Metrological evaluation and testing of robots in international competitions

Deliverable title	D3.1: HEART-MET Evaluation Plan
Deliverable lead	BRSU
Related task(s)	
Author(s)	Nico Hochgeschwender, Santosh Thoduka (BRSU) Mauro Dragone (HWU) Praminda Caleb-Solly, Daniel Bellamy (UWE) Filippo Cavallo (UNIFI)
Dissemination Level	Public
Related work package	WP3 : Healthcare
Submission date	30th June, 2020
Grant Agreement #	871252
Start date of project	1st January, 2020
Duration	36 months
Abstract	This document describes the evaluation plan for HEART-MET, which are robotics competitions in healthcare domain.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Versioning and Contribution History

VERSION	DATE	MODIFIED BY	MODIFICATION REASONS
V0	June 2020	Santosh Thoduka	First Version

List of Abbreviations and Acronyms

ABBREVIATION	MEANING
FEC	Field Evaluation Campaign
CEC	Cascade Evaluation Campaign
FBM	Functionality Benchmark
TBM	Task Benchmark
IoU	Intersection over Union
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Executive Summary

This document describes the evaluation plan for HEART-MET, which are robotics competitions in healthcare domain. The methodology for defining the competition and benchmarking procedures builds on previous robotics competitions such as RoCKIn, RoCKEU2 and SciRoc. Two types of competitions are held: namely field evaluation campaigns which are conducted at physical test-beds, and cascade evaluation campaigns which are conducted online using datasets collected during the field evaluation campaigns. The format of the field evaluation campaign consists of two types of benchmarks: Functionality and Task benchmarks (FBMs and TBMs). While FBMs benchmark individual standalone functionalities of the robot, TBMs benchmark the ability of the robot to combine several functionalities to complete a full task.

The benchmarks are defined based on their relevance to healthcare, and will be validated by a survey answered by relevant stakeholders about their relevance, criteria for evaluation etc. Examples include object detection, object handover, activity recognition, speech understanding and item delivery.

For each benchmark, several variations are defined to account for varied outcomes, particularly when the functionality is prone to failure, or is dependent on external factors such as interactions with humans. The variations aim to assess the robot's robustness, and, by evaluating all teams with the same variation ensures fairness across teams. The benchmarks are evaluated both using metrics such as accuracy and F1-score, and achievements, which serve as checkpoints for different aspects of the task or functionality. Data collected during the field campaigns will be used for benchmarking individual functionalities on static datasets and allows teams to participate online without the need for a robot. The data will additionally be used to evaluate usability and social acceptance of robots.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Contents

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

1 Introduction

The METRICS project organises robotics competitions in four priority areas, namely: Healthcare, Inspection and Maintenance, Agri-Food and Agile Production. The main goal of the competitions is to evaluate robots using objective metrological principles in each of the priority areas.

This document describes the evaluation methodology for healthcare competitions, HEART-MET, the primary objectives of which are:

- identification of relevant tasks which can be performed by a robot in a healthcare setting
- evaluation of standalone functionalities of robots which are required to perform these tasks
- evaluation of robotic systems as a whole for performing these tasks
- derivation of requirements for robotic systems to perform healthcare-related tasks based on the evaluation

Figure 1 illustrates the overall workflow of the HEART-MET evaluation and competition.

Healthcare was identified as one of the high-level market domains for robotics technology in the Multi-Annual Roadmap for robotics in Europe[1]. This is the primary motivation for selecting healthcare as one of the primary areas in the METRICS project. While the roadmap identifies three different sectors of robotics in the healthcare domain, HEART-MET primarily focusses on the assistive robotics sector.

Past and current projects such as RoCKIn, RockEU2 and SciRoc have already defined methods for conducting competitions for the purpose of benchmarking robots. In RoCKIn [1], two classes of evaluation were introduced: Task Benchmarks (TBM) and Functionality Benchmarks (FBM). These are explained in more detail in Section 2. The methodology aims to ensure both reproducibility and repeatability of the benchmarks conducted in the context of a competition. It was also recognized that the popularity and already established infrastructure for robotics competitions provides an opportunity to introduce more rigorous benchmarking of both robot systems and their individual functionalities.

In WP2, the overall methodology adopted in METRICS for conducting competitions has been defined. A common evaluation framework specifies several aspects, such as relevance to industry, modularity, and repeatability and reproducibility, all of which must be considered when defining an evaluation plan. The competitions from all four priority areas will adopt this framework in the creation of the evaluation plans.

In line with the primary objectives and the common evaluation framework, the relevance of defined benchmarks will be evaluated through a survey (see Appendix B). The survey targets several stakeholders such as healthcare workers, family, and people with physical, sensory or cognitive impairments. Feedback obtained from the survey will help refine existing benchmarks and define new ones. The survey will potentially also guide the data annotation process indicated in Figure 1.

The conducted competitions will be in two stages comprising of a Field Evaluation Campaign and a Cascade Evaluation Campaign, with the Field Evaluation Campaign consisting of FBMs and TBMs. These are described in more detail in Section 2.

The remainder of the document is structured in two parts, namely the evaluation methodology and the rulebook. In Section 2, we briefly describe the structure of the competitions, including the field and cascade evaluation campaigns and their tentative schedules. Section 3 describes the overall methodology for the creation of the evaluation plan, including industry relevance, aspects of user experience, scoring and dataset collection. In the second part of this document, Sections 4 and 5 describe the characteristics of the test bed and specifications and constraints of robots that can participate in the competition. Sections 6 and 7 describe a series of functionality and task benchmark which will be tentatively evaluated during the field campaigns. These are subject to change, as described in Section 3. Finally, Section 8 describes the process for the cascade evaluation campaigns. Appendix A consists of sample evaluation sheets that would be used by referees during the evaluation of functionality benchmarks. Appendix B includes the survey being sent to healthcare stakeholders for feedback on task relevance and suggestions for robotic tasks in healthcare.

[1] https://www.eu-robotics.net/sparc/upload/Newsroom/Press/2016/files/H2020_Robotics_Multi-Annual_Roadmap_ICT-2017B.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

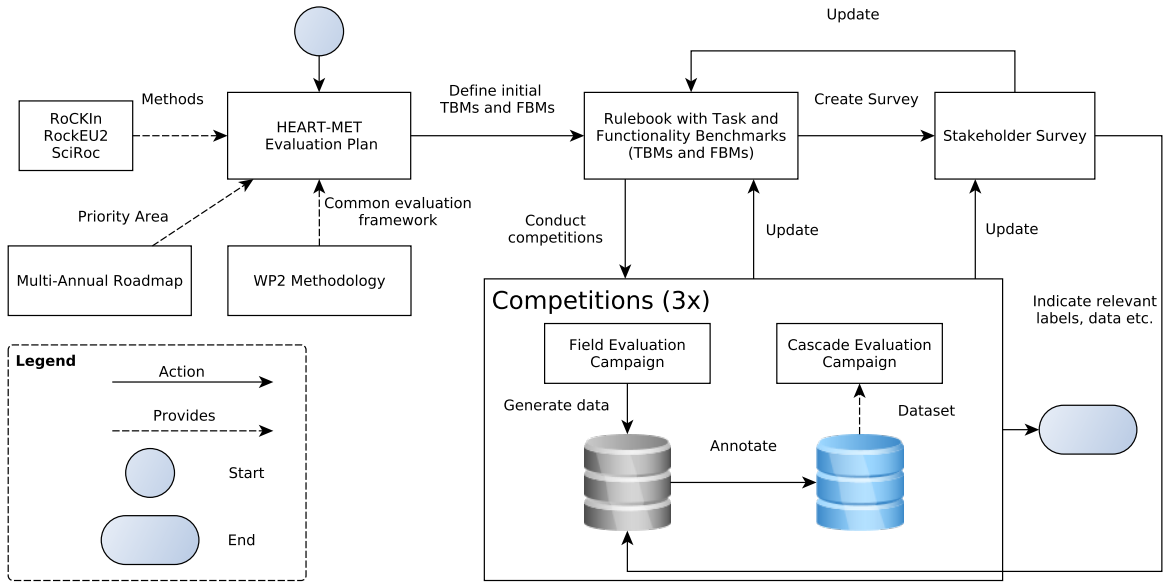


Figure 1: HEART-MET evaluation plan and competition workflow

Part I
Evaluation Methodology

2 Competition Structure

The METRICS competition is composed of two major components: the field evaluation campaigns (FEC) and the cascade evaluation campaigns (CEC). While the FECs evaluate the capabilities of the robots to perform functions and tasks at a physical test bed, the CECs allow teams to evaluate individual functionalities on datasets, without the requirement for a physical robot or test bed. In METRICS, a dry-run of both types of campaigns will be conducted in the first year, with two official FECs and CECs being conducted for the remaining two years.

2.1 Field Evaluation Campaign

The field evaluation campaign takes place at a certified test bed. All teams must be present at the same physical test bed with their robots in order to compete. The typical duration of the campaign is one week, which includes time for setup and preparations.

The benchmarks evaluated during a FEC are of two types: Functionality Benchmarks and Task Benchmarks. The two complementary types of benchmarks evaluate both the individual functionalities of the robot and the ability of the robot to integrate several functionalities into a fully functional system to complete a task. Several past benchmarking projects, such as RoCKIn, RoCKEU2, have introduced and conducted the benchmarking competitions in this format.

The FBMs evaluate individual functionalities of the robot such as object detection, speech understanding, grasping etc. The TBMs evaluate the capability of the robot to execute tasks such as delivery of medicine to a person, which requires the robot to integrate functionalities such as object detection, grasping, navigation and person detection.

A core objective of both the FBMs and TBMs is to collect data which will be used for the cascade evaluation campaigns. This includes data such as RGB images, external video streams, results from the robot's components (such as detected objects) and proprioceptive sensors of the robot.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

2.2 Cascade Evaluation Campaign

The cascade evaluation campaign takes place in the months following the field evaluation campaign. It is conducted entirely online and does not require a physical robot for participation.

The objective of the CEC is to benchmark the performance of teams' software on datasets for different functionalities. All datasets will be those collected during the FEC. For example, images collected by the robots during the execution of the object detection FBM at a FEC will be used to evaluate the performance of all participants of the CEC on the object detection functionality. Additionally, teams competing in the FEC will be encouraged to make available their training data for the CEC.

A fundamental difference between the FEC and CEC is, of course, the absence of a robot in the CEC. This means that the results in the FEC and CEC cannot be directly compared, since in an FEC, teams can take advantage of flexibility of the robot to gain more information and react to dynamic events. For example, for the object detection functionality, the robot can choose to view the object from several viewpoints based on the currently available information. In contrast, in a CEC teams will work with static datasets. Along the same lines, some functionalities, such as the handover functionality cannot be evaluated in a CEC, since it requires actuation and interaction with a human. However, some aspects of this functionality can still be evaluated; for example, the ability of the robot to detect the item falling down during the handover. It is therefore essential to record all relevant data during the FEC, even if a particular task or functionality cannot be fully evaluated in the CEC.

