This is a pre-print (author's version) of the paper. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-77820-0_20.

Benchmarking Robots by Inducing Failures in Competition Scenarios

Santosh Thoduka and Nico Hochgeschwender

Hochschule Bonn-Rhein-Sieg,
Department of Computer Science
53757 Sankt Augustin, Germany
{santosh.thoduka, nico.hochgeschwender}@h-brs.de

Abstract. Domestic service robots are becoming more ubiquitous and can perform various assistive tasks such as fetching items or helping with medicine intake to support humans with impairments of varying severity. However, the development of robots taking care of humans should not only be focused on developing advanced functionalities, but should also be accompanied by the definition of *benchmarking protocols* enabling the rigorous and reproducible evaluation of robots and their functionalities. Thereby, of particular importance is the assessment of robots' ability to deal with failures and unexpected events which occur when they interact with humans in real-world scenarios. For example, a person might drop an object during a robot-human hand over due to its weight. However, the systematic investigation of hazardous situations remains challenging as *(i)* failures are difficult to reproduce; and *(ii)* possibly impact the health of humans. Therefore, we propose in this paper to employ the concept of scientific robotic competitions as a benchmarking protocol for assessing care robots and to collect datasets of human-robot interactions covering a large variety of failures which are present in real-world domestic environments. We demonstrate the process of defining the benchmarking procedure with the human-to-robot and robot-to-human handover functionalities, and execute a dry-run of the benchmarks while inducing several failure modes such as dropping objects, ignoring the robot, and not releasing objects. A dataset comprising colour and depth images, a wrist force-torque sensor and other internal sensors of the robot was collected during the dry-run. In addition, we discuss the relation between benchmarking protocols and standards that exist or need to be extended with regard to the test procedures required for verifying and validating conformance to standards.

Keywords: robotics competitions · benchmarking · assistive robots.

1 Introduction

The Multi-Annual Roadmap for robotics in Europe identifies healthcare as one of the domains in which robotics is expected to play a significant role [7]. Assistive robotics, which is seen as one of the sub-domains along with clinical and rehabilitation robotics, is concerned with providing assistive aid to care givers or to

persons with physical, sensory or cognitive impairments. Their acceptability and impact on such users is increasing, though most robots are still in the research phase [22, 10]. The challenges identified for deploying robots include updating safety standards to ensure a certain level of robustness [22]. Research is also ongoing in areas such as designing the robots to cater to user needs [23], and assessing the impact of personality factors on the acceptance of socially assistive robots [25].

Typical tasks for such robots include fetching items, engaging in conversation, assisting with medicine intake, monitoring a person's activity, etc. They work in close proximity to humans, with some tasks such as object handovers requiring close contact. Therefore the robots must be guaranteed to perform their activities in a safe manner before they can be deployed in care facilities or in homes.

The evaluation of functionalities developed for these robots is often performed in laboratories or in a limited number of settings. Developing benchmarking protocols will enable a more rigorous evaluation of such robotic systems, ensuring that the functionalities of the robot are reproducible in various environments, and repeatable across several trials. Benchmarking is also one of the ways in which the conformance to standards can be improved. The standard ISO 13482 [15], which provides requirements and guidelines for designing personal care robots, requires that all "performance values related to safety of the robot shall be verified and validated." Some of the methods listed for achieving this is through testing the various performance values of the robot under normal and abnormal conditions, injecting faults during tests, endurance tests and observation during operation. The standard ISO 12100 [14], which provides guidelines for risk assessment and risk reduction in machinery, requires the estimation of the probability of occurrence of identified hazardous events. Since the estimation is typically based on historical and statistical data, this is a challenge for robots that have not been put through rigorous testing yet. Several functionalities of assistive robots are still in the research phase; therefore there is a lack of data concerning the likelihood of failure of different components such as perception or grasping. Additionally, since the likelihood of failure is often a function of the robot's morphology and sensor configuration, it is hard to share data between different robots types.

