Trust in Robot Benchmarking and Benchmarking for Trustworthy Robots

Santosh Thoduka, Deebul Nair, Praminda Caleb-Solly, Mauro Dragone, Filippo Cavallo, and Nico Hochgeschwender

Abstract Trustworthy evaluation of robots is necessary for them to be deployed and accepted in society. Scientific benchmarking competitions provide a way to evaluate robots outside of lab conditions. We propose a progressive and iterative benchmarking process through competitions, which incorporates an objective dataset-based evaluation, evaluation on a remote robot, and field evaluations for individual robot functionalities and complete tasks, in a cyclical process similar to the machine learning lifecycle, with a view to achieving trustworthy evaluation. The inclusion of out-of-distribution data, failure scenarios and user studies as part of the benchmarking process addresses the necessity to evaluate robot systems on non-functional qualities such as fault tolerance, adaptability, social acceptance, in addition to their functional abilities to improve trustworthiness.

The original version of this chapter was previously published non-open access. A correction to this chapter is available at https://doi.org/10.1007/978-3-031-55817-7_11

S. Thoduka (✉) · D. Nair
Institute for AI and Autonomous Systems, Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Germany
e-mail: santosh.thoduka@h-brs.de

D. Nair
e-mail: deebul.nair@h-brs.de

P. Caleb-Solly
University of Nottingham, Nottingham, United Kingdom
e-mail: praminda.caleb-solly@nottingham.ac.uk

M. Dragone
Edinburgh Centre for Robotics, Heriot-Watt University Edinburgh, Scotland, UK
e-mail: m.dragone@hw.ac.uk

F. Cavallo
Department of Industrial Engineering, University of Florence, Florence, Italy
e-mail: filippo.cavallo@unifi.it

N. Hochgeschwender
Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany
e-mail: nico.hochgeschwender@uni-bremen.de

M. I. Aldinhas Ferreira (ed.), *Producing Artificial Intelligent Systems*,
Studies in Computational Intelligence 1150,
https://doi.org/10.1007/978-3-031-55817-7_3

Keywords Robot benchmarking · Trustworthiness

1 Introduction

With robots being introduced in society in the form of domestic service robots, agricultural robots, delivery robots, self-driving cars etc., it is essential that they are trusted by the people who use and interact with them. Rigorous evaluation is a crucial aspect of encouraging responsible development of these robots, and increasing the trust that people have in them [19]. Malle et al. [24] reviewed multiple measures of trust, including the performance of the system, in terms of competence and reliability, as well as moral trust, in terms of more qualitative measures of integrity, sincerity and benevolence. Evaluation typically starts with experiments and testing in lab conditions, and, when a robot is closer to deployment, can progress to field testing in the target environment with formal test procedures [28]. The complexity of today's robot systems and the environments in which they are required to operate, makes evaluation challenging. Subsystems need to be evaluated individually, and the overall robot system should be evaluated considering the dependencies between subsystems.

Scientific benchmarking through competitions, such as RoCKIn [1] and the European Robotics League (ERL)[1] [4], is one of the ways in which robots can be evaluated scientifically outside of a lab environment. However these competitions are not so focused on moral trust. The main goal of RoCKIn was to conduct robotics competitions in a manner similar to scientific experiments, such that robots are evaluated in controlled conditions (i.e. certified test beds), and the results are repeatable and reproducible. It introduced the concepts of functional and task benchmarks (FBMs and TBMs), which differentiate between evaluating a standalone functionality of a robot, and evaluating a complete task which might require the integration of multiple functionalities. FBMs allow the evaluation of functionalities even in the absence of a fully developed system, with a focus on repeatability under varying conditions. TBMs require robots to consider dependencies between functionalities, and the focus is on the completion of *achievements*, which are checkpoints within the task, indicating progress towards the final objective, such as successfully reaching a location relevant for the task.

Benchmarking for trustworthy robots should evaluate both functional and non-functional aspects; according to the acceptance model for robots and autonomous systems proposed by [20], a robot is required to be both "*worthy*" (functionally capable) and "*trusty*" (reliable, safe, secure, ethical, etc.). Functional aspects are typically evaluated using quantitative metrics, whereas the evaluation of non-functional aspects is not as straightforward, since quantitative metrics cannot be easily applied to fault tolerance, flexibility, etc. Competitions such as RoboCup,[2] RoCKIn, and the ERL focus heavily on the *functional* aspects of robot evaluation, but ignore or only

[1] https://eu-robotics.net/eurobotics/activities/european-robotics-league/.

[2] https://robocup.org/.

implicitly evaluate *non-functional* aspects such as transparency, reliability, safety, fault tolerance and flexibility. An analysis of rulebooks from competitions in [25] discusses criteria used for evaluating robots. Functional criteria assess the correctness of the robot's behaviour, and other quality aspects not associated with the correctness of the behaviour are considered non-functional. Success rate is provided as an example of a non-functional criteria which indirectly measures robustness. Some competitions introduce penalties for incomplete or incorrect behaviours (such as collisions), which might nevertheless result in a positive evaluation on functional criteria. These penalties indirectly evaluate characteristics such as safety and fault tolerance, but there is no explicit evaluation of these aspects.

In the field of computer vision and machine learning, benchmarking through competitions is also an established method; some examples include the ImageNet challenge [31] and the EPIC-KITCHENS challenge [11]. These challenges are based on datasets, which typically have a fixed (and sometimes hidden) evaluation dataset, against which all submissions are evaluated. In robotics, it is not possible to fully evaluate a system purely based on fixed datasets. Robots' physical embodiment, their ability to act on their environment and the dynamic feedback loop of autonomous sensing, perception, cognition and action are hard to capture through dataset-based evaluations. On the other hand, machine learning models are ubiquitous in several robotics capabilities such as scene perception, human-robot interaction, and navigation. These models can thus be partially evaluated through dataset-based competitions, with the caveat that they must *also* be evaluated on physical robots in real scenarios.