2.3 Schedule

The tentative schedule for the field and cascade evaluation campaigns is shown in Figure 2. Due to the travel restrictions imposed by COVID-19, the dry-run of the FEC will take place in the laboratories of BRSU, UWE, HWU, and UNIFI, performed by robots in the respective labs. It will not include external teams. The two official campaigns in the subsequent years are expected to take place as scheduled. The CECs take place entirely online and take place a few months after the FECs.

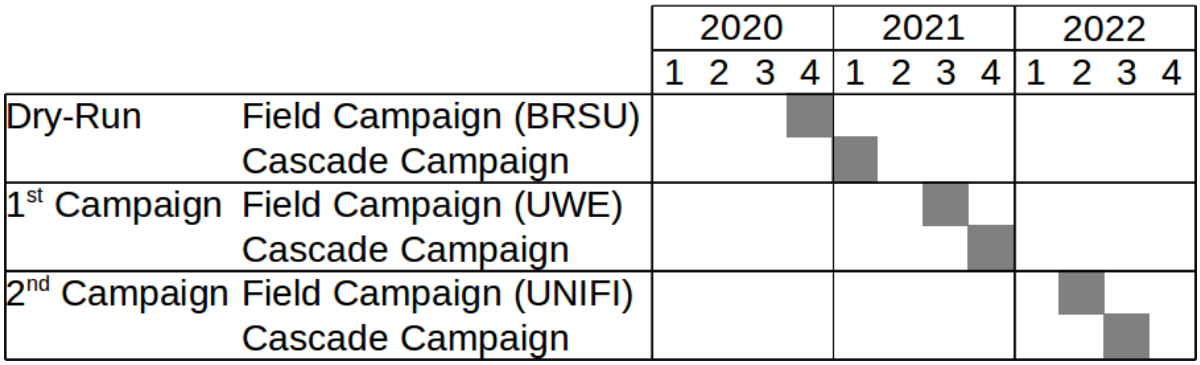


Figure 2: Estimated schedule of competitions

3 Evaluation Plan

In line with the evaluation framework described in D2.1, we define in this document the method of evaluation of each of the functionality and task benchmarks. In the definition of the benchmarks and their evaluation, we try to maximize the repeatability of the evaluations by clearly defining the allowed variations of independent variables for that benchmark. While it is necessary to evaluate the robots in a wide variety of conditions, it is also necessary to maintain fairness by evaluating all teams as equally as possible. Therefore, for a given run of a benchmark, controllable variations in the benchmark will be fixed beforehand and will remain uniform for all teams.

Since all benchmarks for the HEART-MET campaigns are conducted indoors, factors such as lighting, weather and ambient noise are expected to remain sufficiently uniform for all teams as long as they all compete during the same period of the day.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

During a competition, multiple runs of a benchmark can be performed, provided there is sufficient time. Depending on the benchmark, a given run can consist of multiple trials. For example, one run of the Object Detection benchmark can consist of ten trials, each of which requires the robot to detect one object. The set of variations for a run (in this case the objects that need to be detected and the configuration of other objects on the detection surface) will be fixed for all teams, though the order of the trials may differ across teams.

3.1 Industry Relevance

The set of benchmarks that have been defined for this competition have been selected based on their relevance for healthcare-related applications. Delivering medicines, preparing and transporting drinks, and assessing a person's activity state are all tasks that an assistive robot might be expected to perform, particularly to assist older adults with physical, sensory and cognitive impairments. In addition, patrolling or ensuring coverage of a given area is a task performed by a disinfectant robot. In order to accomplish these tasks, the robot must be able to detect objects, recognize humans, understand and generate speech, recognize human activities, grasp, receive and handover objects, pour drinks, open cupboards etc.

The functionality variations described in Section 3.2 ensure that the particular challenges that come with assisting persons with impairments are benchmarked. For example, one variation of the handover task involves a non-responsive human (i.e. the human does not reach out to receive the object). The ability of the robot to deal with such unexpected and safety-critical situations in a healthcare setting is crucial.

In order to validate the relevance of the functionalities, tasks and evaluation criteria, a survey has been prepared for the different stakeholders: namely end-users, healthcare workers, and family and friends of potential end-users. The survey (see Appendix B), consists of some general questions regarding the relevance of tasks, the importance of different robot-human interaction methods, levels of autonomy (based on the amount of interaction necessary), and open ended questions about possible tasks that would be useful for an assistive robot. Additionally, a video showing a typical task (delivery of an item to a person) is shown, and the user is asked about how the robot should respond to certain situations, and to list potential safety-critical situations that might occur.

The results of the survey will help refine the benchmarks and their evaluation criteria. Once the results of the survey are available, the evaluation plan will be updated to reflect the results. This includes adding or removing benchmarks and updating the metrics for evaluation if necessary.

3.2 Functionality Variations

In the FBMs defined in Section 6, two types of functionalities are defined: those in which the robot is simply sensing the environment (such as object recognition or speech understanding), and those in which the robot is actively interacting with the environment (such as grasping or handover). While the first category could still involve the robot moving in the environment (for example, to get a different viewpoint), there is no explicit need to interact with objects in the environment. This is relevant because in the second category, failures due to unexpected situations are more likely to occur. The dynamic nature of the environment increases the uncertainty involved in the functionality, leading to several branches of execution in case a failure were to occur. For example, during the handover functionality, the behaviour of the human significantly affects the success of the functionality. If the human does not reach out for the object, or if the object accidentally slips during the handover, the objective of the functionality will not be achieved.

In order to control and evaluate such events, we explicitly incorporate them into both the execution and the evaluation of the benchmarks. For example, in one (or more) trials of the handover functionality, the human will not reach out for the object. In these trials, the robot is awarded an achievement if it is able to detect that the human did not reach out for the object. In the FBM, this event signals the end of the benchmark, whereas in a TBM, the robot will be expected to continue with the task by, for example, initiating a dialogue with the human. All teams will be evaluated on the same set of variants of the functionality, which will be instantiated shortly before the start of the benchmark. The list of variations for each functionality are detailed in the individual sections in Sections 6 and 7.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

3.3 Autonomy Levels

Robots can operate in a healthcare setting with various levels of autonomy, ranging from fully autonomous to teleoperated. For task benchmarks, teams will be allowed to participate with varying levels of autonomy. However, fully-autonomous and non-autonomous executions of tasks will be categorised separately.

For example, in the Assess Activity State TBM, one robot could perform the task fully autonomously by recognizing the human, visually determining their activity state and engaging in a natural language dialogue to confirm their activity state. Another robot could be remotely controlled, with team members assessing the human's activity via the robot's camera and conversing by using a speaker and microphone on the robot. Both methods of executing the task are feasible and likely in a healthcare setting, though they address slightly different use cases. In the second case, a human is directly involved in assessing a person's activity state and can make immediate decisions based on their observations.

Since different autonomy levels require different skills and target different use-cases, the evaluations of the task benchmarks will categorized based on the autonomy level chosen by the teams. The procedure has been exemplified only in the Assess Activity State TBM 7.1, but is applicable to all task benchmarks.

3.4 Interactions with Humans

Some functionalities, such as activity recognition, speech understanding, handover etc., involve interactions with or observations of humans.

Whether it is performing an activity or behaving in a certain manner for a functionality variation, the human(s) involved must behave as similarly as possible for all teams to maintain fairness. We maintain uniformity for one run of a benchmark by selecting the set of volunteers beforehand, instructing them on how to behave and speak, and use the same set of volunteers for all teams. If several variations are benchmarked during a run, the same volunteers will perform their allotted variations for all teams.

While this guarantees, to some extent, the uniformity across teams during one benchmarking run, there is no guarantee of reproducibility at subsequent FECs in following years. In fact, in such cases where variations in behaviour are expected, it is actually desirable not to reproduce exact behaviour across FECs to evaluate the robustness of the robots.

3.5 Usability, Social Acceptance and User Experience

The ambition of METRICS is to maximise the compliance of robots with ethical, legal, social, economic requirements. In order to do so we aim to evaluate the potential societal impact of assistive robots used in the competition through conducting large-scale acceptability studies in EU countries, and integrating these findings into the competition evaluation plans. In the context of the HEART-MET competition, it is imperative to investigate and address societal barriers that may hinder the widespread adoption of robotic solutions employed in healthcare and social care settings. Research is needed to address concerns that currently limit uptake and acceptance by patients, clinical staff and care workers in general, such as affordability, trust, usability, possible reduction of the patient-carer relationship, loss of control, stigma, lack of familiarity, availability and lack of evidence to support their use.

The HEART-MET competition brings market offer and demand closer together, by providing guarantees regarding the technical and economic performance of these robotic systems, and well as addressing user acceptance issues and considering quality of the interaction between the robot and end-user.

Partners involved in the HEART-MET competition have home-like test environments designed to facilitate user-driven design and evaluation of innovative robotic solutions for healthy ageing and independent living. They also have well-established networks with beneficiaries and national and regional residential care providers, housing associations and Third Sector organisations. Utilising the laboratories and their supporting networks, the design of the HEART-MET competition follows the research concept methodology of a user-centred living lab, integrating concurrent research and innovation processes within a public-private-people partnership.

This methodology is embedded in the competition through a three-step process that will be refined throughout the different stages of the competition, iteratively:

- **Participatory Design**, through instruments such as the online questionnaire described in Section 3.1 and reported in Appendix B, to assess the approach and relevance of tasks performed by robots in

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

healthcare and social care settings, and build an understanding of factors influencing their acceptance by clinicians, nurses, caregivers and family, in addition to people in need of assistance. Findings will provide guidance on successive refinement of tasks and benchmarks. The use of online tools has been necessitated by the Covid-19 lockdown and the temporary closure of research laboratories. In ongoing work, feedback on the HEART-MET competition will be sought through actual workshops, focus groups and other participatory design exercises, including demonstrations of HEART-MET tasks to relevant stakeholders, which will be normally planned as part of laboratory activities once those the restrictions are relaxed.

- **Data Collection** during competition campaigns. The testbeds used in the HEART-MET competitions are designed to facilitate the creation of datasets capturing complex, interleaved and hierarchical naturalistic activities, collected in environments instrumented with a rich variety of sensors. These infrastructures will be used to collect benchmarking data (from external RGB cameras in addition to logs from robots' sensors, including proprioceptive sensor data from the robots' manipulators and base). The data will also include recordings of images and videos suitable for supporting participatory design and the evaluation of human-robot interaction, in addition to dissemination and outreach activities. A further opportunity will include exploiting the existing infrastructures and the replicability offered by our standardised set of test benchmarks, to provide more data on the humans interacting with the robots. This data will be enhanced, by using both commercial and research prototypes of ambient and environmental sensors, wearables including health and fitness devices, and also wireless device-free sensing (WDFS) prototypes (such as WiFi and Radar) for unobtrusive monitoring of behaviour and biometrics, which are commonly envisaged to work together with robots in the target application domain.