It is even more challenging to quantify failures in the case of functionalities which require interaction with humans or the environment. A failure, leading to a hazardous event, could be caused by several factors, including the behaviour of the human. For example, during a robot-to-human handover, the robot might release the object shortly after feeling a pulling force, but the human might also release the object immediately if they feel the robot is not releasing the object. This would result in a potentially hazardous event (such as a knife falling down), but identifying the exact cause of the failure and estimating the likelihood of the event is a challenge.

Hence, incorporating failures and hazardous events in the benchmarking process is necessary to fully evaluate a robotic system and is a step towards ensuring compliance with standards for safety and risk reduction.

In the context of scientifically and rigorously evaluating assistive robots, we attempt to answer the following questions:

1. *How should benchmarking protocols for social and physically assistive robots be defined?* We propose employing scientific competitions as the means to benchmark functionalities and task execution of robots, which has been shown to be a viable method in prior work such as RoCKIn [4]. We elaborate on this in Sects. 2 and 3.
2. *What is the methodology to define benchmarks which incorporate failure conditions in the interaction between robots and humans?* We propose a procedure which defines the expected function and corresponding failure modes, and incorporates these as independent variables in the experiment design. Failure modes which involve interaction are induced by the human volunteer participating in the experiment. We discuss this with the help of a use case in Sect. 4.
3. *How can benchmarking robots with this methodology help them conform to existing standards?* In Sect. 5, we discuss existing standards which define safety guidelines and risk assessment and reduction procedures and how they relate to the methodology discussed.

2 Related Work

Benchmarking in robotics is an active field of research, whether focused on benchmarking complete systems [29, 21], benchmarking tasks such as pick-and-place [28, 24] and navigation [26], or benchmarking hardware components such as end-effectors [11]. Typically, they define a benchmarking framework by specifying the task, environment and evaluation metrics, allowing users to replicate the setup in their own facilities. In the case of [29], the authors develop a simulation-based benchmarking platform using software containers to enhance reproducibility.

Benchmarking has also been carried out through robotic competitions. The DARPA robotics challenge [1], RoboCup [3] and World Robot Challenge [6] are all robotics competitions which target robot applications in different domains such as rescue, agile manufacturing, service robotics, etc. More recently, projects such as RoCKIn [4] and RockEU2 [5], which have now been brought under the umbrella of the European Robotics League [2], have brought competitions closer to scientific experiments by comparing the performance of robots in certified test beds such that the results are reproducible and repeatable [8]. They target three domains, namely, domestic service robots, industrial robots working in a smart factory environment and rescue robots. Benchmarks for domestic service robots include functionalities such as object recognition and navigation, and tasks such as *Getting to know my home*, *Catering for Granny Annie's comfort*, in which the robot is operating in the home of *Granny Annie* performing various domestic tasks [9].

RoCKIn [4] introduced the concept of functionality and task benchmarks, separating the evaluation of standalone functionalities and complete tasks which

require the integration of several functionalities. This was a differentiating factor from previous competitions, which typically evaluated robots on their performance in a complete task. By introducing the benchmarking of standalone functionalities, the evaluation is more fine-grained, while limiting the influence of one functionality on another. However, while these competitions evaluate functionalities of service robots, there is no explicit benchmarking of factors such as safety and resilience to failures. Unsafe actions such as collisions result in a penalty or disqualification, but is not the focus of the benchmarks.

Safety and autonomy are two of the benchmarks proposed in [12] for evaluation socially assistive robots, amongst others such as scalability, privacy, impact on user's care etc. The authors regard both safety of the robot itself, and ensuring safety of the user as important factors. Autonomy is considered from the viewpoint of whether the robot can effectively perform its tasks, and whether the user can trust the robot to perform them. Resilience to failure conditions or unexpected situations is therefore a crucial aspect of autonomy.