The different means by which robots can be benchmarked have their advantages and disadvantages. Participation in benchmarking competitions and field tests in the target environment can be expensive and time-consuming. On the other hand, such evaluations are necessary for subjecting the robot to real-world conditions, which can help identify corner cases and safety risks which might appear during deployment [28]. A simulation or dataset-based benchmark is comparatively less expensive, but the results may be trusted less than a field test. The difference in difficulty of dataset-based benchmarks and field benchmarks can enable robots at different stages of development to be evaluated, and would support their development.

We are therefore interested in how to structure benchmarking competitions for robots at different stages of development, such that the results of the benchmarks are trusted, with the goal of developing robots that are trustworthy. One of the key challenges we address in our work is to incorporate non-functional criteria in the benchmarks. As mentioned earlier, these are only implicitly evaluated in existing benchmarking competitions, although they are crucial to increasing the trust in benchmarking results, and for the development of trustworthy robots. The cost of field tests needs to be considered; field tests are necessary for trustworthy benchmarking, but alternatives could be considered for different stages of development and technology readiness levels of robots. For robots in human-centered environments, their social acceptance is also an important aspect; therefore robot benchmarks should consider the perception of robots and their behaviour by the people interacting with them.

We propose a progressive and iterative robot benchmarking procedure which builds on the framework developed in RoCKIn. The evaluation of a robot starts with benchmarking functionalities on datasets or simulation and progresses to benchmarking complete tasks in the field. By developing datasets and benchmark protocols to include failures, out-of-distribution, and unexpected situations as independent variables, we aim to explicitly evaluate various non-functional aspects such as robustness and fault tolerance into the evaluation procedure. We exemplify this through describing the staged and iterative benchmarking process we have defined in the context of the HEART-MET[3] competition.

2 Progressive and Iterative Robot Benchmarking

The RoCKIn framework decomposes tasks into functionalities and therefore allows evaluation at different levels of granularity. Both benchmarks types are evaluated with robots at a physical test-bed, and thus require deployment on a robot and interfacing with its sensors and actuators. In order to further focus on a functionality and to lower expense and deployment requirements, we propose two additional formats for benchmarking, i.e.

(i) dataset-based evaluation and
(ii) evaluation on a remote robot,

which have been developed in the METRICS project [3].[4] METRICS builds on previous projects on robot benchmarking through competitions (such as RoCKIn), and focuses on developing an evaluation framework, based on metrological principles, for robots in four domains–healthcare, inspection and maintenance [13, 29], agri-food [6], and agile manufacturing. METRICS introduces *cascade campaigns*, which are online, dataset-based competitions, in addition to the already established *field campaigns* (consisting of TBMs and FBMs). Field campaigns enable the evaluation of robots outside of lab conditions, yet allow for control of the experimental conditions for an objective comparison between competing robots. Cascade campaigns help participants to improve their algorithms on datasets, which can be deployed on the robot for the next field campaign. They lower the barrier to entry to the competitions, but also enable a focused evaluation on a single component in a manner similar to dataset-based computer vision benchmarks.

Given the differing levels of difficulty and costs between field and cascade campaigns, we propose four levels of evaluation for robot functionalities and tasks, consisting of dataset-based evaluation, FBM on a remote robot, and field FBMs and TBMs, as illustrated in Fig. 1. Each of the levels introduces additional variables that challenge the robustness of the functionalities and tasks being evaluated. The

[3] Healthcare Robotics Technologies–Metrified (HEART-MET) is a competition for assistive robots in the healthcare domain.

[4] https://metricsproject.eu/.

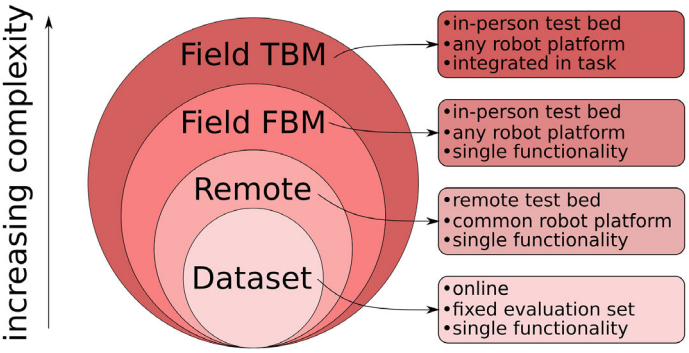

Fig. 1 A robot can be evaluated at different levels of complexity, from evaluation of a functionality on a dataset, to evaluation of a complete task which includes multiple functionalities.

progressive complexity of testing introduced in each of the levels can be compared to test-driven development (TDD), and unit, integration and system tests in software testing [5]. Dataset-based evaluation can be compared to a unit test, in which a single component is tested on a fixed set of input combinations. Remote and field FBMs require integration of the component on a physical robot, akin to integration tests, whereas TBMs evaluate the complete robot system on a given task, similar to system tests. Similar to TDD, the proposed evaluation process is iterative. The results from each evaluation level can inform both the development of new testing criteria, and the development of the functionalities and tasks. Thus a field TBM can be followed by a dataset-based evaluation with new data, and the following remote and field evaluations can include new testing conditions based on results from the previous evaluations, so that the test coverage of the robot is increased incrementally. A similar concept is explored in [27], where complexity in robot behaviour, obstacles, and terrain for field robots are gradually increased during test iterations to ensure an efficient testing strategy and to determine the performance envelope of the robots for future tests. An experience report for testing service robots [28] discusses the iterative nature of testing; early system integration tests for a disinfection robot were followed by months of component-level testing of person detection since it was critical for the safe operation of the robot.

The iterative process is similar to formative assessment which is performed during the learning or development phase, and thus can be geared towards robot functionalities that are still being developed. In comparison to other robotics competitions, which might expect a fully developed robot system, the earlier stages of the benchmarking process proposed here the lower barrier to entry for participating teams. Although a linear progression through the levels is recommended, it is not strictly necessary that all levels are used to benchmark a robot. Progressively evaluating the robot at each level will, however, provide insights about which components or system-level features require improvement. Considering the cost involved in field benchmarks, it

Table 1 HEART-MET cascade and field campaigns

Date	Type	FBMs	TBMs
April–June 2021	Dataset	– Gesture Recognition – Activity Recognition	
August–October 2022	Dataset	– Gesture Recognition – Activity Recognition	
October 2022	Remote FBM	– Gesture Recognition – Activity Recognition	
October 2022	Field FBM & Field TBM	– Object Detection – Person Detection – Gesture Recognition – Activity Recognition	Assess Activity State
October–November 2022	Dataset	– Activity Recognition	
May 2023	Field TBM		Physically Assistive Robot Challenge

is highly recommended to perform a dataset-based or remote evaluation prior to the field FBMs and TBMs.

