in %) were affected by at least one change to save energy after the use of Eco-Bot. This indicator takes into account changes that were recommended by Eco-Bot (behavioural change or investments) as well as other energy saving events registered independently of Eco-Bots recommendations.
Unit	%
Measurement Method	The number of commercial buildings (facilities) that were affected by at least one energy saving measure are counted and put in relation to the total number of commercial buildings (facilities).
Time of evaluation	This parameter will be evaluated in two phases, i.e. in the middle of the pilots (M34) and at the end of the pilots (M40).
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	From Section 2.1 Expected impacts KPI_3.2: Quantification of energy savings in sample due to behaviour change, pre- and post- Eco-bot feedback and interaction Target: More than 30% of users will have changed their behaviour towards energy efficiency in the selected measurable sample of the first phase
Adapted or additionally created targets	Adapted target: 80% of the commercial buildings (facilities) Assuming 20 participating facilities, this would mean that in a total of 16 facilities energy saving measures were implemented in the course of the pilot *Note: The reason for the creation of the adapted target is that we expect commercial businesses to be more likely to implement energy saving measures than private households,*

Table 29: P29 - Implemented energy saving measures recommended by Eco-Bot

P29 - Implemented energy saving measures recommended by Eco-Bot	
Parameter Description	This indicator describes how many of the energy saving measures recommended by Eco-Bot, the user implemented.
Unit	Number
Measurement Method	Eco-Bot asks a follow-up question after having given a recommendation to check whether the user implemented the suggestion or not. The number of implemented recommendations are counted.
Time of evaluation	This parameter will be evaluated in two phases, i.e. in the middle of the pilots (M34) and at the end of the pilots (M40).
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	From Section 1.1.2. Project objectives KPI 3.3: "Number of implemented energy saving measures recommended by the chat-bot "
Adapted or additionally created targets	**640 Energy saving measures recommended by the chat-bot were implemented by residential users (pilot participants of EYPESA and SEN)** *Calculation: Assuming that 30% of the residential users implemented at least one energy saving measure (see P26) this means 64 users changed their behaviour in the pilot. Further assuming that on average they will implement 10 recommendations proposed by the Eco-Bot in the course of the project, a total of 640 recommendations are implemented.* **160 Energy saving measures recommended by the chat-bot affected commercial buildings (participating facilities of the DEX pilot)** *Calculation: Assuming that 80% of the commercial buildings (facilities) were affected by at least one energy saving measure (see P28) this means 16 buildings were affected by a change. Further assuming that on average buildings are affected by 10 changes (implemented recommendations), a total of 160 recommendations are implemented.*

4.4.2 Green impact related parameters

This section presents the green impact related parameters, i.e. the parameters that evaluate the environmental impact of Eco-Bot. All parameters are operationalizations of the KPIs from the DoW. To calculate the CO_2 emissions avoided, the average coefficient of carbon intensity

was researched. The two residential pilots (EYPESA and SEN) resorted to the respective country specific coefficient, while the business pilot (DEX) used a European average, given that the buildings are located in different European countries. EYPESA identified the most recent Spanish coefficient of carbon intensity according to a peer reviewed scientific article (Moro/Lonza 2018). SEN identified the German the most recent coefficient of carbon intensity from a publication of the German Federal Environment Agency (UBA 2018). DEX took the most recent European average coefficient of carbon intensity as provided by the European Environment Agency (EEA 2016). A target was created of the CO2 emissions avoided per participating user based on the expected energy savings achieved in the course of the pilot. All parameters have quantifiable targets. Each of the presented parameters also enables group evaluations (see Chapter 5).

Table 30: P30 – Overall energy savings achieved (in MWh)

P30 - Overall energy savings achieved (in MWh)	
Parameter Description	This indicator quantifies the amount of energy (in MWh) that the users saved throughout the Eco-Bot pilot.
Unit	GWh
Measurement Method	The total energy consumption of the users during the 12 months of the pilot phase will be compared to their total energy consumption of the corresponding reference period before the pilot phase (baseline), so as to calculate the total increase of energy savings achieved (in MWh) during the demonstration and validation phase. The comparison of the energy consumptions before and during the use of Eco-Bot and the resulting calculation of the energy savings will be performed by each of the three pilots, which have at their disposal the previous energy consumptions of their users.
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	From Section 2.1. KPI_1.4 (individual consumers SEC): Quantified GWh/year of energy savings Target Value: 22.8 GWh/year KPI_1.5 (utility consumers EYPESA): Projected GWh/year of Energy savings Target Value: 4.95 GWh/year KPI_1.6 (energy managers DEXMA) Projected GWh/year of Energy savings Target: 250 MWh/year
Adapted or additionally created targets	Adapted targets for the pilot phase: SEC: 90 MWh/year for electricity and 112.5 MWh/year for space heating *Calculation for electricity saving:*