Tolmeijer et al. [27] propose a taxonomy of failures which affect trust during human-robot interaction. Four failure types are identified: *(i)* design: failures caused by the robot performing as designed, but not as expected by the user, *(ii)* system: failures caused by the robot not performing as designed, *(iii)* expectations: similar to the design failure except that the behaviour of the robot is still considered correct, and the user's expectations should be corrected to mitigate the failure, and *(iv)* user: failures caused by actions by the user which were not expected by the robot. Several mitigation strategies are suggested for regaining trust of the user, such as apologizing, proposing alternate actions, or asking the user for a justification of their behaviour.

Using competitions as a means to benchmark robots has been shown to be successful. However, in particular for assistive robots, they must be updated to incorporate aspects such as resilience to failure, trust, safety, etc. We utilize the benchmarking procedures from RoCKIn [4] and extend it to explicitly include failure conditions as part of the variability of the benchmark.

3 Scientific Competitions for Robot Benchmarking

The domestic service robotics competitions in RoboCup, RoCKIn, and RockEU2 target robots operating in a home environment performing various domestic tasks, including iteracting with people. Assistive robots in healthcare environments are closely related but with the additional caveat that they could interact with persons with impairments. This makes safety a crucial aspect in designing the robots, and by extension competitions, and benchmarking protocols.

3.1 HEART-MET: Healthcare Robotics Technologies - Metrified

The METRICS project[1] aims to address the need for benchmarking robots in the four priority areas healthcare, inspection and maintenance, agri-food and agile

[1] https://metricsproject.eu/

production by organizing robotics competitions in the four areas under a common evaluation framework. The healthcare competition, Healthcare Robotics Technologies - Metrified (HEART-MET), targets assistive robots which perform care-related tasks by benchmarking typical tasks in a care facility or private home. Since the tasks involve the robot operating in dynamic and unstructured environments, and interacting with older adults with physical, sensory and cognitive impairments, it is necessary to validate the safety of the robot. The HEART-MET evaluation plan [13] describes the competition in more detail and includes definitions of various functionality and task benchmarks.

The competitions are of two types: (a) field evaluation campaigns (FEC), in which robots compete at a certified physical test bed; and (b) cascade evaluation campaigns (CEC), which are online competitions evaluated on datasets. The datasets for the cascade evaluation campaign are collected from robots competing in the field evaluation campaigns. The CEC thus allows teams to improve specific functionalities which do not require the use of a robot, but nevertheless benefit from realistic datasets which have been generated by several different robot platforms.

Following the benchmark types introduced in RoCKIn [4], the FEC comprises of two types of benchmarks: (a) functionality benchmarks (FBM), which evaluate individual functionalities of the robot; and (b) task benchmarks (TBM), which evaluate the execution of a complete task. Functionality benchmarks include object detection, human recognition, activity recognition, human-to-robot and robot-to-human handover, task-oriented grasping, opening a cupboard etc. Task benchmarks include the delivery of a requested item to a person, preparing a drink, and assessing the activity state of a person. Robots are expected to run multiple trials, particularly for FBMs, and are evaluated based on aggregated metrics (such as true positive rate) over all trials. In benchmarks where quantitative metrics are not applicable, scoring is based on intermediate achievements during the trial. For example, for item delivery, intermediate achievements include navigating to the pickup location, detecting the item, grasping the item, etc.

For both FBMs and TBMs, there is a special emphasis on the feature variations introduced for each trial. Some variations are simply the configuration of the task, such as the object to be detected, location of the task, or the person involved in handover task. Some variations are introduced to evaluate the resilience of the robot to unexpected situations. This applies in particular to tasks which involve interaction of the robot with a human or the environment. For example, while opening a cupboard, the robot might encounter a stuck door, or items could fall out of the cupboard.