- **Evaluation of Human-Robot Interaction (HRI)** will follow widely accepted frameworks for the analysis and the evaluation of HRI which will contribute to the evaluation report after each competition campaign. In the first instance, the evaluation will consider different types of users and their different roles in interacting with the robots [2]. The primary user group considered in the analysis will include volunteers being assisted and directly interacting with the robot, e.g. through speech or, physically, as part of joint actions, as in tasks such as handover of objects. Additional volunteers in non-interacting roles will be recruited as appropriate, with designated observation spots, transparent walls and video streaming enabling the participation and involvement of people (members of the public) in a *bystander* role, i.e. people who have not received any formal training with the robots but who might normally co-exist in the same environment. Videos collected from the perspective of the bystander/observer roles will be used to enable video-based HRI evaluation methods [3]. All external volunteer and participant involvement will only proceed following the requisite risk assessment procedures. Furthermore, the team members participating to the competition with their robots will be themselves subjects of the evaluation, as they will be considered as *supervisors* and *operators/mechanics*. Information recorded during their participation will include; the type and the frequency of the interaction they had with their robot during each task, including instances of remote-teleoperation for situations that the robot was not able to address fully autonomously, and any re-start / intervention / re-programming on the fly that was necessary to recover from hardware, software and network failure or any other unanticipated events.

  Finally, a multi-level evaluation model, following the theoretical and methodological evaluation framework USUS [4] for user-centred evaluation, will be used to evaluate usability, social acceptance, user experience, and societal impact for task benchmark scenarios. The main goal of the USUS framework is to guide research in addressing issues relating to user experience of robots within the context of providing support for collaborative work. This enhances and their acceptability within different social and assistive contexts. The framework will be implemented by considering task trialled as part of user studies, where classic usability metrics as defined in ISO9241-11:2018[2] such as task completion rate, error rate and task duration, can be used to measure the effectiveness and the efficiency of the tested system, and combined with other methods, like the analysis of behavioural and biometric data, and also questionnaires or qualitative interviews to investigate factors related to usability (such as learnability,

[2] https://www.iso.org/obp/ui/#!iso:std:63500:en

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

flexibility, robustness, safety [5]), social acceptance (performance and effort expectancy, attitude toward technology, self-efficacy [6]) and user experience (emotional attachment, feeling of security, trust [7]). In addition, we will also ensure that the competing teams also have an opportunity to reflect on how their approaches, and robot designs and behaviours, address the ethical issues relating to assistive robots, such as those highlighted in BSI 8611:2016 Robots and robotic devices[3], which is a guide to the ethical design and application of robots and robotic systems. This will be achieved through organising a public debate during the competition period with members of the team and invited speakers on a panel.

3.6 Scores, Achievements, Penalties and Disqualifying Behaviours

For some benchmarks, such as object detection, human recognition etc., the accuracy or success is directly measurable using metrics such as intersection-over-union (IOU), precision, recall, etc. However, for other benchmarks, such as opening a cupboard or receiving an object, the measure of success is simply the successful execution of the task. In such cases, we define a set of intermediate checkpoints, each of which receives a number of achievement points on completion. For example, for receiving an object, one achievement is awarded for movement of the arm towards the human and one for grasping the object. Additional achievements are awarded for varying the height of the end-effector based on the pose of the human, and for detection of unexpected outcomes, such as the human not releasing the object.

In addition to scores or achievements, penalties and disqualifying behaviours are defined for each benchmark. These are typically undesired behaviours such as uncontrolled collisions, damaging objects and unresponsiveness. Penalties are considered when the scores or achievements result in a tie between teams. Disqualifying behaviours invalidate the trial. Teams will have the opportunity to rerun the benchmark in case of disqualification.

The metrics used for evaluation of each functionality and task benchmark are specified in Section 6 and 7.

3.7 Cascade Evaluation Campaign

The benchmarks evaluated during the cascade evaluation campaign are described in the individual functionality benchmarks in Section 6. The metrics for each benchmark in the CEC remain the same as those used in the FBM. The data collected for all benchmarks executed during the FEC will be annotated and used as a test set for the CEC.

3.8 Datasets

For FBMs and TBMs which require interaction with the environment or a human, the datasets will include both successful and failed executions by the robot. This will make the dataset useful for execution monitoring, particularly since the datasets will contain multimodal data. Failure cases are particularly relevant since the monitoring and correction of failures is critical when the robot is operating autonomously amongst humans.

For visual tasks, and tasks which involve physical interaction, RGB streams from the perspective of the robot, and an external camera stream will be a part of the dataset. Multi-view perception methods, evaluation of human-robot interaction, and visual execution monitoring will be possible with such data. With the participation of multiple teams, the datasets will be accumulated from robots with different morphologies and sensor configurations, hence encouraging robot-agnostic approaches to be explored during the CECs.

Since the number of teams participating and the amount of data collected is uncertain, it is not clear whether sufficient data will be available for both training and test purposes. Teams participating in the FEC will be encouraged to submit their training data; hence the datasets for training will be released depending on their availability.

[3]Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems, BS 8611:2016, British Standards Inst., 2016; `https://shop.bsigroup.com/ProductDetail?pid=000000000030320089`

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

3.9 Compliance with Common Evaluation Framework

Table 1: Compliance with common evaluation framework

Topic	Addressed	Details
Evaluation Plan		
The first occurrence is a dry run	Yes	specified in Section 2
The evaluation plan is formalized	Yes	evaluation plan presented in Section 3
Each evaluation task is relevant for industry	Yes	stakeholder survey
The dependent and independent variables of each evaluation are identified	Yes	defined in the individual FBMs and TBMs in Sections 6 and 7
The evaluation is modular	Yes	Sections 6 and 7
Testing Environment		
Repeatability and reproducibility of the observations are maximized	Yes	fixed set of variations defined in Section 3.2
The accessibility of the test beds is maximized	Yes	there is no specific requirement for test bed; any lab or home like environment is sufficient. Official test-beds are available for use as well.
The quality procedure is defined and implemented	Yes	see Sections 3.2 and 3.4
Measurements and estimates are clearly identified	Yes	see Section 3.6 and evaluation metrics under each FBM and TBM in Sections 6 and 7
Subjectivity is addressed in an appropriate way	Yes	see Sections 3.4 and 3.5
Metrics are properly designed	Yes	see Section 3.6 and evaluation metrics under each FBM and TBM in Sections 6 and 7

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Part II
Rulebook

4 The METRICS Testbed

4.1 Testbed Environment

The testbeds resemble typical living environments with areas such as a living room, dining room and kitchen. For example, the testbed at Heriot Watt consists of a 60m^2 simulated apartment with an open-plan living, dining and kitchen area, along with a bedroom and bathroom. Several sensors such as a motion capturing system, CCTV cameras, RFID floor and a device-free-sensing system for indoor localization and monitoring are also available. At UWE, the Assisted Living Studio is a modular apartment instrumented with a network of Z-wave sensors, Wi-Fi cameras PIR sensors etc. At BRSU, the test bed consists of a living room, dining room, a lounge area and a fully-functional kitchen.

While participating teams are not required to have access to such facilities for preparation, such test beds are open for teams to test their robots.

4.2 Objects in the Environment

The objects in the environment which the robot has to interact with or recognize include both general domestic objects and healthcare-related objects. Some examples of domestic objects include towels, cups, plates and cutlery, general food items, pillows, etc. Healthcare related objects include medicine boxes or bottles, insulin pen, first-aid kit, inhaler, crutches etc. The exact list of objects which need to be recognized or interacted with will be released along with the individual calls for participation for each field evaluation campaign.

4.3 Benchmarking Equipment in the Environment

4.3.1 Referee Box

Communication with the testbed will be done via a referee box. In particular, this will be used to indicate the start and end of a benchmark to the robot, and for the robot to send feedback if required by a benchmark.

The features of such a referee box are as follows:

- is able to communicate wirelessly to the robot, both to send and receive messages
- can be controlled by the referee to initiate or end a benchmark
- stores feedback sent by the robot
- records start and end time of each run

4.3.2 Benchmark Data Collection

METRICS benchmarking is based on the processing of data collected in two ways:

- **internal benchmarking data**, collected by the robot system under test, such as video and proprioceptive sensors ;
- **external benchmarking data**, collected by the equipment embedded into the testbed

The external benchmarking data collection equipment will include video cameras recording the runs, ambient sensors (if available) in the testbed and messages and timestamps recorded by the referee box.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

5 Robots and Teams

The content of this section has been adapted from the RoCKIn@Work rulebook[4].

The purpose of this section is twofold:

1. It specifies information about various robot features that can be derived from the environment and the targeted tasks. These features are to be considered at least as desirable, if not required for a proper solution of the task. Nevertheless, we will try to leave the design space for solutions as large as possible and to avoid premature and unjustified constraints.

2. The robot features specified here should be supplied in detail for any robot participating in the competition. This is necessary in order to allow better assessment of competition and benchmark results later on.

The description of the robot should be included in the team description paper.

5.1 General Specifications and Constraints on Robots and Teams

Robot Specification 5.1 (*System*)

A competing team may use a single robot or multiple robots acting as a team. It is not required that the robots are certified for industrial use. At least one of the robots entered by a team is capable of:

- *mobility and autonomous navigation.*
- *manipulate and grasp at least several different task-relevant objects. The specific kind of manipulation and grasping activity required is to be derived from the task specifications.*

The robot subsystems (mobility, manipulation and grasping) should work with the environment and objects specified in this rule book.

Robot Specification 5.2 (*Sensor Subsystems*)

Any robot used by a team may use any kind of ***onboard*** *sensor subsystem, provided that the sensor system is admitted for use in the general public, its operation is safe at all times, and it does not interfere with other teams or the environment infrastructure. A team may use the sensor system in the environment provided by the organizer by using a wireless communication protocol specified for such purpose. Sensor systems used for benchmarking and any other systems intended for exclusive use of the organizers are not accessible by the robot system.*

Robot Specification 5.3 (*Communication Subsystems*)

Any robot used by a team may ***internally*** *use any kind of communication subsystem, provided that the communication system is admitted for use in the general public, its operation is safe at all times, and it does not interfere with other teams or the environment infrastructure. A robot team must be able to use the communication system provided* ***as part of the environment*** *by correctly using a protocol specified for such purpose and provided as part of the scenario.*

Robot Specification 5.4 (*Power Supply*)

Any mobile device (esp. robots) must be designed to be usable with an onboard power supply (e.g. a battery). The power supply should be sufficient to guarantee electrical autonomy for a duration exceeding the periods foreseen in the various benchmarks, before recharging of batteries is necessary. Charging of robot batteries must be done outside of the competition environment. The team members are responsible for safe recharging of batteries. If a team plans to use inductive power transmission devices for charging the robots, they need to request permission from the event organizers in advance and at least three months before the competition. Detailed specifications about the inductive device need to be supplied with the request for permission.