	➔ *total electricity consumption of 150 German Eco Bot users: 150 users * 4 MWh/year = 600 MWh/year.* ➔ *15% energy consumption decrease: energy savings: 0,15*600MWh/year = 90 MWh/year* *(Assumption that average electricity consumption per household in Germany is 4MWh/year.)* *Calculation for space heating:* ➔ *Total heat energy consumption of 150 German Eco Bot users: 150 users*15MWh/year = 2250 MWh/year* ➔ *5% energy consumption decrease: energy savings: 0,05*2250 MWh/year = 112.5 MWh/year* *(Assumption that average energy consumption for space heating of an average 100 sqm per household consuming 150kWh/sqm: 15 MWh/year)* EYPESA: 34,521 MWh/year *Calculation and assumptions for energy saving:* ➔ *Aim to have 66 eco-bot users; Assuming that each user on average consumes: 3487 kWh/year (https://www.idae.es/uploads/documentos/documentos_Documentacion_Basica_Residencial_Unido_c93da537.pdf), Assumption during the pilot: 15% savings, After the pilot: 20% savings* ➔ *3487 kWh/year*66 users*15% savings= 34521,3 kWh/year → 0,0345 GWh/year* DEXMA: 1500 MWh/year (assuming 20 participating facilities) *Calculation and assumptions for energy saving:* ➔ *each participating facility will achieve 7500 kWh of energy savings (20*7500 kWh= 1500MWh/year)* ➔ *(Assuming that 37500 kWh/year is the energy consumption for an average building.)* Note: This target was adapted in accordance with a realistic scenario applicable to the scope of the pilot. Taking into account the actual number of Eco-Bot users in the pilot (with the DEXMA pilot of 7 Facility managers and around 20 participating facilities, the EYPESA pilot with 66 households and SEN pilot with 150 households). Furthermore, contrary to the projected impact that the DoW indicates, the adapted target will measure the actual achieved energy savings in the course of the pilot.

Table 31: P31 - Average amount of avoided CO_2 emissions of each user

P31 - Average amount of avoided CO_2 emissions of each user	
Parameter Description	This indicator describes how much CO_2 emissions have been reduced by each user through the use of Eco-Bot.
Unit	kg
Measurement Method	Calculated by multiplying the energy consumption by the carbon intensity coefficient of the electricity consumed, in the respective country.
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5

WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	From 1.1.2. Project objectives KPI_5.4: Average amount of avoided CO_2 emissions in kg of each user
Adapted or additionally created targets	Created target pilot based on energy savings achieved: **DEX: kg/kWh = 2 220 CO2 kg/year per participating facility** *Calculation for electricity:* *150 000 kWh (overall energy savings achieved) / 20 (participating facilities) = 7500 kWh (energy savings achieved per facility)* *7500 kWh (energy savings achieved per facility) x 0,296 kg CO2/kWh (European average coefficient of carbon intensity per kWh) = 2 220 kg CO2/year per participating facility* **EYPESA: 178,3 kg/year per user for electricity** *Calculation for electricity:* *34 521 kWh/year (overall energy savings achieved) / 66 (participating households) = 523 kWh/year per household (rounded to one decimal place)* *523 kWh (energy savings per household) x 0,341 kg CO2/kWh (Coefficient of carbon intensity per kWh in Spain) = 178,3 kg CO2/year per household (rounded to one decimal place)* **SEN: 484.4 kg/year per user for electricity and 150kg/year user for space heating** *Calculation for electricity:* *90 000 kWh/year (overall energy savings achieved) / 150 (participating households) = 600 kWh/year per household* *600 kWh (energy savings per household) x 4,474 kg CO2/kWh (Coefficient of carbon intensity per kWh in Germany for electricity) = 484,4 kg CO2/year per household* *Calculation for space heating:* *112 500 MWh/year for space heating/ 150 (participating households) = 750 kWh/year per participating household* *750 kWh (energy savings per household) x 0,2 (Coefficient of carbon intensity per kWh in Germany for gas) = 150 kg CO2/year per household*

Table 32: P32 -Turn to sustainable energy: Number of users interested in turning to renewable/sustainable energy

P32 - Turn to sustainable energy: Number of users interested in turning to renewable/sustainable energy	
Parameter Description	This indicator describes how many users (in %) are interested in turning to renewable energy (e.g. by becoming a prosumer or changing energy plans or provider etc.)
Unit	%
Measurement Method	Question in the user survey inquiring the interest of the users to turn to renewable energy by one or multiple of the following options: • switching to a green energy plan/provider • producing/supporting renewable energy yourself e.g. by investing in solar panels or investing in a windmill
Time of evaluation	End of the pilot
Presentation of evaluation results	D5.5

WPs that the results will feed into and further utilization of results	This will be used to grasp the potential further impact that Eco-Bot after the end of the project. And how Eco-Bot can contribute to realize the energy transition.
Linkage to DoW's KPIs and targets	KPI from Section 2.1 Expected Impact: KPI_1.2: Measuring the decisions of consumers with regards to energy efficiency (e.g. purchase of higher energy efficiency products, use of renewable energy sources, and tendency for "prosumer" attitude). Target: More than 15% of the total users have been routed towards more energy efficient behaviour
Adapted or additionally created targets	No new or adapted target (target stays 15%, as specified in the DoW).

4.4.3 Economic parameter

This section presents the economic parameter measuring the economic gain (EUR saved) per participating household or facility. This parameter is an operationalization of a KPI from the DoW. Targets for each pilot were created through a calculation based on the achieved energy savings (P30).

(For Senercon, the prices for one kW/h for heating were taken from the internet comparison portal (www.gasauskunft.de) and for electricity from the Federal Ministry for Economic Affairs and Energy (https://www.bmwi.de/Redaktion/DE/Artikel/Energie/strompreise-bestandteile.htm). EYPESA used their common tariff/price that their customers have. DEXMA used the average European electricity price from Eurostat https://ec.europa.eu/eurostat/statistics-explained/index.php/Electricity_price_statistics#Electricity_prices_for_household_consumers.)