3.2 Defining a Benchmark

In order to specify the benchmarking procedure for a given functionality or task, we first identify and specify several aspects of the benchmark, i.e *(i)* define the objective of the benchmark, *(ii)* identify dependent and independent variables, *(iii)* identify failure modes associated with the execution of the benchmark,

(iv) specify evaluation metrics (or achievements for non-quantitative evaluation), and *(v)* specify minimum data sources to be recorded.

An example of this process for the activity recognition benchmark can be seen in Fig. 1. The objective must concisely state the desired objective of the robot

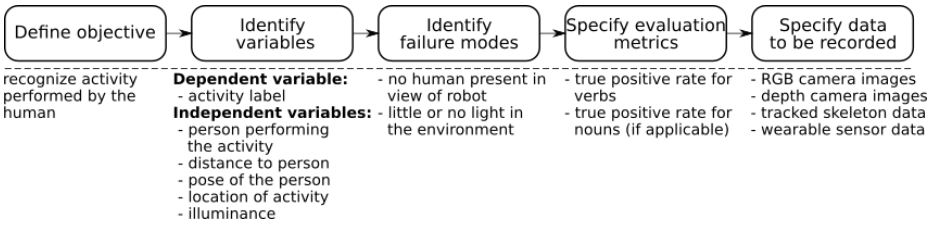


Fig. 1: Procedure for defining a benchmark using activity recognition as an example

when performing the benchmark. This is later used to identify failure modes by considering different aspects of the objective which are not able to be fulfilled.

Variables for the benchmark include the dependent variables (i.e. those which are the target of the benchmark), and independent variables which are varied across trials. The failure modes are a particular type of independent variable, which present the robot with some type of abnormal condition. In the example shown in Fig. 1, the robot does not interact with a human or environment, hence the failure modes correspond to the configuration of the environment. The absence of a human renders the objective of the benchmark unachievable while a low light environment possibly makes the objective unachievable for most robots. Failure modes can also be caused internal hardware and software faults (such as a broken camera), but our focus here is on failure modes arising from external factors (i.e. through configuration of the scene, or behaviour of humans during interaction). When such a situation is encountered during the execution of a benchmark, the robot scores an achievement if it is able to identify that it is an abnormal condition.

The final step is to identify data which is to be recorded during the execution of the benchmark. The actual data recorded may differ based on the robot platform and available sensors. Recording the right data is important since it will allow the careful design of challenges for the cascade campaign, which target the same benchmark (e.g. activity recognition in the case of Fig. 1), or challenges that address the detection of failure conditions (as in the case for handovers described in Sect. 4).

3.3 Execution and Dataset Collection

The execution of trials is controlled using a referee box, which communicates with the robot by sending a start signal and type of benchmark to be executed.

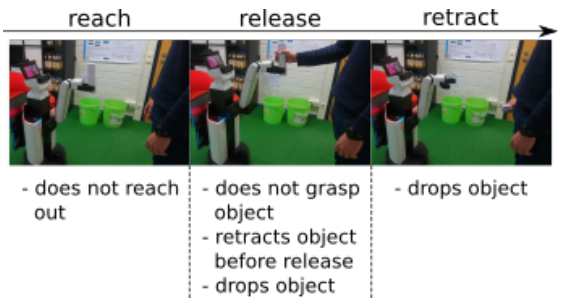
The robot sends back the result of the trial once it has completed. For example, for a trial of the activity recognition FBM, the robot is placed near a person, and the referee box sends a start signal. The person is instructed to perform the activity specified by the referee box, which the robot recognizes and finally returns a result message with the recognized activity.

Collecting data during the execution of a benchmark serves two purposes: *(i)* it serves as a method of recording the test conditions, including available sensors on the robot, a third-person view of the functionality, and results, all of which can be analyzed at a later time; and *(ii)* it aids in the creation of realistic datasets which can be used for improving the performance of algorithms related to that benchmark. In HEART-MET, data collected during the execution of a benchmark in a FEC will later be labelled and used as test sets in the CEC. The recorded data can include sensors on the robot (such as cameras, force-torque sensors, laser scanner etc.), external cameras, smart home sensors, wearable sensors etc. For each trial, the robot begins recording internal sensor data in the form of ROS bagfiles when triggered by the start signal from the referee box, and stops once the trial is complete. The recorded data, result messages, and trial configurations on the referee box are collated to form a partially labelled dataset. For some FBMs, such as object detection, additional annotation of the recorded data is needed to create a fully labelled dataset.