[4] http://rockinrobotchallenge.eu/rockin_d2.1.6.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Robot Constraint 5.1 (*Computational Subsystems*)

> *Any robot or device used by a team as part of their solution approach must be suitably equipped with computational devices (such as onboard PCs, microcontrollers, or similar) with sufficient computational power to ensure safe autonomous operation. Robots and other devices may use external computational facilities, including Internet services and cloud computing to provide richer functionalities, but the safe operation of robots and devices may not depend on the availability of communication bandwidth and the status of external services.*

Robot Constraint 5.2 (*Safety and Security Aspects*)

> *For any device a team brings into the environment and/or the team area, and which features at least one actuator of any kind (mobility subsystems, robot manipulators, grasping devices, actuated sensors, signal-emitting devices, etc.), a mechanisms must be provided to immediately stop its operation in case of an emergency (emergency stop). For any device a team brings into the environment and/or the team area, it must guarantee safe and secure operation at all times. Event officials must be instructed about the means to stop such devices operating and how to switch them off in case of emergency situations.*

Robot Constraint 5.3 (*Operation*)

> *In the competition, the robot should perform the tasks autonomously. An external device is allowed for additional computational power. It must be clear at all times that no manual or remote control is exerted to influence the behavior of the robots during the execution of tasks.*

Robot Constraint 5.4 (*Environmental Aspects*)

> *Robots, devices, and apparatus causing pollution of air, such as combustion engines, or other mechanisms using chemical processes impacting the air, are not allowed. Robots, devices, and any apparatus used should minimize noise pollution. In particular, very loud noise as well as well-audible constant noises (humming, etc.) should be avoided. The regulations of the country in which a competition or benchmark is taking place must be obeyed at all times. The event organizers will provide specific information in advance, if applicable. Robots, devices, and any apparatus used should not be the cause of effects that are perceived as a nuisance to humans in the environment. Examples of such effects include causing wind and drafts, strong heat sources or sinks, stenches, or sources for allergic reactions.*

5.2 Benchmarking Equipment in the Robots

Hardware

- Teams might have to install a USB-stick during the runs for storing the data.
- The robots need to have WiFi-connectivity for communication with the Referee Box

Software

- The robot needs to have the software packages to run *rosbag record*
- A ros package will be provided which makes it easier to trigger starting and stopping the recording of bagfiles autonomously

Recorded Data The data required to be recorded internally by the robot is dependent on the FBM or TBM. Some common data streams that must be recorded (if available) include the following:

- base velocity commands
- odometry
- TF tree
- joint states

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Table 2: Relation between TBMs and FBMs and Cascade Campaign

FBM	TBM	Cascade Campaign
Object Detection	Item Delivery Prepare A Drink	Object Detection
Human Recognition	Assess Activity State Area Coverage Receive and Transport a Drink	Human Recognition
Activity Recognition	Assess Activity State	Activity Recognition
Gesture Recognition		Gesture Recognition
Task-oriented Grasping	Item Delivery Prepare a Drink	Grasp Verification
Handover	Item Delivery	Detection of (un)successful handover
Receive Object	Receive and Transport Drink	Detection of (un)successful receive
Speech Understanding	Assess Activity State	Speech Understanding
Opening Cupboard		Detection of (un)successful opening
Pouring	Prepare a Drink	Detection of (un)successful pour
Area Coverage	Area Coverage	

- camera (RGB, depth, camera calibration)
- sound
- laser (if applicable for the benchmark)
- other sensors such as force-torque, tactile, IR

While a particular sensor might not be used by the teams for a particular task, it is preferable to record it since it might be used by competing teams during the cascade evaluation campaigns. It is expected that participating robots will be heterogeneous, therefore the exact set of variables and sensors that will be recorded will be determined on site per robot.

6 Functionality Benchmarks

This chapter describes the functionality benchmarks which are meant to evaluate specific, standalone capabilities of a robot. We provide a general description of the functionality, the variations in terms of the independent variables for that functionality, communication with the referee box, procedure for conducting the benchmark and finally the evaluation criteria for the benchmark.

The variations mentioned in the benchmark are specific controllable aspects of the task that will be selected by the referee before the benchmark begins and remains fixed for all teams. Besides these variations, it is expected that factors such as lighting conditions, locations and volunteers used during the benchmark are also varied without prior specification.

While the evaluation criteria mentioned in each benchmark is the primary method of comparing performance, the penalties and eventually the duration for executing the benchmark is taken into account in case there is a tie between teams; i.e. if the scores are tied, the team with lower penalties is ranked higher and if the penalties are also tied, the team with lower execution time is ranked higher.

The following sections describe several functionality benchmarks with their associated metrics and procedure for execution. For a particular campaign, a subset of the FBMs will be selected based on expected team participation and healthcare relevance. Some of the following FBMs have not been fully specified. Their metrics and procedures will be defined once their relevance and metrics for benchmarking have been clarified through the survey to stakeholders.

Table 2 shows the mapping between FBMs and TBMs and benchmarks evaluated in the cascade evaluation campaign.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.1 Object Detection Functionality

6.1.1 Functionality Description

This functionality benchmark assesses the robot's capability of locating a target object in a given location. A common task for a healthcare robot is to locate a particular object, possibly in a particular location. In typical object detection benchmarks, the task is to detect all objects in a given image. Here, we instead require the robot to find a particular object among a set of objects. Therefore, one of the possible variations in this benchmark is one in which the target object is not present at the target location.

Several secondary objects are placed on a flat surface with no minimum distance between objects. The target object is either included in this set of objects or not, depending on the variation selected for a particular trial. The robot is placed in front of the location, and must locate the target object if it exists, or indicate that the target object is not present.

6.1.2 Healthcare Relevance

Several tasks for a healthcare robot require the robot to locate and fetch an item. Such items could include medicines, reading glasses, a walking cane etc. The task would typically involve moving to a known location where the item usually is, detecting it, grasping it and delivering it to a person. This functionality hence deals with only detecting a target item, once the robot is already at the location where it expects the item to be.

6.1.3 Feature Variation

The independent variables for this functionality are:

- the set of objects and their poses
- is the target object present [yes, no]

The dependent variable is the detected pose of the object, or the detection of no object.

6.1.4 Communication with the Referee Box

For each trial in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the location of the object (or indicate the object is not found)
- In case of a timeout, the referee box sends a stop message to the robot

6.1.5 Procedures and Rules

A set of target objects, secondary objects and their poses is specified and fixed for all teams.

The maximum time allowed for one trial is 20 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot sends the result message. If 20 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

The robot has the option of specifying the location of the object in one of two ways: a) the 3D bounding box of the object with respect to the robot's base b) the 2D bounding box of the object in an image. In the first case, the robot must provide, in the benchmarking data, a point cloud of the scene (transformed to the robot's base frame) corresponding to the provided 3D bounding box of the object. The 3D bounding box must also be in the robot's base frame. In the second case, the robot must provide the raw RGB image of the scene corresponding to the provided 2D bounding box. Only a single bounding box per point cloud / image must be specified. If multiple bounding boxes are specified, only the first one is considered.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.1.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Point cloud of scene (if 3D bounding box is provided)
- RGB image of the scene (if 2D bounding box is provided)

In addition, the teams should provide the data used for training machine learning models (if any).

6.1.7 Metrics for Evaluation

We calculate the following four metrics for evaluating the performance of the robot:

- True Positive (TP): detects target object when it is present
- False Positive (FP): detects wrong object whether target object is present or not
- False Negative (FN): does not detect any object when target object is present
- True Negative (TN): does not detect any object when target object is not present

These metrics are illustrated in Figure 3.

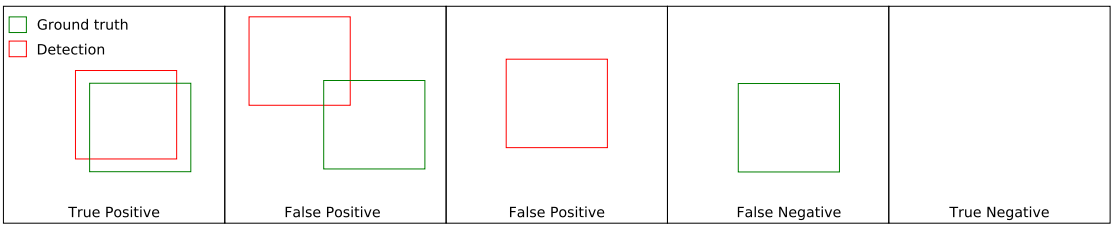


Figure 3: Metrics for object detection

For TP and FP, a measurement of the degree of similarity between the ground truth and detected bounding boxes is required. Here we use the Jaccard Index, also known as the Intersection over Union (IoU),

$$IoU = \frac{B_{GT} \cap B_{DET}}{B_{GT} \cup B_{DET}} \tag{1}$$

where B_{GT} and B_{DET} are the ground truth and detected bounding boxes respectively, and the intersections are calculated as areas and volumes for 2D and 3D bounding boxes respectively. In the example in Figure 4, the green box shows the ground truth bounding box, and the red box shows the predicted bounding box. If the IoU is greater than a threshold, the detection is considered to be a true positive, and a false positive if the IoU is lower than the threshold. The same applies to 3D bounding boxes, except that the areas will be replaced by volumes.

Since the IoU threshold is a configurable parameter, we consider a range of thresholds from 0.5 to 0.95 with a step size of 0.05 (this is the approach taken by the COCO challenge[5]). The metrics TP, FP are then averaged over all IoU thresholds.

Teams will be ranked based on:

- the sum of the TP and TN;
- in case of a tie, the team with a lower FP count is ranked higher;
- in case teams are still tied, the team with the lower FN count is ranked higher.

[5] http://cocodataset.org/#detection-eval

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

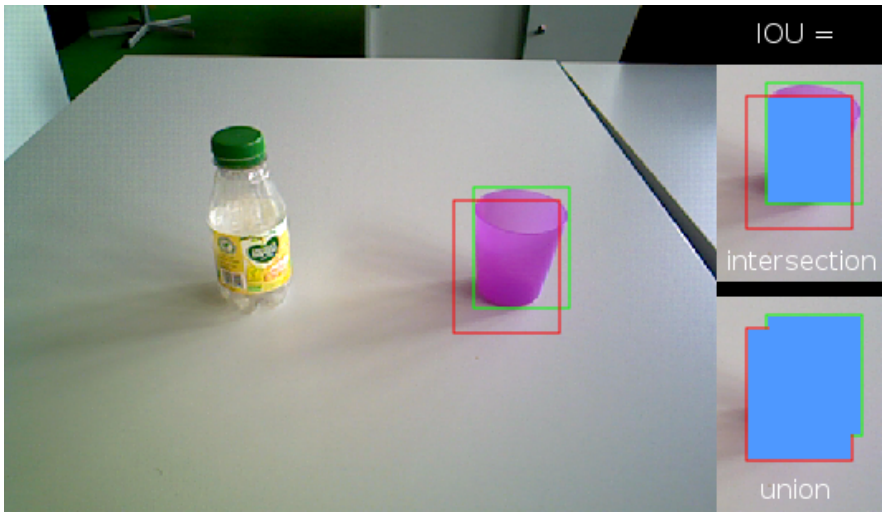


Figure 4: Example object detection; the green box shows the ground truth, and the red box shows the detection

It must be noted that the reason for considering FN a lower priority compared to FP is that it is preferable for a robot not to detect an object than to incorrectly detect an object. In the former case, the robot can simply retry detecting the object, whereas in the latter case, it might result in a task failure (for example, the robot might deliver an incorrect medicine to the person).