Table 33: P33 - Amount of money saved per household/facility

P33 - Amount of money saved per household/facility	
Parameter Description	This indicator describes how much money has been saved per participating household/facility by the use of Eco-Bot.
Unit	EUR
Measurement Method	Calculation of energy savings in kWh x price of energy of the respective country and region in EUR per kWh. Note in the SEN pilot the savings are also calculated for space heating: the energy savings achieved in space heating x price of gas in Germany.
Time of evaluation	End of pilot
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot.
Linkage to DoW's KPIs and targets	The importance of the economic gain is mentioned in the DoW however not linked to a specific KPI or target.

Adapted or additionally created targets	**DEX: 1125 EUR/year per facility** *Calculation for electricity savings:* *150 000 kWh (overall energy savings achieved) / 20 (participating facilities) = 7500 kWh* *7500 kWh (energy savings achieved per facility) x 0,12 EUR (average price of one kWh in Europe) = 900 EUR* **EYPESA: 78,45 EUR/year per household** *Calculation for electricity savings: 34 521 kWh/year (overall energy savings achieved) / 66 (participating households) = approx. 523 kWh/year per household* *523 kWh (energy savings per household) x 0,15 EUR (average price of one kWh in Spain) = 78, 45 EUR/year per household* **SEN: 180 EUR/year per household for electricity (and 45 EUR/year per household for space heating** *Calculation for electricity:* *90 000 kWh/year (overall energy savings achieved) / 150 (participating households) = 600 kWh/year per household* *600 kWh (energy savings per household) x 0,3 EUR (average price of one kWh in Germany) = 180 EUR/year per household* *Calculation for space heating:* *112 500 MWh/year for space heating/ 150 (participating households) = 750 kWh/year per participating household* *750 kWh (energy savings per household) x 0,06 EUR (average price of one kWh in Germany for gas) = 45 EUR/year per household*

4.4.4 Rebound effect related parameter

This section presents the rebound effect related parameter. The rebound effect (i.e. overcompensation of achieved energy savings through increased energy usage) is a complex and multi-layered phenomenon. There are direct rebound effects, which are direct changes in the product use (e.g. purchase a more energy efficient appliance but run it more and longer) and indirect rebound effects (e.g. use the money saved from energy savings to purchase a large plasma screen). While the existence of the rebound effect and its importance have been demonstrated (Freire-González 2017) it is very hard to quantify the amplitude of the rebound effect. This is partly due to the fact, that it is hard to isolate the rebound effect from other factors that increase energy demand (such as weather conditions or economic growths). Therefore the projections regarding the amplitude of the rebound effect vary widely across the studies (UBA 2014). For instance, Wang et al. (2016) suggest that rebound effects in the residential electricity use in Beijing are between 46% and 56% in the long-term. An evaluation of the magnitude of the rebound effect that occurred in the three Eco-Bot pilots is beyond the scope of the project. However, tackling the rebound is within the projects realm.

Eco-Bot tackles the rebound effect by making users aware of the existence and its effect and gives information on how to avoid it, thus to act preventatively. Therefore, the following parameter was created:

Table 34: P34 – Sensibilization of the users for the rebound effect

P34 – Sensibilization of the users for the rebound effect	
Parameter Description	This parameter addresses the sensibilization and awareness for the rebound effect among Eco-Bot users since the challenge is that users are often unaware of this issue. When a user registered the purchase of a new appliance, Eco-Bot will automatically inform the user about the risk of the occurrence of a rebound effect and give recommendations on how to avoid it. Another message will be sent two months after the purchase of a new appliance.
Unit	%
Measurement Method	User survey
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5
WPs that the results will feed into and further utilization of results	Contribute the data to the research on the rebound effect.
Linkage to DoW's KPIs and targets	The importance of the rebound effect is mentioned in the DoW however not linked to a specific KPI or target.
Adapted or additionally created targets	50% of users who received information on the rebound effect found it useful

4.5 Pilot specific evaluation metrics

This chapter presents the parameters that are specific to each of the pilots. Each sub chapter presents the interest and expectations of a pilot. The parameters are logically connected to some KPIs of the DoW but do not have a direct link and do not constitute an operationalization of the KPIs. These parameters were created upon the reviewers' suggestion since their evaluation will be valuable for further exploitation and dissemination of the Eco-Bot project. The parameters have been suggested by the pilots.

4.5.1 EYPESA pilot specific parameters

This section presents the (B2C) pilot of residential users in Spain. Spain is among the countries, that committed to perform a large-scale smart meter roll-out by 2020 or earlier (EUR-lex 2014). In 2018, the country completed the smart meter roll out (Ledo 2018). Consequently, all EYPESA customers have at least basic smart meters, transmitting data on an hourly basis.

In this pilot there are three different groups of enrolled EYPESA customers. A group of 33 Eco-Bot users with the basic smart meter, a group of 33 Eco-Bot users with advanced smart meters (transmitting data every minute) and a control group of 33 Non-Eco-Bot users. Having these three groups allows EYPESA to compare the energy savings achieved and other related outcomes (e.g. money saved, and CO_2 emissions avoided) among the different groups. The comparison between users with an hourly meter and users with a minute meter

helps to answer the critical questions: are users with a more detailed feedback thanks to the minute meter more engaged? Do they save more energy? Or does it make little to no difference? Answering these questions will be invaluable for the dissemination and future energy saving projects with smart meters. Comparing the control group of non-Eco-Bot users with the Eco-Bot users will provide information that goes beyond the comparison of the energy consumption before and after the bot, since both groups are monitored in the same time period with the same conditions. Consequently, EYPESA will be able to deduce which percentage of the energy savings must be related to external factors that affected all pilot participants across the three groups (such as weather conditions and economic situation). This parameter will ensure the soundness of the scientific approach and help contextualize the evaluation results in other parameters. Since it is merely a control there is no target associated with it.