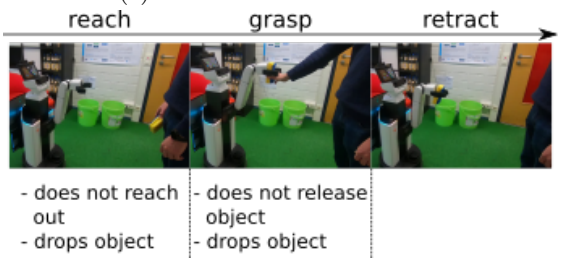
4 Use Case: Handover FBM

In this section, we use the robot-to-human and human-to-robot handover as a use case to describe the benchmark specification process in more detail and to exemplify the data collection process which includes interaction failures. Receiving and giving objects are essential skills for an assistive robot. Several challenges exist, including variablity in the type of objects, coordinating the interaction with the human in a natural manner, and ensuring a safe and fault-free execution. In defining the benchmark for the handover functionality, we focus primarily on evaluating the robot's capability of handling failure scenarios. We follow the procedure defined in Sect. 3.2 to define the benchmarking protocol for this functionality.

Objective The objective of the handover functionality is to safely transfer an object from the giver to the receiver. "*Safely*" refers to ensuring the safety of the robot itself, the human, the object being handed over and the surrounding environment. An additional requirement is that the handover occurs in a manner that is intuitive to the person, which includes executing the handover at a comfortable position between the human and robot, comfortable grasp pose on the object and timely release or grasp of the object. These additional requirements are subjective and require feedback from the persons involved for evaluation. We focus here on the primary objective, though the subjective requirements can be included in the benchmarking protocol using the same process.



(a) Robot-to-human handover



(b) Human-to-robot handover

Fig. 2: The phases and interaction failure modes caused by the human for (a) robot-to-human and (b) human-to-robot handover

Variables The location, time of day and lighting conditions are variables common to several benchmarks, and are included for this benchmark as well. For this benchmark, the object, the person and the person's pose (such as sitting, standing, or laying) are specified as independent variables as well.

Failure Modes Figure 2 illustrates the phases and the associated failure modes for both handover functionalities. The failure modes do not indicate the *cause* of the failure, but simply the type of failure that can occur. For example, for the failure mode *does not grasp object*, the cause could be that the human no longer wants the object, the object is too far away, or that there is no suitable grasp position available. For the purposes of benchmarking, we are currently only interested in the failure mode, and not the causes. Once the failure modes are enumerated, they are used to generate variations in the benchmark trials by instructing the human to either behave nominally or to induce one or more of the failure modes. Some sample trial configurations for a robot-to-human handover benchmark are shown in Table 1. The volunteer receives an instruction for each phase of the handover.

Evaluation Metrics The evaluation of this benchmark is based on intermediate achievements. In the nominal case, the achievements are successfully reaching

No.	Object	Location	Time	Reach	Release	Retract
1	bottle	kitchen	morning	reach out	grasp object	-
2	towel	bedroom	evening	do not reach out	-	-
3	pill box	living room	afternoon	reach out	do not grasp object	-
4	book	living room	morning	reach out	grasp object	drop object

Table 1: Sample trial configurations for a robot-to-human handover with instructed behaviour for the human volunteer

out, grasping or releasing the object, and retracting. In the case where a failure mode is induced, the robot instead scores an achievement if it detects the failure mode. For example, if the human drops the object during the handover, the robot scores an achievement if it detects that the object has dropped.