This process was implemented in the HEART-MET competition, where we have conducted competitions at all four levels, as seen in Table 1. As its precursors, such as the ERL, METRICS' HEART-MET is designed to evaluate robot systems for healthcare applications: robots are evaluated on functionalities such as localisation and mapping, object detection, gesture and activity recognition, object handover, task-oriented grasping and speech understanding and human-robot interaction, which are required to complete tasks such as assessing the activity state of their human users, interpreting their instructions, delivering items such as medicines to a person and preparing and transporting a drink. Dataset-based evaluations focused on video-based gesture and activity recognition, which were subsequently evaluated on a remote robot, and during field FBMs. The TBMs *Assess Activity State* and *Physically Assistive Robot Challenge* required the integration of FBMs such as gesture and activity recognition, person detection and object detection in order to complete the task.

2.1 Dataset-Based Evaluation

Cascade campaigns were introduced as a way to evaluate sub-components of a robotic system in a cost-effective manner through evaluation on benchmark datasets. For example, in the video-based gesture recognition benchmark in HEART-MET competition, participants are provided with labeled training and validation sets for the cascade campaign, and submit results on an unlabeled test set using the Codalab platform[5] (see Fig. 2). This functionality is required to complete tasks such as the

[5] https://codalab.lisn.upsaclay.fr/.

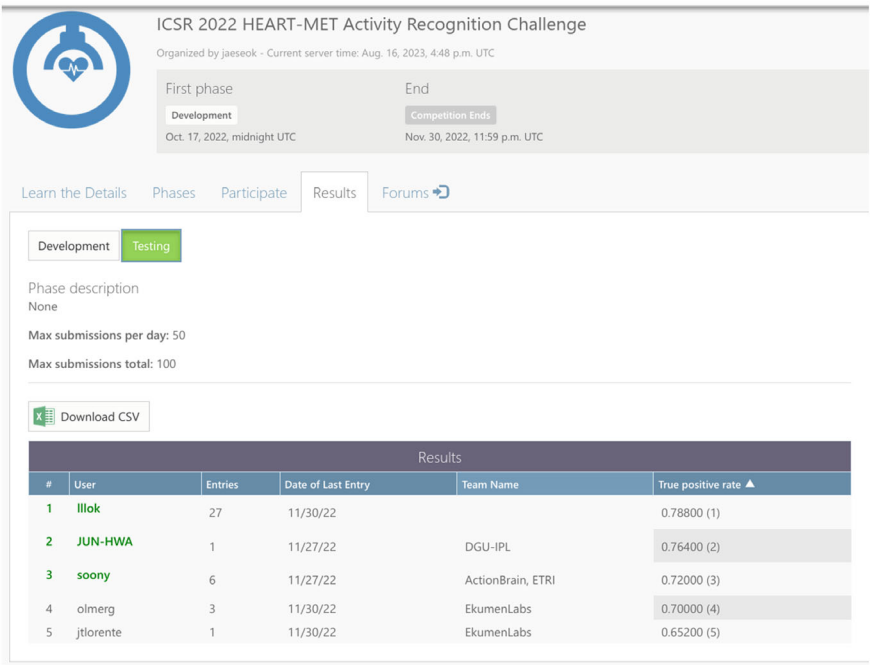

Fig. 2 A dataset-based competition for Activity Recognition hosted on the Codalab platform

Physically Assistive Robot Challenge for communicating with a person. By focusing the evaluation on a single functionality and on a fixed dataset, the cascade campaign is an opportunity for participants to develop and improve their algorithms without a physical robot involved, while still being able to compare their results to other participants. In Fig. 1, this represents the lowest level of evaluation complexity. Some functionalities, such as navigation, may not require datasets for evaluation. In such cases, a simulation-based evaluation can take the place of datasets, in which the evaluated scenarios can be fixed.

The quality of the dataset determines the extent to which the evaluation result is representative of the expected performance of the functionality when deployed in the real world. Trust in a dataset-based benchmark of a robot functionality is thus dependent on the quality of the dataset. An important characteristic of a good dataset is that the variations of independent variables in the dataset are representative of the real world. The dependent and independent variables should be clearly defined and the distribution of the variables in the dataset should be documented. For the validation set, out-of-distribution samples are also desirable in order to evaluate robustness. We discuss some of these characteristics further in Sect. 3.1, applied to datasets collected during the HEART-MET competitions.

2.2 Remote Evaluation

Deploying a functionality on a robot introduces additional variables, which might reduce performance. Examples include noisy sensors and actuators, new environments, the need to process sensor data online, considerations about the embodiment of the robot, etc. Evaluation on a remote robot provides the benefit of ensuring all methods are evaluated on a common platform, while keeping costs low for participants. The Real Robot Challenge 2022[6] is an example of such a competition, in which participants are provided with datasets for reinforcement learning and imitation learning for manipulation tasks, and submissions are evaluated remotely on a set of real robots. Robothon [35] and the remote format of the Robotic Grasping and Manipulation Competition (RGMC) by the National Institute of Standards and Technology [22] both take a different approach by providing a standard task board to participants for performing the tasks in their own labs. The submission includes video recordings of the robot performing the task, and a live demonstration through video conferencing.