Table 35: P35 – EYPESA Energy savings achieved by users with basic smart meters compared to users with advanced smart meters

P35 – EYPESA Energy savings achieved by users with basic smart meters compared to users with advanced smart meters	
Parameter Description	This indicator quantitatively demonstrates the differences in energy savings achieved by users with a basic smart meter (hourly data transmission) and with a more advanced smart meter (data transmission per minute) in per cent. The underlying hypothesis is that users with advanced smart meters have more detailed insights in their energy consumption on an appliance level and thus achieve greater energy savings then the basic smart meters.
Unit	%
Measurement Method	Total energy savings achieved in kWh are compared among the users with basic and with advanced smart meters.
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5 (M43)
WPs that the results will feed into and further utilization of results	Feed into the lessons learned (WP7) of the project and will be useful for similar future projects to build on the experience. Furthermore, the difference in energy savings achieved between these two groups can be translated into economic gain (how much money can be saved) and how much CO_2 can be avoided. This information is invaluable when consulting customers on different smart meters. It will also show whether it is worth having the more advanced meter or whether it suffices to have the standard smart meter.
Linkage to DoW's KPIs	None
Target	None

Table 36: P36 – EYPESA: Energy savings achieved by Eco-Bot users compared to the control group of non-Eco-Bot users

P36 – EYPESA: Energy savings achieved by Eco-Bot users compared to the control group of non-Eco-Bot users	
Parameter Description	This parameter compares the energy savings achieved by Eco-Bot user with the energy savings achieved by the control group of non-Eco-Bot users to control for context depended factors that have repercussions on the energy consumption such as weather conditions or the economic situation.
Unit	%
Measurement Method	Total energy savings achieved in kWh are compared among the group of Eco-Bot users and the control group of non-Eco-Bot users.
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5 (M43)
WPs that the results will feed into and further utilization of results	Exploitation (WP7)/Dissemination (WP6): This will be used for exploitation purposes of the Eco-Bot and will be valuable for the dissemination activities of Eco-Bot. It will show which percentage of energy savings are linked to other external factors and not Eco-Bot. Thereby a scientifically sound approach will be ensured.
Linkage to DoW's KPIs	None
Target	None

Furthermore, EYPESA expects Eco-Bot to achieve high rates of user satisfaction (see P8), so to ensure that Eco-Bot is an attractive additional service feature for EYPESA customer, which would also be a good basis to enlarge the customer basis of EYPESA. Moreover, EYPESA expects Eco-Bot to relieve the customer service. EYPESA would find Eco-Bot of very high interest, if users were not referred to the customer service in more than 50% of their Eco-Bot sessions.

Table 37: P37 – EPEYSA: Self Service Rate

P37 – EYPESA: Self Service Rate	
Parameter Description	% of sessions and % of inquiries that the bot was able to go through without directing the user to the customer service office.
Unit	%
Measurement Method	% of conversations/inquiries where the user was not passed the customer service contact Number of conversations (inquiries) where the user was not passed to the customer service contact (under any circumstance)/Total number of conversations (inquiries)
Time of evaluation	The evaluation results will be recorded in three different phases: a) at the end of the small-scale validation, b) in the middle of the pilots, and c) at the end of the pilots.

Presentation of evaluation results	The small-scale validation phase results will be documented in D4.5 and the pilot results (both 6-month and 12-month) will be documented in D5.5 (M43)
WPs that the results will feed into and further utilization of results	These results are interesting for further exploitation of Eco-Bot, and for evaluating Eco-Bot as a tool that could be interesting for customer support in the electricity sector.
Linkage to DoW's KPIs	None
Target	More than 50%

4.5.2 SEN pilot specific parameters

This section presents the (B2B2C) pilot of residential users in Germany. Germany is moving slowly on the smart meter roll out since it figures among the countries where the cost-benefit-analysis was negative or inconclusive (EUR-lex 2014). In 2016, Germany passed a legislation (Messstellenbetreibergesetz, MsBG) to ensure large-scale smart meter roll-out in Germany. However, the roll-out will not be completed before the end of the pilot in 2020 (Bundesnetzagentur 2019). Consequently, none of the SEN customers have a smart meter to begin with and SEN will distribute 50 advanced smart meters (data transmission every 10 seconds) among the 150 Eco-Bot users. This allows SEN to compare smart meter users versus non smart meter users and thus evaluate whether smart meter users achieve greater energy savings than non-smart meter users. Such evaluations are invaluable when consulting customers on smart meters, hence the following parameter was created:

Table 38: P38 – SEN Energy savings achieved by users with smart meters compared to users without smart meters

P38 – SEN Energy savings achieved by users with smart meters compared to users without smart meters	
Parameter Description	This indicator quantitatively demonstrates the differences in energy savings achieved by users with and without smart meters. The underlying hypothesis is that users with smart meters have insights in their energy consumption on an appliance level and thus achieve greater energy savings then the ones without smart meters.
Unit	%
Measurement Method	Total energy savings achieved in kWh are compared among the users with and without smart meters.
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5 (M43)
WPs that the results will feed into and further utilization of results	Feed into the lessons learned (WP7) of the project and will be useful for similar future projects to build on the experience. Furthermore, the difference in energy savings achieved between these two groups can be translated into economic gain (how much money can be saved) and avoided CO2. This information is valuable when consulting customers on getting a smart meter and how much money they could save and how much CO_2 they could avoid.
Linkage to DoW's KPIs	