The trial configurations, each of which consists of an instantiation of the defined variables, are fixed beforehand, and all teams must execute all trials. Some variants, such as the lighting conditions, are difficult to reproduce for all teams, but a best effort is made to achieve uniformity in such cases. The failure modes are easily induced by instructing the human volunteer to behave in a certain way for a particular trial.

Fig. 3: Trials the handover functionalities consist of successful handovers, failed trials including unreleased objects, dropped objects, and ignoring the robot (top to bottom)

Data The data collected during this benchmark will be used for improving the detection of the failure modes. Therefore, any available sensors that the robot might use for detecting the failures are recorded. Figure 3 shows a subset of the frames captured from the robot's camera during several trials, in which the volunteers completed the handover successfully, did not release the object, dropped objects and did not respond to the robot. Other data recorded includes depth images, RGB images from the end-effector camera, force-torque sensor at the wrist, and proprioceptive sensors of the joints. Figure 4 shows force measurements by the force-torque sensor and corresponding image frames from a second camera on the robot. The force caused by the interaction is evident in the top plot at the start of the *release* phase, and the reduced downward force once the object leaves the robot arm is visible in the bottom plot towards the end of the *release* phase.

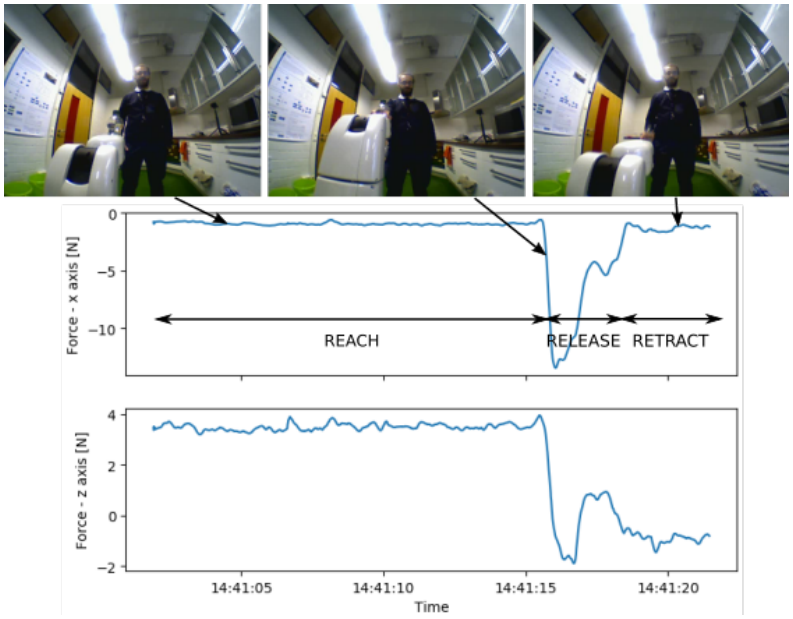


Fig. 4: Force measurements from the wrist force-torque sensor during a robot-to-human handover

The videos and other sensor data are extracted and labelled with the outcome of each phase of the handover. Since the handover functionality cannot be evaluated in a cascade campaign, the dataset recorded during the field campaign is used to create two related dataset challenges: *(i)* to detect whether the handover was successful or not; and *(ii)* to detect the failure mode if the handover was unsuccessful. During a dry-run of the field evaluation campaign conducted in our lab, 150 trials of the robot-to-human handover and 144 trials of the human-

to-robot handover were executed with nine volunteers, of which 93 and 117 trials respectively resulted in a failure. Since the robot was not equipped with the functionality to detect certain failures, it only achieved points for successful handovers and for detecting when a person did not respond to the robot. However, the dataset will enable the development of functionality to detect dropped objects and unreleased objects for future campaigns.

The benchmark protocol defined here and the data collection process can be easily replicated and extended by the community. A tool which coordinates execution of the benchmark, and generates trial configurations based on the variations defined for the benchmark has also been made available[2].