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

6.2 Human Recognition Functionality

6.2.1 Functionality Description

This functionality benchmark assesses the robot's capability of recognizing people that it might regularly encounter. The set of people that the robot must recognize is known beforehand, as they represent household occupants, caretakers, relatives etc. For each trial, the robot is expected to recognize the person (or identify them as unknown) by observing them from a predefined maximum distance.

6.2.2 Healthcare Relevance

Robots operating in a home or healthcare environment must be able to recognize people that they are assisting and other frequent visitors or inhabitants of the home. This is necessary since the robot must adapt its behaviour or task based on the person whom it is interacting with.

6.2.3 Feature Variation

The robot is expected to recognize a fixed number of people (such as 4 - 8 people). Additional unknown persons can be presented for a trial. However, for each trial only one person will be present in front of the robot. The independent variables that will be varied for each execution are:

- Distance of the person to the robot

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

- Pose of the person [standing, sitting, laying]
- Pose of the person's face with respect to the robot [straight, 30° left, 30° right]
- Presence of eye wear
- Presence of face mask
- Presence of head covering (such as a cap)

The dependent variable is the identity of the person.

6.2.4 Communication with the Referee Box

For each trial in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the identity of the person (or unknown) in front of it
- In case of a timeout, the referee box sends a stop message to the robot

6.2.5 Procedures and Rules

The teams are allowed to take photos of the people who need to be recognized at least two hours before the start of the benchmark. There is no guarantee that the person will wear the same clothes or other accessories such as head coverings during the official runs.

The configurations of all trials in a run are selected and fixed for all teams. If multiple runs are executed during a competition, new configurations are selected.

The maximum time allowed for one trial is 10 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot sends a message with the identity of the person. If 10 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

6.2.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot

In addition, the teams should provide the data used for training machine learning models (if any).

The data recorded during the runs will be used in the Cascade Evaluation campaigns to evaluate the capability of teams to recognize people.

6.2.7 Metrics for Evaluation

The performance of the robot is measured by the F1-score. In the following definitions, we use GT to notate ground truth and RB to notate the output of the robot. We also use X to indicate the class whose metric is being calculated, and $\bar{X}$ to indicate any class apart from X. The classes include the N persons whom the robot should recognize and a class for unknown persons. The following are calculated per class:

- True positive (TP): total count where $GT = X$ and $RB = X$
- False positive (FP): total count where $GT = \bar{X}$ and $RB = X$
- False negative (FN): total count where $GT = X$ and $RB = \bar{X}$

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Based on these definitions, we calculate the F1-score as:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{2}$$

where

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \tag{3}$$

If multiple runs are performed, the F1-score is calculated using the sum of the TP, FP and FN over all runs.

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

6.3 Activity Recognition Functionality

6.3.1 Functionality Description

This functionality benchmark assesses the robot's capability of recognizing the activities of a human. The robot is placed in front of a human who performs an activity. The robot needs to recognize the activity being performed by the human.

6.3.2 Healthcare Relevance

When a robot is operating in an assistive capacity, it is often required to monitor the state of the person who has caring needs. Such monitoring could include detecting when the person is sleeping, detecting a fall, or simply other daily living activities. The robot could then make decisions based on the detected activities; for example by calling for help if a person has fallen down.

6.3.3 Feature Variation

The human activity (dependent variable) will be chosen from a list consisting of:

- All classes in the Charades dataset[6] [8], which consist of daily living activities
- A set of activities which are relevant in the healthcare context, listed below:
  - Coughing
  - Walking on crutches
  - Sitting in a wheelchair
  - Moving from a couch to a wheelchair
  - Falling down
  - Limping
  - Hopping
  - Reacting to getting hurt
  - Colliding against furniture

For a benchmark run, a fixed number of activities will be selected and performed by a set of actors, which the robot has to recognize. To maintain uniformity, the same actors will perform a given activity for all teams. The independent variables that could be varied for each execution are:

- Distance of the person to the robot
- Pose of the person [standing, sitting, laying]

[6] https://prior.allenai.org/projects/charades

6.3.4 Communication with the Referee Box

For each activity in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the classified activity label
- In case of a timeout, the referee box sends a stop message to the robot

6.3.5 Procedures and Rules

The referee selects the list of activities, their locations, distances to the robot, and poses of the person. The human actor performs the activity when the robot confirms it has received the start message.

The maximum time allowed for classifying one activity is 20 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the activity class. If 20 seconds is exceeded, a timeout is recorded for that activity, and the robot must prepare for the next activity.

6.3.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Classified activity label

In addition to the recorded internal robot data, an external RGB camera will record each run.

6.3.7 Metrics for Evaluation

The performance of the robot is based on the following metrics:

1. True positive verbs (TP_n): correctly identified verbs
2. True positive nouns, if any (TP_n): correctly identified nouns

In addition, the overall true positive rate are calculated as:

1. True positive rate for verbs $TPR_v = \frac{TP_v}{N}$
2. True positive rate for nouns $TPR_n = \frac{TP_n}{N_n}$

where N is the total number of trials, and N_n is the number of trials where the activity includes a noun.

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

6.4 Gesture Recognition Functionality

6.4.1 Functionality Description

This functionality benchmark assesses the robot's capability of recognizing gestures performed by a human. The robot is placed in front of a human who performs a gesture. The robot needs to recognize the gesture being performed by the human.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.4.2 Healthcare Relevance

Gestures are a form of interaction that a human might use to communicate with the robot. Especially when the human has an impairment, it might be necessary to communicate via gestures as opposed to speech. In such a case, a robot needs to be able to recognize common gestures such as waving, pointing etc.

6.4.3 Feature Variation

The gestures (dependent variable) will be chosen from a list consisting of:

- Waving for attention
- Waving to come closer
- Waving to go away
- Pointing to object
- Nodding
- Shaking head

For a benchmark run, a set of gestures will be selected and performed by actors, which the robot has to recognize. To maintain uniformity, the same actors will perform a given gesture for all teams. The independent variables that could be varied for each execution are:

- Distance of the person to the robot
- Pose of the person [standing, sitting, laying]

6.4.4 Communication with the Referee Box

For each gesture in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the classified gesture label
- In case of a timeout, the referee box sends a stop message to the robot

6.4.5 Procedures and Rules

The referee selects the list of activities, their locations, distances to the robot, and poses of the person.

The maximum time allowed for classifying one gesture is 20 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the gesture class. If 20 seconds is exceeded, a timeout is recorded for that gesture, and the robot must prepare for the next gesture.

The human actor performs the gesture when the robot confirms it has received the start message.

6.4.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Classified gesture label

In addition to the recorded internal robot data, an external RGB camera will record each run.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.4.7 Metrics for Evaluation

The performance of the robot is based on the following metrics:

1. True positive (TP): correctly identified gestures

In addition, the overall true positive rate are calculated as:

1. True positive rate $TPR = \frac{TP}{N}$

where N is the number of trials.

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment

6.5 Task-oriented Grasping Functionality

6.5.1 Functionality Description

This functionality benchmark assesses the robot's capability of grasping objects for a given task. Depending on the task, the robot might want to grasp an object from a certain orientation, and keep the object oriented a certain way. For example, in order to pour from a glass, it is most common to grasp the glass from the side, without covering the opening of the glass. Similarly, for transporting a glass of water, in most cases, it is safest to keep the glass vertically oriented.

Therefore, for this functionality benchmark, the robot will be tasked with grasping an object with some constraint (such as grasp orientation, or grasp location). While no constraints are placed on the orientation of the object, the robot must report its estimate of the object's orientation.

An object is placed on a flat surface, and the robot is placed in front of it. The grasp constraints are specified, and the robot needs to grasp the object and lift it for a minimum duration before placing it back on the surface.

Each benchmark run will consist of multiple trials, each with a different object and constraints.

6.5.2 Healthcare Relevance

Several domestic and healthcare related tasks require the robot to interact with objects in the environment. This could be, for example, picking up a medicine bottle from the cabinet, pouring water into a glass, transporting the glass of water etc. Depending on the task, different grasp configurations might be desirable. Therefore the robot must be capable of grasping objects with different constraints placed on grasping location and orientation. Additionally, to prevent accidents, the robot must be aware of the orientation of the object while executing the task (for example, to avoid spilling the water during transportation).

6.5.3 Feature Variation

The independent variables are the object to be grasped, the pose of the object on the flat surface and the constraints for the grasp.

The constraints will be selected from the following:

- Grasp orientation [top, side, front]
- Grasp location (object-dependent; for example, cup handle)

6.5.4 Communication with the Referee Box

For each trial in the run:

- The robot waits for a message specifying the object and the grasp constraints from the referee box
- The robot sends a confirmation that it has received the specification message
- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box once it has placed the object back on the surface
- In case of a timeout, the referee box sends a stop message to the robot

6.5.5 Procedures and Rules

A set of objects, their poses and grasp constraints are specified and fixed for all teams.

The maximum time allowed for one trial is 30 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot sends a message confirming the end of the trial. If 30 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

6.5.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Detected pose of the object with respect to the robot
- Selected grasp pose with respect to the object
- Continuous estimated pose of the object from when it is grasped

In addition to the recorded internal robot data, an external RGB camera will record each run, and an IMU attached to the object will measure the orientation of the object.

6.5.7 Metrics for Evaluation

The performance of the robot is based on the following achievements for each execution:

1. Grasp object [Achievements: 1]
2. Grasp object from correct orientation [Achievements: 1]
3. Grasp object at correct location [Achievements: 1]

In addition, we calculate the difference between the ground truth orientation (only with respect to the vertical axis) of the object and the estimated orientation of the object. The last two achievements will not be measured, but will be judged subjectively by the referees. For example, the referees must decide whether a cup was indeed grasped by the handle based on the observed contact points of the gripper with the cup.

Penalties

- the robot collides with environment in an uncontrolled manner
- the robot drops the object causing it to drop to the floor
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.6 Handover Functionality

6.6.1 Functionality Description

This functionality benchmark assesses the robot's capability of handing over objects to a person. An object is placed in the robot's gripper and the robot is placed in front of a person. The robot needs to hand over the object to the person and verify that the object has been received. A single run of the benchmark consists of ten trials, each with different variations.

6.6.2 Healthcare Relevance

An assistive robot is often tasked with bringing an item (such as medicine, reading glasses etc.) to a person. When delivering items to a human, the handover is a complicated interaction that requires awareness from both the human and the robot about each other's intentions in order to be successful.