	None
Target	5% more savings

Furthermore, SEN defined what they want to achieve through Eco-Bot. Since the iESA system relies on users to enter energy data and energy saving events, SEN hopes that Eco-Bot motivates users to make more data entries to the system and created a respective parameter with a target of a 15% increase of entries to the system. Moreover, the iESA system can evolve from an energy monitoring system to an energy advising system with tailored recommendations. Eco-Bot could also foster a tighter link between SEN and their customers by providing more information of the customers' building and appliances and suitable energy saving measures. In the same vein, Eco-Bot constitutes a modernization of iESA: thanks to the NILM (see P3-P7), SEN can offer their customers precise information on their energy consumption without high submeter costs. SEN is interested in their customer's perception of Eco-Bot as a new channel of communication. Consequently, a parameter has been created to evaluate which of the three communication channels (hotline, user forum, Eco-Bot) that SEN provides is the preferred one. The evaluation can be performed for all Eco-Bot users in the SEN pilot, as well as along the lines of demographic criteria. Since SEN is hoping to reach out and engage users that are underrepresented for now, such as female and/or younger users, they can verify whether Eco-Bot was popular among this group of users. SEN defined that if Eco-Bot was the preferred channel of communication for more than 30% of the female or younger users (aged < 45) then it would be very interesting for further exploitation. Eco-Bot could be rolled out to more customers of SEN or according to the evaluation results to a specific group of customers. In addition, when Eco-Bot is revealed as a popular communication channel, more tasks could be transferred to the realm of Eco-Bot.

Table 39: P39 – SEN increase of energy saving events in percent entered into the iESA system

P39 – SEN increase of energy saving events in percent entered into the iESA system	
Parameter Description	This indicator describes the increase in per cent of energy saving events entered into the iESA system as result of using Eco-Bot.
Unit	%
Measurement Method	Total number of energy saving events entered by iESA users into the system within the pilot phase divided by the total number of energy saving events entered by the same users during a corresponding (annual) period before the pilot started (multiplied by 100). Note: users may also fill in energy saving measures that were not recommended by the bot but as a result of raised awareness due to communication features of the Eco-Bot system (less technical and easier to handle approach than the conventional use of iESA)
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4

Presentation of evaluation results	D5.5 (M43)
WPs that the results will feed into and further utilization of results	Feed into the lessons learned (WP7) of the project and will be useful for similar future projects to build on the experience.
Linkage to DoW's KPIs	none
Target	15 % increase

Table 40: P40 – SEN Evaluation of Eco-Bot as channel of communication

P40 – SEN Evaluation of Eco-Bot as channel of communication	
Parameter Description	This quantitative indicator demonstrates which of the three possible communication channels was preferred by the users (if applicable according to their segment or demographics).
Unit	%
Measurement Method	Question in the user survey. All pilot participants are asked which channel of communication they preferred/ found the most effective (Eco-Bot, the service hotline or the user forum).
Time of evaluation	After the completion of the demonstration activities of Tasks 5.2-5.4
Presentation of evaluation results	D5.5 (M43)
WPs that the results will feed into and further utilization of results	If the user survey reveals that the bot is the preferred communication channel, it could be rolled out to broader user group. Additionally, further tasks could be carried out by Eco-Bot, e.g. information on and testing of new features could be spread by the bot.
Linkage to DoW's KPIs	None
Target	30 % share of younger (< 45) and/or female users preferred Eco-Bot as a channel of communication over the other options

4.5.3 DEX pilot specific parameters

Table 41: P41 – NPS comparison of "DEXCell EM with Eco-Bot" and "DEXCell EM without Eco-Bot"

P41 - NPS comparison of "DEXCell EM with Eco-Bot" and "DEXCell EM without Eco-Bot"	
Parameter Description	This parameter compares the user's NPS rating of DEXCell EM with and without Eco-Bot. NPS is rated in a 0-10 scale where users rating: • 9 to 10 are promoters • 7 and 8 are passives/indifferent • 0 to 6 are detractors

Unit	Number (%)
Measurement Method	The Facility Manager (user) is regularly asked to give the NPS rating via DexCell EM Service. At the end of the pilot the FM/user will be asked to give the NPS rating for "DexCell EM with Eco-Bot" through the following question in a user survey: How likely is it that you would recommend "DEXCell EM with Eco-Bot" to a friend or colleague? This rating is compared with his previous NPS rating of "DEXCell EM without Eco-Bot" to see the improvement or deterioration on the score.
Time of evaluation	End of the pilot
Presentation of evaluation results	D 5.5 (M43)
WPs that the results will feed into and further utilization of results	WP7
Linkage to DoW's KPIs	None
Target	• More than 50% of users should give higher NPS rating for 'DEXCell EM with Eco-Bot' than for 'DEXCell EM without Eco-Bot' • 50% are promoters of Eco-Bot (NPS 9 to 10)

5 Enabled group level evaluations

While many of the parameters of the Eco-Bot impact evaluation metrics are geared towards evaluation on an individual level and constitute the basis of Eco-Bot's feedback to the user (e.g. energy savings achieved, money saved, CO_2 avoided etc.) these parameters also enable interesting group level evaluations.