5 Related Standards

Benchmarks and standards are closely related since benchmarks can provide a means for measuring the conformance to standards. In defining the benchmarking protocols, we want to identify synergies with existing benchmarks and ways to define them in such a way that they can be used as verification and validation procedures required by standards.

The International Organization for Standardization (ISO) has defined several standards for safety of machines. ISO 12100:2010 [14] provides a framework for designing safe machines, including guidelines for risk assessment and reduction. The process of risk assessment and reduction begins by identifying limits of the machine and identifying the risk of potential hazards and the likelihood of their occurrence. Removing the hazard, or reducing the risk of the hazard is performed by incorporating protective measures. The standard is applicable to all machines, and hence to robotic systems as well. Benchmarking protocols will primarily aid in the development of risk reduction strategies and to some extent help in estimating the likelihood of the occurrence of hazards.

The technical committee ISO TC/299 is concerned with standards related to robotics including safety, performance criteria and test methods. Among the standards developed by this committee is the safety standard ISO 13482 [15], which provides requirements and guidelines for designing personal care robots. It defines safety requirements for hazards caused by robot motion, a charging battery, environmental conditions, localization errors etc., and guidelines for protective measures against each type of hazard. Following ISO 12100, these requirements are to be used to perform a risk analysis of the robot, with the application of protective measures if necessary. Several options for verifying and validating that the robot conforms to the requirements are also specified. For example, for hazards due to incorrect autonomous decisions and actions, the verification and validation methods are practical tests, measurement, observation during operation, examination of software and review of task-based risk assessment. Practical tests and observation during operation both involve subjecting the robot to abnormal conditions in addition to normal operating conditions.

[2] https://github.com/HEART-MET/metrics_refbox

ISO/TR 23482-1 [19] defines test methods that can be used ensure compliance with ISO 13482, and ISO/TR 23482-2 [20] provides additional guidelines to design robots according to ISO 13482. Similarly, ISO 18646-1 and ISO 18646-2 [17, 18] define several performance characteristics of the locomotion, and navigation of service robots, such as rated speed, stopping characteristics, turning width etc., along with recommendations on how to test them. They include specific details of the recommended test facility, test procedure and reported results for each of the performance characteristics. ISO 15066:2016 [16] is concerned with safety requirements for collaborative robots and identifies the collaboration types safety-rated monitored stop, hand guiding, speed and separation monitoring, and power and force-limiting. For tasks which involve a physical interaction between the robot and human, the power and force-limiting collaboration type is most applicable since intentional or unintentional contact with the robot is expected.

Our goal in defining benchmarking protocols is to enhance and extend the test procedures already defined, with a focus on robot-human interaction. We define the test facility (a certified test-bed), test procedure incorporating the failure modes, and the evaluation metrics to report results for each benchmark. The manner in which the robot responds to failure conditions is also an aspect which can benefit from standardization. Even though some standards have similar activities which can be found in the proposed benchmarking protocols (e.g. defining the scope and task as defined in ISO 12100) the existing standards lack, to the best of our knowledge, test procedures to assess the failure conditions occurring in human-robot interaction tasks such as object handover.

6 Conclusions

In this paper we proposed scientific competitions as a means to benchmark functionalities for assistive robots with a particular focus on failure modes, especially in tasks that involve human-robot interaction. The process for defining a benchmark comprises of defining the objective, identifying the variables and failure modes, and specifying the evaluation metrics and data to be recorded. We demonstrated the feasibility of the proposed approach with the help of a handover use-case, which incorporates several failure modes associated with the interaction between the human and the robot. While related standards share several activities as those included in the benchmarking protocol the investigation of failure conditions are not yet present in robotic standards. For the future we aim to establish further synergies and harmonize the activities between standards and benchmarking protocols developed in competitions as both share the common goal to systematically assess the performance of robotic systems.