In the HEART-MET competition, the Robotic Assisted Living Lab (RALT)[7] at Heriot-Watt University has developed a framework for enabling Docker-based submissions, which can be executed on a Toyota Human Support Robot (HSR) in their lab (see Fig. 3), which follows a similar approach to the Real Robot Challenge. As with datasets, for each benchmark, the dependent and independent variables are identified, and several trials are performed with different sets of dependent and independent variables. Compared to the dataset-based evaluation, remote evaluation increases the complexity and validity of evaluation, since the functionality is evaluated on a physical robot in a new environment (in case of remote robots), and the software must interface with the robot's sensors and actuators such that it can run in a time-constrained manner. However, unlike a dataset-based evaluation, it may not be possible to evaluate all variations of independent variables of the functionality due to time constraints. Special consideration should be given to independent variables that challenge the robustness of the functionality (such as extreme lighting conditions, noisy environments, etc.). A remote evaluation on a real robot may increase the trust in the benchmark result, given that the functionality has already been evaluated on a representative dataset beforehand. Inclusion of out-of-distribution and other anomalous conditions will enable evaluation of non-functional aspects as well.

[6] https://www.real-robot-challenge.com/.

[7] https://ralt.hw.ac.uk/.

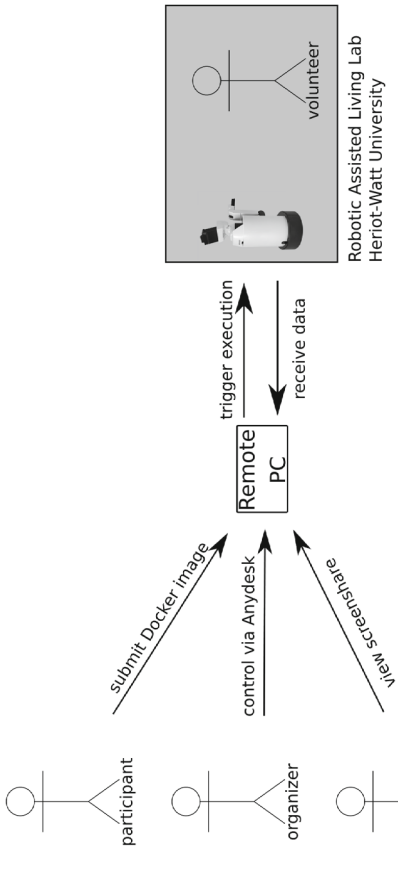

Fig. 3 Participants can submit Docker images to evaluate their algorithms remotely on a robot in the Robotic Assisted Living Lab.

2.3 Functional Benchmark

A functional benchmark is used to evaluate a standalone functionality of a robot at a certified test-bed [15, 33].[8] These benchmarks can be used to compare the performance of different robots, or to track the performance of a single robot over time on a single functionality. Some examples of the functionalities that are used in robots for their operations are: speech recognition, person detection, object detection, object manipulation, gesture recognition, navigation, etc. In addition to the metrics that directly measure success of the functionality, metrics like speed, accuracy, robustness and energy efficiency are also measured. An execution of the Object Detection FBM at the certified test-bed in the Cobot Maker Space at the University of Nottingham[9] can be seen in Fig. 4.

Since an FBM requires the evaluation of the robot at a certified test-bed, participants and their robots should be present at the test-bed. Thus the robots are all evaluated in the same environment, but participants may use any robot, which introduces differences in performance caused by the particular robot that is used. Similar to a remote evaluation, a functional benchmark may not be able to evaluate all variations of independent variables due to time or environmental constraints. For example, weather conditions might limit the types of lighting conditions in which an outdoor robot is evaluated. Nevertheless, functional benchmarks are an important tool for the development and evaluation of trustworthy robots, since they are one step closer to testing robots in their target environment.

Fig. 4 Object Detection FBM at the certified test-bed in the Cobot Maker Space at the University of Nottingham

[8] For example, ERL certified test-beds for domestic service robots can be found here: https://old.eu-robotics.net/robotics_league/consumer/certified-test-beds/certified-test-beds.html.

[9] https://cobotmakerspace.org/.

2.4 Task Benchmark

Task benchmarks are used to compare the performance of the robot in solving complete tasks. They have a close resemblance to system integration testing (SIT), in which the complete system, composed of many sub-components, is tested. TBMs help in identifying the integration challenges of the different components developed for solving the respective tasks. They also test the interoperability of the different functionalities, either on a single device or in distributed devices. Some examples of task benchmarks in METRICS include socially acceptable item delivery in the healthcare domain, (Fig. 5) pipeline area inspection or repetitive inspection in the inspection and maintenance domain, intra-row weeding in the agricultural domain, and collaborative assembly between human and robot in the agile manufacturing domain.

A semi-autonomous TBM can be introduced as an intermediate evaluation method between an FBM and TBM. The semi-autonomous mode provides the option for a human operator to perform one or more functionalities through tele-operation, although the high-level task, and most of the functionalities are planned and executed by the robot. The goal is to evaluate the integration of all functionalities in a task even when one or more of the functionalities have not been fully developed or evaluated yet.

Task benchmarks have the utmost impact in acceptance of the robot in operations. A degraded performance from a single functionality can cause a complete task failure; for example, if the object detection functionality does not detect a target object, the

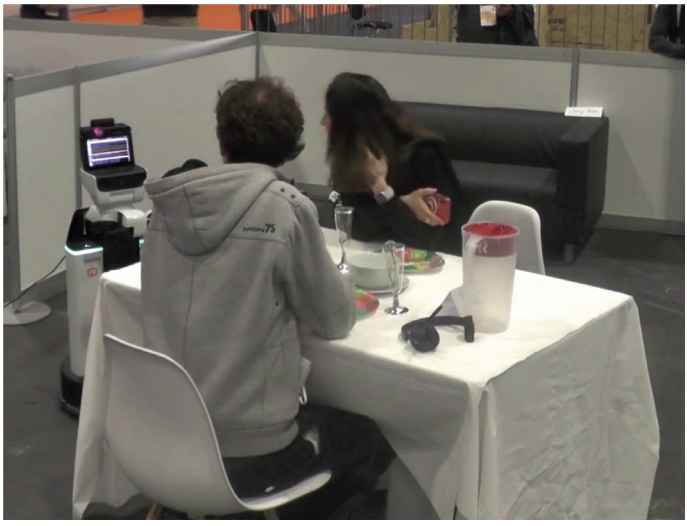

Fig. 5 Physically Assistive Robot Challenge TBM at the 2023 international Conference on Robotics and Automation

robot will not be able to proceed with the task of fetching it for a person. Therefore, a task benchmark additionally evaluates a robot's ability to recover from failures, and requires the robot to adapt to new situations that might be caused by unreliable functionalities.