6.6.3 Feature Variation

The configuration for a single trial consists of the assignment of the following independent variables:

- Object to be handed over
- Human actor
- Human pose [standing, sitting, laying down]
- Human behaviour before grasp [reaches out, does not reach out]
- Human behaviour during grasp [grasps object, does not grasp object]
- Human behaviour after grasp (up to 5 seconds) [keeps object in hand, lets object fall]

6.6.4 Communication with the Referee Box

For each trial:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the outcome of the execution (specified below)
- In case of a timeout, the referee box sends a stop message to the robot

The message defining the outcome of the execution should include the following information:

- human pose [standing, sitting, laying]
- human reached out for object [yes, no]
- object was grasped successfully [yes, no, undefined]
- object fell down after grasp [yes, no, undefined]

6.6.5 Procedures and Rules

The maximum time allowed for one trial is 30 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the outcome of the action. If 30 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

The human will place themselves at most 1 m in front of the robot.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.6.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Position of end-effector with respect to the robot base at the time of the hand-over

In addition to the recorded internal robot data, an external RGB camera will record each run.

6.6.7 Metrics for Evaluation

The performance of the robot is based on the following achievements for each execution:

1. Initiation of movement of arm towards human [Achievements: 1]
2. Detection of human pose [Achievements: 0.5]
3. Variation of end-effector pose based on human pose [Achievements: 0.5]
4. Detection of human (not) reaching out for object [Achievements: 0.5]
5. Object is released at most 5 seconds after human reaches out [Achievements: 0.5]
6. Detection of (un) successful grasp [Achievements: 1]
7. Detection of object (not) falling after grasp [Achievements: 1]

In addition the following subjective evaluations will also be recorded by the referees:

- natural motion of the arm
- smooth / intuitive handover of object to person

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops the object causing it to drop to the floor through no fault of the human (note: this is a subjective evaluation, and it will be up to the referees to decide if the robot was at fault)
- the robot does not release the object
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

6.7 Receive Object Functionality

6.7.1 Functionality Description

This functionality benchmark assesses the robot's capability of receiving objects from a person. The robot is placed in front of a person who is holding an object. The robot must receive the object from the person when it is offered. The robot is allowed to initiate a dialogue to facilitate the exchange. In case no dialogue is initiated, the person will hand over the object without prompting.

A single run of the benchmark consists of ten trials, each with different variations.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.7.2 Healthcare Relevance

Similar to the handover task, a robot in a healthcare setting is often required to receive objects from a human. The interaction must be as smooth and intuitive as possible, as would be expected from a robot interacting with persons with impairments.

6.7.3 Feature Variation

The configuration for a single execution consists of the assignment of the following independent variables:

- Object to be handed over
- Human actor
- Human pose [standing, sitting, laying down]
- Human behaviour before grasp [reaches out with object, does not reach out, drops the object]
- Human behaviour during and after grasp [releases the object, does not release the object]

6.7.4 Communication with the Referee Box

For each trial in the run:

- The robot waits for a start message from the referee box
- The robot sends a confirmation that it has received the start message
- The robot sends a message to the referee box with the outcome of the trial (specified below)
- In case of a timeout, the referee box sends a stop message to the robot

The message defining the outcome of the trial should include the following information:

- human pose [standing, sitting, laying]
- human reached out with object [yes, no]
- object fell down [yes, no]
- object was grasped successfully [yes, no, undefined]
- object was released [yes, no, undefined]

6.7.5 Procedures and Rules

The maximum time allowed for one trial is 30 seconds. The time is calculated from the moment the robot confirms the start message has been received, until the robot replies with the outcome of the action. If 30 seconds is exceeded, a timeout is recorded for that trial, and the robot must prepare for the next trial.

The human will place themselves at most 1 m in front of the robot.

6.7.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Position of end-effector with respect to the robot base at the time of the hand-over

In addition to the recorded internal robot data, an external RGB camera will record each run.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.7.7 Metrics for Evaluation

The performance of the robot is based on the following achievements for each trial:

1. Movement of arm towards human [Achievements: 1]
2. Detection of human pose [Achievements: 0.5]
3. Variation of end-effector pose based on human pose [Achievements: 0.5]
4. Detection of human (not) reaching out with object [Achievements: 1]
5. Detection of fallen object [Achievements: 0.5]
6. Grasping of object [Achievements: 1]
7. Detection of human (not) releasing object [Achievements: 1]

In addition the following subjective evaluations will also be recorded by the referees:

- natural motion of the arm
- smooth / intuitive receipt of the object from the person

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops the object causing it to drop to the floor through no fault of the human (note: this is a subjective evaluation, and it will be up to the referees to decide if the robot was at fault)
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

6.8 Speech Understanding Functionality

6.8.1 Functionality Description

This functionality benchmark assesses the robot's capability of understanding speech. Since the focus is on understanding (not just recognition) of speech, the benchmark will take the form of a question and answer session. The robot is placed in front of a human, who will ask the robot a set of questions. The robot must answer each question after it has been asked.

6.8.2 Healthcare Relevance

Speech is one of the methods of interaction between a robot and a human, particularly in a healthcare setting in a home or care facility. The speech could be slow, slurry or incoherent. The robot will hence be evaluated on its capability of understanding several speech patterns.

6.8.3 Benchmarking Data

The internally recorded data must include (at minimum):

- Audio stream of the robot

In addition to the recorded internal robot data, an external microphone will record each run.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.9 Opening Cupboard Functionality

6.9.1 Functionality Description

This functionality benchmark assesses the robot's capability of opening a cupboard or drawer. The robot is placed in front of a cupboard or drawer, such as a medicine cabinet or a kitchen drawer, and must open it.

6.9.2 Healthcare Relevance

A robot in a home or care facility might tasked to retrieve items such as medicine, cups, plates etc. from a cupboard or drawer.

6.9.3 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Force/Torque sensors (if available)
- Tactile sensors (if available)

In addition to the recorded internal robot data, an external RGB camera will record each run.

6.10 Pouring Functionality

6.10.1 Functionality Description

This functionality benchmark assesses the robot's capability of pouring a fluid from one container into another. The robot is placed in front of a table with a container (such as a cup). A container with fluid is placed in the robot's hand. The robot should pour the fluid into the container on the table until it is full or the source container is empty. To allow for safe execution of the benchmark, plastic beads will be used in place of an actual fluid.

6.10.2 Healthcare Relevance

A person, who has caring needs, might request the robot to bring a glass or water, or other drinks. For this task, the robot be able to pour liquids from one container to another.

6.10.3 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base

In addition to the recorded internal robot data, an external RGB camera will record each run.

6.11 Area Coverage Functionality

6.11.1 Functionality Description

This functionality benchmark assesses the robot's capability of exploring a given area, with the intention of disinfecting the area with a UV lamp.

Disinfection with a UV lamp is dependent on both the distance to the target surface and the duration for which the UV lamp is pointed to the surface. Hence, the robot must ensure that a sufficient time:distance ratio is achieved for all surfaces in the area.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

6.11.2 Healthcare Relevance

During the Covid-19 pandemic, several robotic systems have been built to automatically disinfect surfaces with the use of UV-lamps. Such robots typically move through a facility when humans are not present and disinfect visible surfaces.

7 Task Benchmarks

The task benchmarks aim to evaluate the performance of a robot in the execution of a full task. The full task includes several subsystems of the robot, which have been individually evaluated in the functional benchmarks. Hence, the focus is on the integration of the functionalities and the capability of the robot to account for failures in individual functionalities to successfully complete the task.

The following sections describe some task benchmarks and the evaluation procedure. The evaluation is typically in the form of achievements for having reached certain checkpoints in the task. Not all task benchmarks have been fully specified yet. Their metrics and procedures will be defined once their relevance and metrics for benchmarking have been clarified through the survey to stakeholders.

7.1 Assess Activity State Task

7.1.1 Task Description

This task benchmark assesses the robot's capability of integrating several FBMs to assess a person's activity state through both visual cues and a natural language dialogue. The robot must locate a particular person in a given location and initially visually assess their activity state. The robot must then approach the person and initiate a natural language dialogue to verify their activity state. The functionalities required to complete this task include Human Recognition 6.2, Activity Recognition 6.3 and Speech Understanding 6.8. All variations of the individual functionalities will considered for the task benchmark as well.

7.1.2 Healthcare Relevance

In addition to visually monitoring the activity state of a person, this task requires the robot to confirm the assessed state by engaging in a dialogue. In addition to increasing the level of engagement between the robot and the human, this task is a more comprehensive way for the robot to evaluate the activity state of a human, instead of simply observing visually.

7.1.3 Communication with the Referee Box

- The robot waits for a start message from the referee box. This message contains the identity of the target person and their location.
- The robot sends a confirmation that it has received the start message
- The robot sends a feedback message to the referee box to indicate its progress, when it has:
  - located the person
  - visually assessed their activity state
  - completed the assessment via natural language dialogue with the person
- The robot sends a message indicating the completion of the benchmark

7.1.4 Procedures and Rules

For each trial, the referee selects a person, the location, and activity of the person. The configuration for a particular trial is fixed for all teams.

The maximum time allowed for one task execution is 5 minutes. The time is calculated from the moment the robot confirms the start message has been received, until the robot indicates the end of the benchmark.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

If 5 minutes is exceeded, a timeout is recorded for that execution, and the robot must prepare for the next execution.

7.1.5 Autonomy Level

Teams are allowed to choose any level of autonomy for this task, from fully autonomous to fully remote-controlled. The teams must specify the autonomy level beforehand for each functionality. For this task benchmark, autonomy levels could include:

- Navigation: [full remote controlled, remote waypoint specification, fully autonomous]
- Visual Activity Recognition: [remotely assessed by team member, autonomous]
- Natural Language Dialogue: [remote conversation via microphone and speaker, autonomous]

7.1.6 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Odometry and global pose of the robot
- Output of human or face detection
- Output of activity recognition
- Audio used for recognizing speech
- Recognized speech
- Transcript of spoken text

In addition to the recorded internal robot data, an external RGB camera will record each run.
The data recorded during the runs will be used in the Cascade Evaluation campaigns to evaluate:

- human recognition
- activity recognition
- speech understanding

7.1.7 Scoring and Ranking

The performance of the robot is based on the following achievements for each execution:

1. Successful navigation to the person [Achievements: 1]
2. Successful identification of the person [Achievements: 1]
3. Successful visual recognition of the activity [Achievements: 1]
4. Initiation of dialogue with the person [Achievements: 1]
5. Verification of activity state through dialogue [Achievements: 1]

In case multiple runs are executed during a competition, the sum of achievements for all runs is calculated.
Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

7.2 Item Delivery Task

7.2.1 Task Description

This task benchmark assesses the robot's capability of integrating several FBMs to safely deliver a healthcare related item to a human. The robot must locate and grasp a specified item, transport the item to a human, hand over the item and verify that the item has been received. The functionalities required to complete this task include Object Detection 6.1, Task-oriented Grasping 6.5 and Handover 6.6. All variations of the individual functionalities will considered for the task benchmark as well.