These different groups can be the Eco-Bot segments[6]. In the business sector we can find out which type of building (supermarkets, hotels or restaurants) were affected most by the change. The evaluation can also be based on other groups than the segments. In the business sector a comparison could be made between the different countries in which the facilities are located. Likewise in the residential sector users could be grouped according to one shared characteristic (e.g. climate setting, income range or their status as owner or tenant). It would also be possible to compare the group of people from the EYPESA pilot with the smart meter (transmitting data every minute) to the group of people from the SEN pilot with the smart meter (transmitting data every 10 seconds) to see whether there are significant differences when it comes to the energy savings and thus conclude whether it is worth it to have the smart meter transmitting data every 10 seconds. Also a more exploratory approach could be chosen: Instead of verifying hypothesis, the group statistics could be performed accordingly to the situation of the data, which can be based on the answers of the following possible questions:

- Can a group be identified that achieves the most energy savings?
- Can a group be identified that achieves the least energy savings?
- Can a group be identified that interacts the most with Eco-Bot?
- Can a group be identified that interacts the least with Eco-Bot?

Moreover, group level evaluations could allow to draw conclusions relevant for the further roll out of Eco-Bot. For instance, if it turned out that the German households made more energy savings than the Spanish households and a closer examination shows that this is related to the fact that there is more electric heating in Germany and thus more energy saving potential in heating (which has a big impact on energy consumption), the conclusion would be that Eco-Bot might be of particular interest for Northern countries. Similarly, country based comparisons could allow to create a more culturally sensitive Eco-Bot. It could be, for example, that the evaluation shows that there are more behavioral changes carried out in Spain and more investments carried out in Germany. Or a particular recommendation or type of recommendation is popular in one or the other country and could thus in a more culturally sensitive update of Eco-Bot be more extensive. (The latter would not be implemented within the project duration but in a possible Eco-Bot 2.0).

[6] Please note that the segmentation aims to allow Eco-Bot to provide tailored advice, thus focuses on aspects like motivation and ability to enact change, rather than similar patterns of energy usage. Therefore, considering other groups in the evaluation opens the door for further findings and conclusions.

List of References

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003, April). Consolidating the ISO usability models. In Proceedings of 11th international software quality management conference (Vol. 2003, pp. 23-25).

AIMultiple (2019, January 1): 15 Key Metrics for Chatbot Conversational Analytics in 2019. Retrieved on July 22, 2019 from https://blog.aimultiple.com/chatbot-analytics/.

Bargas-Avila, J.A. and Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In Proc. Conference on Human Factors in Computing Systems. ACM.

Bundesnetzagentur (2019): Jahresbericht 2018. 20 Jahre Verantwortung für Netze. Retrieved on 14 August, 2019 from https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Allgemeines/Bundesnetzagentur/Publikationen/Berichte/2019/JB2018.pdf?__blob=publicationFile&v=6

ChatbotPack (2019). Why Chatbots Need Quality Standards? Retrieved on July 4, 2019 from https://www.chatbotpack.com/chatbot-quality.

Debecker, A. (2017, April 7): Chatbots don't actually have an engagement problem. Venturebeat. Retrieved on July 4, 2019 from https://venturebeat.com/2017/04/07/chatbots-dont-actually-have-an-engagement-problem/.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., & Lucas, G. (2014). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (pp. 1061-1068).

Discover.Bot (2019, January 1): Improving Your Chatbot Performance Through A/B Testing. Retrieved on July 2, 2019 from https://discover.bot/bot-talk/improving-chatbot-performance-through-ab-testing/.

EUR-Lex (2014): REPORT FROM THE COMMISSION Benchmarking smart metering deployment in the EU-27 with a focus on electricity./* COM/2014/0356 final */ Retrieved 14 August, 2019, from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1403084595595&uri=COM:2014:356:FIN

European Environmental Agency (EEA) (2016): Data and Maps. CO2 emission intensity from electricity generation, Retrieved on July 10, 2019 from https://www.eea.europa.eu/data-and-maps/daviz/sds/co2-emission-intensity-from-electricity-generation-2/download.table

Freire-González, J. (2017). Evidence of direct and indirect rebound effect in households in EU-27 countries. Energy Policy, 102, 270-276.

Georgila, K., Gordon, C., Choi, H., Boberg, J., Jeon, H., & Traum, D. (2018). Toward Low-Cost Automated Evaluation Metrics for Internet of Things Dialogues. In Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS).

Hassenzahl, M., Schöbel, M., & Trautmann, T. (2008). How motivational orientation influences the evaluation and choice of hedonic and pragmatic interactive products: The role of regulatory focus. Interacting with Computers, 20(4-5), 473-479.

Hoque, M.E., Courgeon, M., Martin, J.C., Mutlu, B., & Picard, R.W. (2013, September). Mach: My automated conversation coach. In Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (pp. 697-706). ACM.

Jain, M., Kumar, P., Kota, R. and Patel, S.N. (2018). Evaluating and Informing the Design of Chatbots. In Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18). ACM, New York, NY, USA, 895-906.

Kaleem, M., Alobadi, O., O'Shea, J., & Crockett, K. (2016). Framework for the Formulation of Metrics for Conversational Agent Evaluation. In RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme, May 28 (p. 20).

Kocaballi, A. B., Laranjo, L., & Coiera, E. (2018). Measuring user experience in conversational interfaces: a comparison of six questionnaires. In R. Bond, M. Mulvenna, J. Wallace, & M. Black (Eds.), Proceedings of 32nd BCS Human Computer Interaction Conference (HCI 2018) (pp. 1-12). Belfast: BCS, The Chartered Institute for IT.