Since task benchmarks constitute complete tasks in the target environment, they are also an opportunity for end-users to visually inspect and understand the expected behaviour of robots. They offer opportunities to evaluate aspects such as usability, social acceptance and trust from the point of view of the end-user. Feedback regarding usability, social acceptance and user experience can be obtained through questionnaires conducted during and after the TBM, based on the USUS framework [39]. As further described in Sect. 3.3, the results of the questionnaire can be incorporated into the overall score for a TBM.

2.5 Completing the Cycle

The development, evaluation and deployment phases in software development, robotics [34] and machine learning [2] are often iterative, in which the results of the evaluation and deployment phases are used for further development. The competition framework followed in METRICS is similarly iterative. The structure of cascade and field campaigns is closely related to the machine learning (ML) lifecycle [2], as illustrated in Fig. 6. The ML lifecycle starts with the collection of a dataset in the data management phase. The model learning and verification stages consist of the initial training and validation phase of a new model, which is performed during the cascade campaigns. The verified model is applied in the real world during the model deployment stage, which, in METRICS, is performed during the field campaigns. The data management phase receives feedback from each of the three remaining phases, which informs decisions about new data that should be collected to improve the model. Once the model is deployed, new conditions might cause the verified model to perform poorly, thereby restarting the cycle through the collection of additional data in the new environment. During the field campaigns in METRICS, data is recorded from the robots for the development of new datasets for the next cascade campaign.

The iterative process of evaluation ensures that results from previous benchmarks are taken into consideration for future development. The benchmarks themselves can be updated by the inclusion of new data in the datasets, and increasing focus on areas of poor performance. Robot functionalities and tasks can also be improved by teams through similar means. The iterative and progressive nature of the evaluation is thus suited for robots with technology readiness levels (TRL) 4 through to 7; namely for a component being validated in a lab environment (TRL 4) to a full scale demonstration of the system in the target environment (TRL 7) [12].

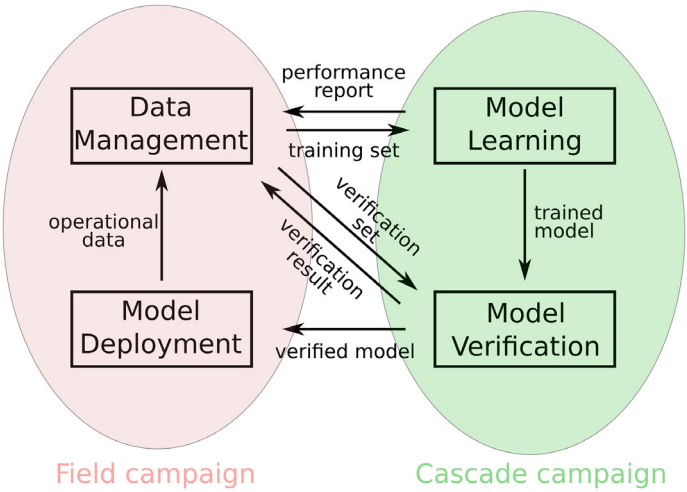

Fig. 6 Relation between the machine learning lifecycle (diagram based on [2]), and the field and cascade campaigns

3 Developing Benchmarks for Trustworthy Robots

In this section, we describe benchmark datasets and benchmark definitions for field campaigns developed in the context of the HEART-MET competition. We describe the characteristics of the datasets which encourage the development of algorithms and models which have desirable non-functional properties such as fault tolerance, adaptability, and robustness. Similarly, we also describe the reasoning for including non-nominal and failure scenarios in the benchmark definitions for the field campaigns, with the goal of evaluating robots' resilience to unexpected conditions. With the target of developing and benchmarking robots operating in human-centric environments, we also emphasize the importance of evaluating them based on social acceptance criteria.

3.1 Datasets

ML models on robots are particularly well suited to be evaluated at all four levels illustrated in Fig. 1 since the underlying machine learning models are developed using datasets, but need to be evaluated in the context of tasks once they are integrated into a robot component. Although dataset-based evaluations are at the lowest level of complexity of evaluation, they form the basis for evaluating robots in field campaigns; i.e. if a dataset-based evaluation does not properly evaluate a functionality, the impact of a field evaluation will be low.

Fig. 7 Sample frames from a video clip of (top) *Drinking water* in the Activity Recognition dataset and (bottom) *Thumbs down* in the Gesture Recognition dataset

Fig. 8 The person induces a failure during a handover by not letting go of the object

In the HEART-MET competition, we recorded three datasets which were used in the context of cascade and field campaigns:

(i) video-based activity recognition,
(ii) video-based gesture recognition, and
(iii) handover failure detection,

which are relevant for a healthcare robot that interacts with persons. Activity recognition (see Fig. 7) may be used by the robot to offer assistance for certain activities, or to respond to unexpected actions such as falling to the floor. Video-based gesture recognition may be used as a modality for human-robot communication (in addition to speech). The handover failure detection dataset (see Fig. 8) focuses on potential failures that could occur during human-robot interaction, so that the robot can respond to the failures appropriately.

A summary of the datasets is shown in Table 2.

Table 2 Summary of datasets collected for the HEART-MET competitions

Dataset	Locations	Subjects	Samples	Classes
Gesture Recognition	4	47	668	9
Activity Recognition	4	52	1178	19
Handover Failure Detection	1	17	591	4

3.1.1 Dataset Characteristics

The datasets were recorded for the purpose of evaluating algorithms which would eventually run on robots in real environments. Therefore, the environments in which they are recorded should reflect the types of environments that the robot will encounter during deployment. The datasets in Table 2 were recorded at four laboratories which were setup as simulated apartments with a living room, dining room, kitchen, etc. Several volunteers performed the actions, gestures and object handovers, and data was recorded using the robot's camera(s) and external cameras. The robot's proprioceptive sensors, such as joint positions and velocities and force-torque sensors, were recorded for the handover failure detection dataset since they were relevant for the task. Depth camera data, inertial sensors [14], and ambient sensors in the simulated apartment, such as switches on doors, motion sensors and power measurement of electrical devices [30], were recorded if available and considered relevant for the activities being recorded. We describe some of the characteristics of these datasets in the following paragraphs, and also some additional desirable characteristics of robotics datasets which would assist in the development of trustworthy robots.