7.2.2 Healthcare Relevance

An assistive robot can aid a person by fetching items from around the living area, which would be particularly helpful for persons with physical impairments.

7.2.3 Communication with the Referee Box

- The robot waits for a start message from the referee box. This message contains the required item and its location, and the location of the human
- The robot sends a confirmation that it has received the start message
- The robot sends a feedback message to the referee to indicate its progress, when it has:
  - located the item
  - grasped the item
  - reached the human
  - handed over the item

  The feedback messages will be identical to the ones specified in the individual functionality benchmarks, if applicable. For example, the feedback for handing over the item should include the human pose, whether the human reached out for the item, whether the item was successfully grasped, and whether the item fell down after the grasp.
- The robot sends a message indicating the completion of the benchmark

7.2.4 Procedures and Rules

For each trial, the referee selects an item, the locations, pose of the human and intended human behaviour. The configuration for a particular trial is fixed for all teams.

The maximum time allowed for one task execution is 5 minutes. The time is calculated from the moment the robot confirms the start message has been received, until the robot indicates the end of the benchmark. If 5 minutes is exceeded, a timeout is recorded for that execution, and the robot must prepare for the next execution.

7.2.5 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Output of object detection (if any)
- Output of human or face detection (if any)
- Position of end-effector with respect to the robot base at the time of the hand-over

In addition to the recorded internal robot data, an external RGB camera will record each run.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

7.2.6 Scoring and Ranking

The performance of the robot is based on the following achievements for each execution:

1. Successful navigation to item location [Achievements: 1]
2. Successful detection of item [Achievements: 1]
3. Successful pick of item [Achievements: 1]
4. Successful navigation to human location [Achievements: 1]
5. Successful detection of human [Achievements: 1]
6. Initiation of movement of arm towards human [Achievements: 1]
7. Detection of human pose [Achievements: 0.5]
8. Variation of end-effector pose based on human pose [Achievements: 0.5]
9. Detection of human (not) reaching out for item [Achievements: 0.5]
10. Object is released at most 5 seconds after human reaches out [Achievements: 0.5]
11. Detection of (un) successful grasp [Achievements: 1]
12. Detection of item (not) falling after grasp [Achievements: 1]

In case multiple runs are executed during a competition, the sum of achievements for all runs is calculated. In addition the following subjective evaluations will also be recorded by the referees:

- natural motion of the arm
- smooth / intuitive handover of object to person

Penalties

- the robot collides with environment or human in an uncontrolled manner
- the robot drops the object causing it to drop to the floor through no fault of the human (note: this is a subjective evaluation, and it will be up to the referees to decide if the robot was at fault)
- the robot does not release the object to the human
- the robot stops responding

Disqualifying Behaviours

- the robot damages the environment
- the robot damages the object

7.3 Area Coverage Task

7.3.1 Task Description

This task benchmark assesses the robot's capability of exploring an area while avoiding humans. The robot will mimic one with an active UV lamp, and disable the UV lamp when it encounters humans. The indication of an active or inactive UV lamp can be mocked up with LEDs or some other form of visual or auditory output. This task requires the combination of the FBM Area Coverage 6.11 and is related to the Human Recognition FBM 6.2, though it is not necessary to identify persons in this task.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

7.3.2 Healthcare Relevance

The relevance is identical to the functional benchmark (Section 6.11, with the additional requirement that the robot must be able to disable the UV lamp when in the vicinity of humans.

7.3.3 Benchmarking Data

The internally recorded data must include (at minimum):

- RGB camera stream of the robot
- Proprioceptive sensor data from the robot's manipulator and base
- Output of human or face detection (if any)
- Map of environent

In addition to the recorded internal robot data, an external RGB camera will record each run.

7.4 Prepare Drink Task

7.4.1 Task Description

For this task benchmark, the robot must locate a filled container (such as a water bottle), and an empty cup which will be placed at a given location (such as the kitchen counter top). The robot must then grasp the container and pour its contents into the cup, stopping when the cup is nearly full or when the whole content has been poured.

This TBM requires the functionalities Object Detection 6.1, Task-oriented Grasping 6.5 and Pouring 6.10.

7.4.2 Healthcare Relevance

A person, who has caring needs, might request the robot to bring a glass or water, or other drinks.

7.5 Receive and Transport Drink Task

7.5.1 Task Description

For this task benchmark, the robot must recognize a particular person at a given location, receive a partially-full cup of water from them, transport it to the kitchen and place it on the counter. This TBM requires the functionalities Human Recognition 6.2 and Receive Object 6.7, in addition to being able to navigate with an object, and being able to place an object on a surface.

7.5.2 Healthcare Relevance

This is another typical task that an assistive robot might be tasked to perform; namely restoring items to their original location. In this particular task, the robot must transport a half-full cup safely, after having successfully received it from a person.

8 Cascade Evaluation Campaign

While the field evaluation campaigns, consisting of FBMs and TBMs, require participation with a physical robot, the cascade evaluation campaigns are meant to evaluate the performance of teams on datasets collected during the field evaluation campaigns. The datasets will be made available to any team that wishes to participate, including those that did not compete in the field evaluation campaign. Submissions will be accepted online, and the evaluation will be based on the performance on the test set of the datasets.

The objective of this campaign is to evaluate algorithms independent of a physical robot, and their generalisability across different robot platforms and environments.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

The benchmarks in the cascade evaluation campaign are described in Table 2. The metrics will be identical to those used in the corresponding FBM. In some cases, such as for grasp verification, detection of (un)successful handover etc., the benchmark does not directly mirror the FBM. In these cases, the metric for evaluation is simply the accuracy of detection.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

References

[1] F. Amigoni, E. Bastianelli, J. Berghofer, A. Bonarini, G. Fontana, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci, *et al.*, "Competitions for benchmarking: Task and functionality scoring complete performance assessment," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 53–61, 2015.

[2] J. C. Scholtz, "Human-robot interactions: Creating synergistic cyber forces," in *Multi-Robot Systems: From Swarms to Intelligent Automata*. Springer, 2002, pp. 177–184.

[3] S. N. Woods, M. L. Walters, K. L. Koay, and K. Dautenhahn, "Methodological issues in hri: A comparison of live and video-based methods in robot to human approach direction trials," in *ROMAN 2006-the 15th IEEE international symposium on robot and human interactive communication*. IEEE, 2006, pp. 51–58.

[4] A. Weiss, R. Bernhaupt, M. Lankes, and M. Tscheligi, "The usus evaluation framework for human-robot interaction," in *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction*, vol. 4, no. 1, 2009, pp. 11–26.

[5] J. S. Dumas, J. S. Dumas, and J. Redish, *A practical guide to usability testing*. Intellect books, 1999.

[6] N. Savela, T. Turja, and A. Oksanen, "Social acceptance of robots in different occupational fields: A systematic literature review," *International Journal of Social Robotics*, vol. 10, no. 4, pp. 493–502, 2018.

[7] A. Van Maris, N. Zook, P. Caleb-Solly, M. Studley, A. Winfield, and S. Dogramadzi, "Designing ethical social robots—a longitudinal field study with older adults," *Frontiers in Robotics and AI*, vol. 7, 2020.

[8] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510–526.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

A Evaluation Sheets

These evaluation sheets have been adapted from those used in the RoCKIn competitions[7].

[7] https://github.com/rockin-robot-challenge/evaluation_sheets

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Object Detection Functionality

Team name: _____

Referee I: _____, **Referee II:** _____

Date and time: _____

Execution	Object	IoU	TP	FP	FN	TN	Timeout
1.							
2.							
3.							
4.							
5.							
6.							
7.							
8.							
9.							
10.							
Sum.							

Total execution time [s] _____

Benchmarking data delivered appropriately: ☐ yes / ☐ no

Team leader signature: _____

Referee signature: _____

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Human Recognition Functionality

Team name: _____

Referee I: _____, **Referee II:** _____

Date and time: _____

Notes for Referees:

- The configuration (actor, face pose, etc.) for all ten executions must be selected and fixed for all teams
- If multiple runs are executed, the configuration must be regenerated

Execution	Actor	Human standing	Human sitting	Human laying	Face straight	Face right	Face left	Eye wear	Face mask	Head covering	TP	FP	FN
1.													
2.													
3.													
4.													
5.													
6.													
7.													
8.													
9.													
10.													

Total execution time [s] _____

F1-score _____

Benchmarking data delivered appropriately: ☐ yes / ☐ no

Team leader signature: _____

Referee signature: _____

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Activity Recognition Functionality

Team name: _____

Referee I: _____, **Referee II:** _____

Date and time: _____

Notes for Referees:

- The same actors must be used for all teams

Execution	Activity Verb	Activity Noun	TP Verb	TP Noun	Timeout
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					

True Positive Rate (verbs) _____

True Positive Rate (nouns) _____

Total execution time [s] _____

Benchmarking data delivered appropriately: ☐ yes / ☐ no

Team leader signature: _____

Referee signature: _____

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Task-oriented Grasping Functionality

Team name: _____

Referee I: _____, **Referee II:** _____

Date and time: _____

Notes for Referees:

- The objects, their poses and the succeeding tasks for all ten executions must be selected and fixed for all teams
- If multiple runs are executed, the sets of objects, poses and tasks must be regenerated

Execution	Object	Grasp Orientation	Grasp Location	Grasped	Correct orientation	Correct location	Orientation error	Timeout
1.								
2.								
3.								
4.								
5.								
6.								
7.								
8.								
9.								
10.								

Total execution time [s] _____

Average estimated orientation error [degrees] _____

Total achievements _____

Benchmarking data delivered appropriately: ☐ yes / ☐ no

Team leader signature: _____

Referee signature: _____

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Handover Functionality

Team name: _____

Referee I: _____, **Referee II:** _____

Date and time: _____

Notes for Referees:

- The configuration (object, actor, human pose, human actions during grasp) for all ten executions must be selected and fixed for all teams
- If multiple runs are executed, the configuration must be regenerated

	External configuration										Robot behaviour					
Execution	Object	Actor	Human standing	Human sitting	Human laying	Human reaches out	Human does not reach out	Human grasps correctly	Object falls during grasp	Object falls after grasp	Initiate arm motion towards human	Detects human not reaching out	Release at appropriate time	Detects unsuccessful object grasp	Detects object falling after grasp	Timeout
1.																
2.																
3.																
4.																
5.																
6.																
7.																
8.																
9.																
10.																

Total execution time [s] _____

Total achievements _____

Benchmarking data delivered appropriately: ☐ yes / ☐ no

Team leader signature: _____

Referee signature: _____

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Receive Object Functionality

Team name: _____

Referee I: _____, **Referee II:** _____

Date and time: _____

Notes for Referees:

- The configuration (object, actor, human pose, human actions during grasp) for all ten executions must be selected and fixed for all teams
- If multiple runs are executed, the configuration must be regenerated

	External configuration										Robot behaviour					
Execution	Object	Actor	Human standing	Human sitting	Human laying	Human reaches out	Human does not reach out	Human drops the object	Human releases the object	Human does not release the object	Detects human not reaching out	Detects object falling before grasp	Initiate arm motion towards human	Grasps at appropriate time	Detects human not releasing object	Timeout
1.																
2.																
3.																
4.																
5.																
6.																
7.																
8.																
9.																
10.																