Kuligowska, K. (2015). Commercial Chatbot: Performance Evaluation, Usability Metrics and Quality Standards of Embodied Conversational Agents. Professionals Center for Business Research. 2. 1-16.

Laugwitz, B., Held, T., & Schrepp, M. Construction and evaluation of a user experience questionnaire. (2008) In: Holzinger, A. (Ed.), Proceedings of the 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society (USAB 2008). LNCS 5298, pp. 63-76. Graz, Austria.

Law, E.L.C., Roto, V., Hassenzahl, M., Vermeeren, A.P. and Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In Proc. Conference on Human Factors in Computing systems (pp. 719-728). ACM.

Ledo, S. (2018): España ya mide la luz con contadores inteligentes. Retrieved on 14 August, 2019 from elPériodico.com: https://www.elperiodico.com/es/economia/20181021/espana-mide-luz-contadores-inteligentes-7098751

Lee, J. (2018, June 28): The practical guide to chatbot metrics and analytics. Growthbot. Retrieved on July 4, 2019 from https://blog.growthbot.org/the-practical-guide-to-chatbot-metrics-and-analytics.

Merritt, A. (2017, April 10): Examining the data on chatbot retention. Venturebeat. Retrieved on July 4, 2019 from https://venturebeat.com/2017/04/10/examining-the-data-on-chatbot-retention/.

Moro, A., & Lonza, L. (2018). Electricity carbon intensity in European Member States: Impacts on GHG emissions of electric vehicles. Transportation Research Part D: Transport and Environment, 64, 5-14.Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv:1704.04579.

Neff, A. (2019, February 12): Metrics uncover and track the value of chatbots within customer service. ICMI. Retrieved on July 18, 2019 from https://www.icmi.com/Resources/2019/metrics-uncover-the-value-of-chatbots-within-customer-service.

Newlands, M. (2017, June 13): 10 Metrics Your Chatbot Should Track to Optimize User Experience. Entrepreneur Media, Inc. Retrieved on July 2, 2019 from https://www.entrepreneur.com/article/295447.

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. Software Quality Professional, 19(3), 25-36.

Rauschenberger, M., Schrepp, M., Cota, M. P., Olschner, S. & Thomaschewski, J. (2013). Efficient measurement of the user experience of interactive products. How to use the user experience questionnaire (UEQ). Example: Spanish language version. International Journal of Interactive Multimedia and Artificial Intelligence. Special Issue on Artificial Intelligence and Social Application, 2(1), 39-45.

Sarstedt, M., Diamantopoulos, A., Salzberger, T., & Baumgartner, P. (2016). Selecting single items to measure doubly concrete constructs: A cautionary tale. Journal of Business Research, 69(8), 3159-3167.

Silvervarg, A. & Jönsson, A. (2011). Subjective and objective evaluation of conversational agents in learning environments for young teenagers. In Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems. Barcelona, Spain.

Stefanidi E., Foukarakis, M., Arampatzis D., Korozi, M., Leonidis A., & Antona M. (2019). ParlAmI: A Multimodal Approach for Programming Intelligent Environments. Technologies, Special Issue: The PErvasive Technologies Related to Assistive Environments (PETRA), 7(1), 11.

Umweltbundesamt (UBA 2014): Rebound effects. Retrieved on 14 August, 2019 from https://www.umweltbundesamt.de/en/topics/waste-resources/economic-legal-dimensions-of-resource-conservation/rebound-effects

Umweltbundesamt (UBA 2018): Entwicklung der spezifischen Kohlendioxid- Emissionen des deutschen Strommix in den Jahren 1990 - 2018. Retrieved on 19 June, 2019 from https://www.umweltbundesamt.de/themen/co2-emissionen-pro-kilowattstunde-strom-sinken)

Valtolina, S., Barricelli, B. R, Di Gaetano, S., & Diliberto, P. (2018, May). Chatbots and Conversational Interfaces: Three Domains of Use". In Proceedings of the 5th International Workshop on Cultures of Participation in the Digital Age: Design Trade-offs for an Inclusive Society (CoPDA 2018). Castiglione della Pescaia, Italy.

Vasconcelos, M., Candello, H., Pinhanez, C., & dos Santos, T. (2017, October). Bottester: Testing Conversational Systems with Simulated Users. In Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems (p. 73). ACM.

Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia B., Metallinou, A., Yang, S., Raj, A. & Goel, R. (2018). On Evaluating and Comparing Open Domain Dialog Systems. arXiv preprint arXiv:1801.03625.

Wang, Z., Han, B., & Lu, M. (2016). Measurement of energy rebound effect in households: evidence from residential electricity consumption in Beijing, China. Renewable and Sustainable Energy Reviews, 58, 852-861.

Yao, M. (2017, October 2): Guiding principles to maximize chatbot retention and engagement. Thinkgrow. Retrieved on July 4, 2019 from https://thinkgrowth.org/guiding-principles-to-maximize-chatbot-retention-engagement-c03b168dab1

ANNEX 1

P42 – Number of policies examined	
Parameter Description	This indicator shows the extensiveness of the preparatory work with regards to existing policies (European Union Energy Policies, Policies regarding renewable energy and energy efficiency and Energy Policies of IEA countries (mainly between 2000 and 2018)) in order to create the model of customer behaviour for Eco-Bot purposes.
Unit	Number
Measurement Method	The number of policies that have been examined by KAT are counted.
Time of evaluation	D2.2
Presentation of evaluation results	Annex of D3.3
Foreseen further actions based on evaluation results	None
Linkage to DoW's KPIs and targets	From Section 1.1: KPI_1.1: The number of policies, strategies, scientific papers and energy efficiency models examined together with the number of identified factors relevant for creating the model of customer behaviour for eco-bot purposes Target: More than 100 energy efficiency models, research papers, projects, policy guidelines are examined (Note this target is an overall target for P42-P46)
Result of evaluation	500 different sources were analysed – including academic papers, books, **policy guidelines** and technical reports involving online databases of scientific research. More than 50 existing energy efficiency models were identified and analysed.