Variability The variability in human subjects, location and time of dataset collection, hardware used, lighting conditions, and other task-relevant independent variables increase the likelihood that the dataset is representative of real-world conditions. Methods that use the datasets are hence more likely to be able to generalize to new environments and conditions. During the model verification and deployment phases, algorithms should be evaluated on new conditions not seen during training; hence the train, validation, and test datasets were split based on the subjects present in the dataset. The level of variation included in the dataset may be used to assist arguments that the dataset is more *complete* compared to other datasets, which in turn could increase one's confidence that methods trained on the dataset can generalize to new conditions. It is often not feasible to record datasets with high variability due to lack of resources or time; in such cases, a best-effort approach should be applied to capture as much variability as feasible.

Out-of-Distribution and Failure Samples We include out-of-distribution (OOD) samples in our datasets, such as noisy videos, blurred frames (see Fig. 9), activities and gestures which are not part of the target classes, and other unrelated videos. Since the outputs of learning models are used to perform actions, it is important that they also estimate the degree of confidence in their outputs, such that a higher-level planner

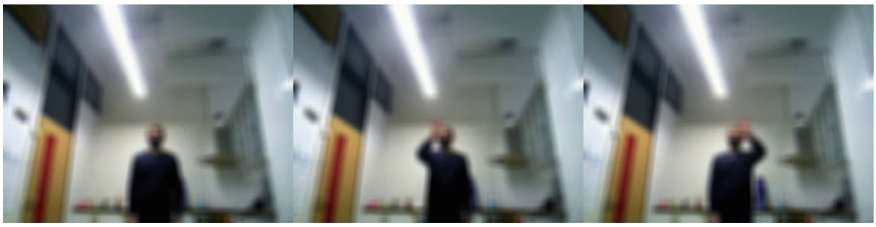

Fig. 9 An out-of-distribution sample in our Gesture Recognition dataset with blurred frames

can take it into account while making decisions [36]. Robots operate in open-world settings but typical learning datasets assume closed-world settings (i.e. only the categories present in the training dataset exist in the world). By including OOD samples in our datasets, we evaluate the methods' ability to estimate their own degree of confidence in a result, and to deal with unexpected inputs. The Fishyscapes benchmark [8] is an example of a dataset which intentionally introduces out-of-distribution, anomalous objects in the Cityscapes dataset [10]. The motivation for the handover failure detection dataset is similar–we want to evaluate the ability of robots to detect, diagnose and respond to failure conditions.

Both OOD and failure detection enable the evaluation of non-functional aspects such as fault tolerance, adaptability, and reliability. Including these instances in benchmark datasets increases one's confidence that the benchmark truly evaluates real world conditions which are expected to contain unseen scenarios and failures. This is especially relevant for robots, since they perform *actions* in the real world based on the results from learning models. Therefore it is important for them to estimate the uncertainty of the outputs of learning models, and to communicate uncertainties and failures to improve trust [7].

Other Desirable Characteristics Robots perceive the world continuously (temporal embodiment) and mobile robots are able to move around to perceive the world from different spatial locations (spatial embodiment). Temporal and spatial embodiment can be exploited to improve reliability and reduce uncertainty [18, 40], or overcome fault conditions such as occlusion. Datasets used for developing and evaluating robot functionalities should thus reflect the type of data available to the robot and include spatio-temporal data where feasible. For example, the instances of the CORe-50 dataset [23] for object detection and segmentation comprise of a sequence of image frames in which an object is moved around in front of the camera, capturing different views of the object at various distances and inducing partial occlusions, rather than static images of single objects.

As suggested by the iterative nature of our proposed evaluation framework, datasets should be continuously updated to reflect new conditions, objects, and behaviours. This would reflect the continually changing nature of the environments that the robots are expected to operate in, as well as their evolving application requirements. This also enables the development of approaches such as incremental learning, class incremental learning and active learning, which are challenges identified

in [36] for deep learning in robotics. Continual updates of the dataset improves the adaptability of the robot, and highlights the importance of logging provenance and metadata about the data, training and model updates, allowing some level of traceability for the outputs of learning models [21]. Another important aspect of dataset development is documentation–the Datasheets for Datasets [17] highlights the need to fully document the collection process and composition of datasets, and identify potential biases, or harmful outcomes of using the dataset.

3.2 Benchmark Definition for Field Campaigns

Just as with benchmark datasets, the definition of the benchmarking protocol for field campaigns plays a role in the level of trust assigned to the benchmark results. The conditions under which a benchmark is to be conducted should be defined, and eventually documented at the time of execution of the benchmark. As with datasets, the specified dependent and independent variables for FBMs and TBMs define the scope of the benchmarks concretely. In Table 3, examples of dependent and independent variables for two functional benchmarks in HEART-MET are shown. During the execution of a benchmark, the independent variables should be varied for each trial and recorded in the benchmark result. As described in [37], independent variables can also be failure modes–for example, the absence of the target object for an Object Detection benchmark, or adversarial behaviour from a person during a human-robot interaction task, which leads to a failure. When the exact conditions under which a certain component was evaluated is documented, it also improves the reproducibility of the results.

Increasing the number of variables increases the effort in conducting a benchmark due to combinatorial explosion. For example, one might want to evaluate an object handover for every pose of the person, and each with different behaviours of the person as described in Table 3. Therefore, although the benchmark definition lists possible independent variables, in practice one might choose a subset of combinations. Due to the iterative nature of the evaluation process, the combinations may be chosen based on previous performances, leaving open the option to test more combinations in future iterations.