Total execution time [s] _____

Total achievements _____

Benchmarking data delivered appropriately: ☐ yes / ☐ no

Team leader signature: _____

Referee signature: _____

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

B Survey on Robotics in Healthcare

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252

Survey Information.

Survey on Robotics in Healthcare

Before you decide to participate in this survey, it is important for you to understand why this research is being done and what it will involve. Please take time to read the following information carefully. Take time to decide whether or not you wish to take part.

What is the purpose of this study?
This questionnaire is meant to assess the relevance of tasks performed by a robot in a healthcare setting. The target group for this questionnaire includes healthcare workers such as nurses, caregivers and family, in addition to people in need of assistance. The tasks performed by the robot are intended to assist older adults with physical, sensory and cognitive impairments.

The results of this questionnaire will help guide the definition of tasks in robotics competitions for healthcare. Therefore the primary aims are to:

- propose tasks based on the observed capabilities of the robot

- assess the relevance of certain tasks, and

- identify safety-critical parts of the tasks, and proposing expected robot behaviour in safety-critical situations

Written material associated with the project (including anonymized direct quotations from participants) may be used in scientific publication in specialist scientific proceedings and journals.

Do you have to take part?
It is up to you to decide whether or not to take part. If you do decide to take part, you will be asked to sign a consent form. If you decide to take part, you are still free to withdraw at any time, without giving a reason. A decision to withdraw at any time, or a decision not to take part, will not have any consequences.

What will happen to me if I take part?
The study will involve the completion of a brief (15-20 minute) online survey. This will include viewing a 5 minute video of a robot performing simple domestic tasks.

Will my participation in the study be kept confidential?
You will **not** be asked for any personally identifiable data such as your name, email or address, or date of birth. As such, all information that you provide in the survey will be anonymous. Any information that is reported or published from this study will therefore not contain information that would reveal your identity. You will be asked for basic demographic information such as age range, gender, occupation etc. Once you have submitted your responses to the survey, it will not be possible to delete them.

Who is organizing or sponsoring the research?
The research is organized by the EU H2020 project METRICS (Metrological Evaluation and Testing of Robots in International CompetitionS, funded by the EU grant agreement H2020-EU.2.1.1-#871252, https://cordis.europa.eu/project/id/871252). The universities involved in this study, who will have access to the data are: Hochschule Bonn Rhein Sieg (Germany), Heriot-Watt University (UK), University of the West of England (UK) and Università degli Studi di Firenze (Italy).

Who has approved this study?
This study has been reviewed and received ethical approval from XXXX Research Ethics committee. You may have a copy of this approval if you request it.

Consent Form

- - I confirm that I have read and understand the information sheet for the above study
- - I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, and without consequences
- - I understand that any of the information collected may be used in the dissemination of results of the project and will remain anonymous
- - I agree that information obtained by the researchers during the study may be analyzed for research purposes, and that anonymized text snippets (quotes) of what I write in my responses may be used in future research publications and to inform the design of robotics competitions. Such analysis is conducted only by members of the research team and collaborators in the present study.

I have read the above and agree to take part in this online study

Profile and Demographics

Q1.
The following questions ask a few brief demographic questions to help us understand your responses. We are interested in knowing more about your work and/or any care and support needs you may have. Please provide as much or as little information as you are comfortable with.

Please select all that apply

- [ ] I am a member of the management staff within a health or social care organisation
- [ ] I am a care worker within a health or social care organisation
- [ ] I am a resident within a health or social care facility
- [ ] I am living independently but I have care and support needs
- [ ] I am an informal carer (e.g. relative or friend) of someone with care and support needs
- [ ] I live independently and don't have any care needs
- [ ] I have another role that relates to the health and social care sector (please describe)

Q2. How would you describe your gender?

- ○ Male (including transgender men)
- ○ Female (including transgender women)
- ○ Prefer to self describe as (non-binary, gender-fluid, agender, please specify):
- ○ Prefer not to say

Q3. What is your age range?

- ○ 18-24
- ○ 25-34
- ○ 35-44
- ○ 45-54
- ○ 55-64
- ○ 65-74
- ○ 75-84
- ○ 85 and over
- ○ Prefer not to say

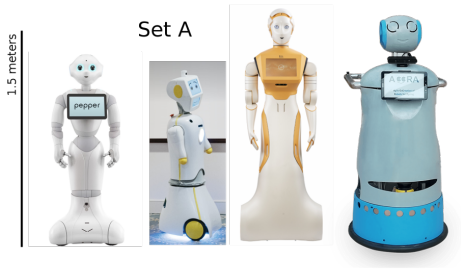
Q15. Please type your responses below. Leave blank if necessary.

What is your preferred spoken/written language?

Which country do you currently live/work in?

Opinions about emerging technologies

Q4.



We would love to know your opinions of robots such as the ones shown above, and of technology in general.

To what extent do you agree with the following statements?

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree
I would feel relaxed with such a robot in my home/care facility	○	○	○	○	○	○	○
I would be worried about robots like these moving their arms around	○	○	○	○	○	○	○
I would find it unnatural talking to these robots in the same way that I speak to people	○	○	○	○	○	○	○
I would feel scared around these robots	○	○	○	○	○	○	○
I can easily learn how to use such a robot	○	○	○	○	○	○	○
I feel the necessity for robots in my daily life	○	○	○	○	○	○	○
I worry about the robot breaking down	○	○	○	○	○	○	○
I expect my family or friends to help me when I use a robot	○	○	○	○	○	○	○
Robots can be used by remote control	○	○	○	○	○	○	○
I worry that robots are not suitable for the layout of my room / care facility	○	○	○	○	○	○	○
I am excited about technology such as robot helpers and other artificial intelligent systems	○	○	○	○	○	○	○
I am comfortable using technology such as computers,	○	○	○	○	○	○	○

Task usefulness

Q5.

Shown above are some examples of robotic platforms that might be used in a healthcare setting. The robot platforms typically have wheels to move around, an arm to grasp items, and sensors such as cameras, force sensors, distance sensors etc. The robots may also be equipped with microphones and speakers to enable speech recognition and generation.

How useful would it be for a robot to perform the following tasks to assist older adults with physical, sensory and cognitive impairments? Include other tasks which might be useful for an assistive robot to perform based on the robot platforms and their capabilities shown above.

	Extremely useful	Very useful	Moderately useful	Slightly useful	Not at all useful
Fetching the right medicine from a cupboard	○	○	○	○	○
Assisting the person with medicine intake	○	○	○	○	○
Preparing, transporting and serving a drink	○	○	○	○	○
Assessing a person's activity state	○	○	○	○	○
Leading a natural language dialogue	○	○	○	○	○
Navigate around the living area	○	○	○	○	○
Find misplaced items in the home	○	○	○	○	○
Give and receive items to/from a person	○	○	○	○	○
Recognize specific persons	○	○	○	○	○
Recognize gestures	○	○	○	○	○
Recognize safety-critical situations	○	○	○	○	○
Be able to tell a person when it is unable to complete a task	○	○	○	○	○
Other tasks (enter multiple if necessary)					
	○	○	○	○	○
Other tasks (enter multiple if necessary)					
	○	○	○	○	○
Other tasks (enter multiple if necessary)					
	○	○	○	○	○
Other tasks (enter multiple if necessary)					
	○	○	○	○	○
Other tasks (enter multiple if necessary)					
	○	○	○	○	○

Interaction methods and how to perform tasks

Q6. For each of the types of impairments below, select the corresponding interaction method(s) that would be most appropriate for a person with such an impairment to interact with a robot.

	Speech	Interactive screen on the robot	Mobile or web application	Gestures
Cognitive impairment	☐	☐	☐	☐
Visual impairment	☐	☐	☐	☐
Auditive impairment	☐	☐	☐	☐
Physical impairment	☐	☐	☐	☐
Other (please specify)	☐	☐	☐	☐
Other (please specify)	☐	☐	☐	☐
Other (please specify)	☐	☐	☐	☐
Other (please specify)	☐	☐	☐	☐

Q7. Please list activities and gestures of a person that the robot should be able to recognize. Examples of activities include laying down, falling down, cooking, etc, and gestures could be waving, pointing, beckoning, etc.

Q8. For each of the following levels of interaction with the robot, indicate your willingness to interact with or operate such a robot.

	Extremely comfortable	Somewhat comfortable	Neither comfortable nor uncomfortable	Somewhat uncomfortable	Extremely uncomfortable
Fully automated: the robot performs tasks by itself without interaction with the person (e.g. cleaning robot)	○	○	○	○	○
Some instruction and interaction: the robot performs tasks by itself, but with instructions from the person (e.g. fetching an item when requested)	○	○	○	○	○
Moderate amount of instruction and interaction: the robot performs tasks by itself but with detailed instructions from the person, including dialogue to establish specific instructions (e.g. fetching a specific item from a specific location)	○	○	○	○	○
Lots of instruction and interaction: the robot performs tasks by itself, but constantly interacts with the person to ensure the task is being performed correctly	○	○	○	○	○
Fully controlled by a human operator: the robot is fully controlled by a human (e.g. telepresence robot)	○	○	○	○	○

Q9. List possible care-related tasks that would be useful for the robot to perform, and select the corresponding level of interaction that would be most appropriate for you. You can enter multiple tasks per choice. The types of tasks can be anything that would assist a person who has caring needs.

	No supervision	Little supervision	Moderate supervision	Constant supervision	c
1		☐	☐	☐	☐
2		☐	☐	☐	☐
3		☐	☐	☐	☐
4		☐	☐	☐	☐
5		☐	☐	☐	☐

Video questions

Q10.
The video below shows a robot delivering an item to a person after some initial dialogue. Please watch the video and answer the following questions.

Note: Once you watch the video, you can click back in your browser to continue with the survey.

https://www.youtube.com/watch?v=jWbw6IUMg7U

What are some common objects that a person might ask a robot to bring?

Q11. What safety-critical events could occur during the execution of this task?

Q12. What should the robot do if:

	Inform the person and ask for further instructions	Continue looking for the item	Try to pick up the item again	Other
it cannot find the requested item	○	○	○	○
it drops the object during pickup	○	○	○	○
it drops the item while handing over to the person	○	○	○	○

Q13. How useful are the following skills of the robot to perform this task?

	Extremely useful	Very useful	Moderately useful	Slightly useful	Not at all useful
Adapting the handover of the item based on whether the person is sitting standing or laying	○	○	○	○	○
Detecting when the person is taking the item from the robot's hand	○	○	○	○	○
Detecting if the item falls to the ground during the handover	○	○	○	○	○