P43– Number of strategies examined	
Parameter Description	This indicator shows the extensiveness of the preparatory work with regards to existing strategies (National Energy Strategies, EU White and Green Papers, Action Plans, European directives, reports and roadmaps) in order to create the model of customer behaviour for Eco-Bot purposes.
Unit	Number
Measurement Method	The number of strategies that have been examined by KAT are counted.
Time of evaluation	D2.2
Presentation of evaluation results	Annex of D 3.3

Foreseen further actions based on evaluation results	None
Linkage to DoW's KPIs and targets	KPI_1.1: The number of policies, strategies, scientific papers and energy efficiency models examined together with the number of identified factors relevant for creating the model of customer behaviour for eco-bot purposes Target: More than 100 energy efficiency models, research papers, projects, policy guidelines are examined *(Note this target is an overall target for P42-P46.)*
Result of evaluation	500 different sources were analysed – including academic papers, books, policy guidelines and technical reports involving online databases of scientific research. More than 50 existing energy efficiency models were identified and analysed.

P44 – Number of scientific papers examined	
Parameter Description	This indicator shows the extent of the preparatory work with regard to the analysis of various material sources - including scientific articles, books and technical reports including online research databases (such as WoS and ScienceDirect). In addition, it also refers to statistical data obtained from the following databases: MURE, OECD, IRENA, EIA, Eurostat and GUS.
Unit	Number
Measurement Method	The number of scientific papers that have been examined by KAT are counted.
Time of evaluation	D2.2
Presentation of evaluation results	Annex of D3.3
Foreseen further actions based on evaluation results	None
Linkage to DoW's KPIs	KPI_1.1: The number of policies, strategies, scientific papers and energy efficiency models examined together with the number of identified factors relevant for creating the model of customer behaviour for eco-bot purposes Target: More than 100 energy efficiency models, research papers, projects, policy guidelines are examined *(Note this target is an overall target for P42-P46)*
Result of evaluation	500 different sources were analysed – including **academic papers, books**, policy guidelines and technical reports involving online databases of scientific research. More than 50 existing energy efficiency models were identified and analysed.

P45 – Number of energy efficiency models examined	
Parameter Description	This indicator shows the extensiveness of the preparatory work with regards to existing energy efficiency models in order to create the model of customer behaviour for Eco-Bot purposes.
Unit	Number
Measurement Method	The number of energy efficiency models that have been reviewed by KAT are counted.
Time of evaluation	D2.2
Presentation of evaluation results	Annex of D3.3
Foreseen further actions based on evaluation results	None
Linkage to DoW's KPIs and targets	KPI_1.1: The number of policies, strategies, scientific papers and energy efficiency models examined together with the number of identified factors relevant for creating the model of customer behaviour for eco-bot purposes Target: More than 100 energy efficiency models, research papers, projects, policy guidelines are examined *(Note this target is an overall target for P42-P46.)*
Result of evaluation	500 different sources were analysed – including academic papers, books, policy guidelines and technical reports involving online databases of scientific research. More than **50 existing energy efficiency models were identified** and analysed.

P46 – Number of identified factors relevant for creating the model of user behaviour for Eco-Bot purposes	
Parameter Description	This indicator shows how many factors were identified as relevant for creating the model of customer behaviour for Eco-Bot. These factors were derived from the above mentioned work steps.
Unit	Number
Measurement Method	The number of factors that have been identified as relevant for creating the model of user behaviour by KAT are counted.
Time of evaluation	D2.2
Presentation of evaluation results	Annex of D3.3
Foreseen further actions based on evaluation results	None
Linkage to DoW's KPIs	KPI_1.1: The number of policies, strategies, scientific papers and energy efficiency models examined together with the number of identified factors relevant for creating the model of customer behaviour for eco-bot purposes

	Target: More than 100 energy efficiency models, research papers, projects, policy guidelines are examined *(Note this target is an overall target for P42-P46.)*
Result of evaluation	500 different sources were analysed – including academic papers, books, policy guidelines and technical reports involving online databases of scientific research. More than 50 existing energy efficiency models were identified and analysed.

P47 – Number of target groups (segments) successfully identified within the use cases and mapped on the behavioural model that was identified as the most relevant for Eco-Bot	
Parameter Description	This indicator measures how many segments were successfully identified and mapped onto the behavioural model that was identified the most relevant for Eco-Bot.
Unit	Number
Measurement Method	The number of target groups (segments) that were successfully identified (D 3.2) and mapped onto the behavioural model as proposed in D 2.2. are counted.
Time of evaluation	D2.3 and D3.2
Presentation of evaluation results	D3.2
Foreseen further actions based on evaluation results	None
Linkage to DoW's KPIs and targets	KPI_2.1: Number of target groups (segments) successfully identified within the use cases and mapped on the behavioural model that was identified as the most relevant for Eco-Bot Target: More than 40 clusters are created through Eco - Bot processing of the pool of the collected data.
Result of evaluation	46, composed of 25 segments for residential households and 21 for commercial facilities