3.3 Social Acceptance

For robots that interact with humans, their appearance, behaviour, and interactions with humans are central to their acceptance by humans (apart from their functional capabilities). A large majority of empirical research on social acceptance of robots is performed through user studies and surveys, but with relatively small sample sizes [32]. Low samples sizes lead to challenges such as choosing a representative group since the results are influenced by factors such as cultural differences [26],

Table 3 Dependent and independent variables defined for two sample FBMs

Benchmark	Dependent Variable	Independent variables
Person Detection	– Location of person in image – No detection of person in case no person is present	– Presence of a person – Distance of person to robot – Pose of person [standing, sitting, laying] – Pose of the person's face [straight, 30° left, 30°right] – Presence of eye wear – Presence of face mask – Presence of head covering – Colour of clothing – Degree of occlusion – Lighting conditions
Object Handover	– Successful handover – Recognition of failure cause in case of failure	– Object to be handed over – Person's pose [standing, sitting, laying] – Person's behaviour before grasp [reaches out, does not reach out] – Person's behaviour during grasp [grasps object, does not grasp object] – Person's behaviour after grasp [keeps object in hand, drops object]

previous exposure to robots, occupational field of the robot [32] and age [9]. Social acceptance can be measured based on different criteria such as safety, impact on care and quality of life in the case of healthcare robots, and user satisfaction [9]. In a competition setting, an effort must be made to recruit people who are the target users for the robots, and to select the criteria relevant for the task. For example, for a healthcare robot, older adults or healthcare professionals would be ideal participants, and measuring all of the factors mentioned above would be relevant.

Social acceptance has been evaluated through user studies in competitions such as the Smart City Robotics Challenge [38], which was conducted in a public mall. The questionnaires were filled by users who were randomly selected to be a part of the scenario involving the robot taking an elevator, and answered questions relating to the social behaviour and collaboration of the robot, and the proxemics between the robot and human. In HEART-MET, the *Physically Assistive Robot Challenge* TBM requires a robot to fetch an item for a person. The evaluation considers both the functional capabilities, and the social acceptance of the robot. The TBM is thus scored using two metrics:

1. a technical score which is based on the robot successfully completing phases of the task (acknowledging which item is to be fetched from where, grasping the item, and handing it over to the person), and

2. a social acceptance score which is based on the robot's behaviour while interacting with the person (including aspects such as not causing surprise, interrupting a conversation at an appropriate time, announcing intentions, socially-aware navigation, etc.).

The social acceptance score is computed from a questionnaire[10] filled out by the person(s) interacting with the robot (and optionally by external observers).

In a competition, limitations of the results of questionnaires might result from

(i) limited time, which makes it difficult to conduct multiple trials,
(ii) availability of target users to answer the questionnaires
(iii) heterogeneity of participating robots in terms of physical appearance and capabilities.

The first two issues may be mitigated by recording videos of the interactions which can be used to conduct questionnaires later with a larger audience. Despite potential limitations, it is still important to evaluate a user's trust in functional and social savvy [16].

4 Conclusions

Benchmarking robots through competitions should place an emphasis on non-functional qualities of robots in addition to functional capabilities in order to encourage the development of trustworthy robots. The benchmarking procedure proposed in METRICS facilitates increasing levels of evaluation complexity while also accounting for the expense involved in physical robot benchmarks. The characteristics of benchmarks at each level, such as the composition of the datasets and the definitions of independent variables and failure conditions in the field campaigns, can influence the trustworthiness of benchmark results. Since robot embodiment is closely linked to perceived trust, the robot platform being used for a benchmark must be taken into consideration as well. Although a large variation in the data, inclusion of OOD conditions, failure scenarios, and user studies for robot acceptance contribute to the evaluation of non-functional qualities, it still remains an open question how one can quantify such qualities in order to objectively evaluate the trustworthiness of different types of robot platforms.

Acknowledgements This work is supported by funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252 (METRICS). We acknowledge National Metrology and Testing Laboratory (LNE) and the other METRICS partners for their work in developing the evaluation framework and evaluation plans for the competitions. Some sections of this chapter appeared in the *RO-MAN 2022 Workshop on Responsible Robotics: Robots with and for society*; we would like the thank the reviewers and attendees for their feedback.

[10] https://drive.google.com/file/d/1wcCQbGgaC9__hcngoeSSPN72wkq7YvnH/view.

References

1. Amigoni, F., Bastianelli, E., Berghofer, J., Bonarini, A., Fontana, G., Hochgeschwender, N., Iocchi, L., Kraetzschmar, G., Lima, P., Matteucci, M., et al.: Competitions for benchmarking: task and functionality scoring complete performance assessment. IEEE Robot. Autom. Mag. **22**(3), 53–61 (2015)
2. Ashmore, R., Calinescu, R., Paterson, C.: Assuring the machine learning lifecycle: desiderata, methods, and challenges. ACM Comput. Surv. (CSUR) **54**(5), 1–39 (2021)
3. Avrin, G., Barbosa, V., Delaborde, A.: AI evaluation campaigns during robotics competitions: the METRICS paradigm. In: 1st International Workshop on Evaluating Progress in Artificial Intelligence (EPAI 2020) in Conjunction with ECAI 2020 (2020)
4. Basiri, M., Piazza, E., Matteucci, M., Lima, P.: Benchmarking functionalities of domestic service robots through scientific competitions. KI-Künstliche Intelligenz **33**(4), 357–367 (2019)
5. Beck, K.: Test-driven Development: By Example. Addison-Wesley Professional (2003)
6. Bertoglio, R., Fontana, G., Matteucci, M., Facchinetti, D., Berducat, M., Boffety, D.: On the design of the agri-food competition for robot evaluation (ACRE). In: 2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). pp. 161–166 (2021). https://doi.org/10.1109/ICARSC52212.2021.9429792
7. Bhatt, U., Antorán, J., Zhang, Y., Liao, Q.V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al.: Uncertainty as a form of transparency: measuring, communicating, and using uncertainty. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 401–413 (2021)
8. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The fishyscapes benchmark: measuring blind spots in semantic segmentation. Int. J. Comput. Vis. **129**(11), 3119–3135 (2021)
9. Broadbent, E., Stafford, R., MacDonald, B.: Acceptance of healthcare robots for the older population: review and future directions. Int. J. Soc. Robot. **1**, 319–330 (2009)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
11. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. Int. J. Comput. Vis. (IJCV) **130**, 33–55 (2022). https://doi.org/10.1007/s11263-021-01531-2
12. European Association of Research and Technology Organisations: The TRL Scale as a Research & Innovation Policy Tool, EARTO Recommendations (2014). https://www.earto.eu/wp-content/uploads/The_TRL_Scale_as_a_R_I_Policy_Tool_-_EARTO_Recommendations_-_Final.pdf. Accessed 11 Aug 2023
13. Ferri, G., Ferreira, F., Faggiani, A., Fabbri, T.: From ERL to RAMI: expanding marine robotics competitions through virtual events. In: OCEANS 2021: San Diego–Porto, pp. 1–8. IEEE (2021)
14. Fiorini, L., Cornacchia Loizzo, F.G., Sorrentino, A., Rovini, E., Di Nuovo, A., Cavallo, F.: The VISTA datasets, a combination of inertial sensors and depth cameras data for activity recognition. Sci. Data **9**(1), 218 (2022)
15. Fontana, G., Matteucci, M., Amigoni, F., Schiaffonati, V., Bonarini, A., Lima, P.U.: RoCKIn benchmarking and scoring system. In: RoCKIn-Benchmarking Through Robot Competitions. IntechOpen (2017)
16. Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., Ivaldi, S.: Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers. Comput. Hum. Behav. **61**, 633–655 (2016)
17. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Commun. ACM **64**(12), 86–92 (2021)
18. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE Trans. Pattern Anal. Mach. Intell. (2022)

19. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors **53**(5), 517–527 (2011)
20. He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T.M., Mehnen, J.: The challenges and opportunities of human-centred AI for trustworthy robots and autonomous systems. IEEE Trans. Cogn. Dev. Syst. (2021)
21. Jentzsch, S.F., Hochgeschwender, N.: Don't forget your roots! using provenance data for transparent and explainable development of machine learning models. In: 2019 34th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW), pp. 37–40. IEEE (2019)
22. Kimble, K., Van Wyk, K., Falco, J., Messina, E., Sun, Y., Shibata, M., Uemura, W., Yokokohji, Y.: Benchmarking protocols for evaluating small parts robotic assembly systems. IEEE Robot. Autom. Lett. **5**(2), 883–889 (2020)
23. Lomonaco, V., Maltoni, D.: CORe50: a new dataset and benchmark for continuous object recognition. In: Levine, S., Vanhoucke, V., Goldberg, K. (eds.) Proceedings of the 1st Annual Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 78, pp. 17–26. PMLR (13–15 Nov 2017)
24. Malle, B.F., Ullman, D.: A Multi-dimensional conception and measure of human-robot trust. In: Trust in Human-Robot Interaction, pp. 3–25. Elsevier (2021)
25. Nguyen, M., Hochgeschwender, N., Wrede, S.: An analysis of behaviour-driven requirement specification for robotic competitions. In: Proceedings of the 5th International Workshop on Robotics Software Engineering (2023)
26. Nitto, H., Taniyama, D., Inagaki, H.: Social acceptance and impact of robots and artificial intelligence. Nomura Res. Inst. Pap. **211**, 1–15 (2017)
27. Norris, W.R., Patterson, A.E.: System-level testing and evaluation plan for field robots: a tutorial with test course layouts. Robotics **8**(4), 83 (2019)
28. Ortega, A., Hochgeschwender, N., Berger, T.: Testing service robots in the field: an experience report. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 165–172 (2022). https://doi.org/10.1109/IROS47612.2022.9981789
29. Pérez-Grau, F.J., Barriga, P.L., Viguria, A.: Lowering the entry barrier to aerial robotics competitions. In: 2023 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 487–492. IEEE (2023)
30. Ranieri, C.M., MacLeod, S., Dragone, M., Vargas, P.A., Romero, R.A.F.: Activity recognition for ambient assisted living with videos, inertial units and ambient sensors. Sensors **21**(3), 768 (2021)
31. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
32. Savela, N., Turja, T., Oksanen, A.: Social acceptance of robots in different occupational fields: a systematic literature review. Int. J. Soc. Robot. **10**(4), 493–502 (2018)
33. Schneider, S., Hegger, F., Hochgeschwender, N., Dwiputra, R., Moriarty, A., Berghofer, J., Kraetzschmar, G.K.: Design and development of a benchmarking testbed for the factory of the future. In: 2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA), pp. 1–7. IEEE (2015)
34. da Silva, A.S., Kreutz, A., Weiss, G., Rothe, J., Ihrke, C.: DevOps in robotics: challenges and practices. In: European Conference on Software Architecture, pp. 284–299. Springer (2022)
35. So, P., Wittmann, J., Ruhkamp, P., Sarabakha, A., Haddadin, S.: Towards Remote Robotic Competitions: An Internet-Connected Task Board and Dashboard (2022). arXiv:2201.09565
36. Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Upcroft, B., Abbeel, P., Burgard, W., Milford, M., et al.: The limits and potentials of deep learning for robotics. Int. J. Robot. Res. **37**(4–5), 405–420 (2018)
37. Thoduka, S., Hochgeschwender, N.: Benchmarking robots by inducing failures in competition scenarios. In: Duffy, V.G. (ed.) Digital Human Modeling and Applications in Health, Safety,

Ergonomics and Risk Management. AI, Product and Service, pp. 263–276. Springer International Publishing, Cham (2021)
38. Wang, L., Iocchi, L., Marrella, A., Nardi, D.: Developing a questionnaire to evaluate customers' perception in the smart city robotic challenge. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1–6. IEEE (2019)
39. Weiss, A., Bernhaupt, R., Lankes, M., Tscheligi, M.: The USUS evaluation framework for human-robot interaction. In: AISB2009: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction, vol. 4, pp. 11–26 (2009)
40. Yang, J., Ren, Z., Xu, M., Chen, X., Crandall, D.J., Parikh, D., Batra, D.: Embodied amodal recognition: learning to move to perceive objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2040–2050 (2019)