7 Acknowledgement

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

References

1. Darpa Robotics Challenge. https://archive.darpa.mil/roboticschallenge/, accessed: January 2021
2. European Robotics League. https://www.eu-robotics.net/robotics_league/, accessed: January 2021
3. RoboCup. https://www.robocup.org/, accessed: January 2021
4. Robot Competitions Kick Innovation In Cognitive Systems and Robotics. http://rockinrobotchallenge.eu/, accessed: January 2021
5. Robotics Coordination Action for Europe Two. https://www.eu-robotics.net/eurobotics/about/projects/rockeu2.html, accessed: January 2021
6. World Robot Challenge. https://worldrobotsummit.org/en/wrs2020/challenge/, accessed: January 2021
7. Robotics 2020 Multi-Annual Roadmap for Robotics in Europe. Technical guide, SPARC - Partnership for Robotics in Europe (2016)
8. 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 Robotics & Automation Magazine **22**(3), 53–61 (2015)
9. 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)
10. Cavallo, F., Esposito, R., Limosani, R., Manzi, A., Bevilacqua, R., Felici, E., Di Nuovo, A., Cangelosi, A., Lattanzio, F., Dario, P.: Robotic Services Acceptance in Smart Environments With Older Adults: User Satisfaction and Acceptability Study. Journal of Medical Internet research **20**(9), e264 (2018)
11. Falco, J., Hemphill, D., Kimble, K., Messina, E., Norton, A., Ropelato, R., Yanco, H.: Benchmarking Protocols for Evaluating Grasp Strength, Grasp Cycle Time, Finger Strength, and Finger Repeatability of Robot End-Effectors. IEEE robotics and automation letters **5**(2), 644–651 (2020)
12. Feil-Seifer, D., Skinner, K., Matarić, M.J.: Benchmarks for evaluating socially assistive robotics. Interaction Studies **8**(3), 423–439 (2007)
13. Hochgeschwender, N., Thoduka, S., Dragone, M., Caleb-Solly, P., Bellamy, D., Cavallo, F.: HEART-MET Evaluation Plan. https://metricsproject.eu/healthcare/ (2020), accessed: February 2021
14. Safety of machinery — General principles for design — Risk assessment and risk reduction. Standard, International Organization for Standardization (2010)
15. Robots and robotic devices - Safety requirements for personal care robots. Standard, International Organization for Standardization (2014)
16. Robots and robotic devices — Collaborative robots. Standard, International Organization for Standardization (2016)
17. Robotics - Performance criteria and related test methods for service robots — Part 1: Locomotion for wheeled robots. Standard, International Organization for Standardization (2016)
18. Robotics - Performance criteria and related test methods for service robots — Part 2: Navigation. Standard, International Organization for Standardization (2019)
19. Robotics — Application of ISO 13482 — Part 1: Safety-related test methods. Standard, International Organization for Standardization (2010)
20. Robotics — Application of ISO 13482 — Part 2: Application guidelines. Standard, International Organization for Standardization (2010)

21. 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 Robotics and Automation Letters **5**(2), 883–889 (2020)
22. Kyrarini, M., Lygerakis, F., Rajavenkatanarayanan, A., Sevastopoulos, C., Nambiappan, H.R., Chaitanya, K.K., Babu, A.R., Mathew, J., Makedon, F.: A Survey of Robots in Healthcare. Technologies **9**(1), 8 (2021)
23. Lin, C.C., Liao, H.Y., Tung, F.W.: Design Guidelines of Social-Assisted Robots for the Elderly: A Mixed Method Systematic Literature Review. In: International Conference on Human-Computer Interaction. pp. 90–104. Springer (2020)
24. Morgan, A.S., Hang, K., Bircher, W.G., Alladkani, F.M., Gandhi, A., Calli, B., Dollar, A.M.: Benchmarking Cluttered Robot Pick-and-Place Manipulation With the Box and Blocks Test. IEEE Robotics and Automation Letters **5**(2), 454–461 (2019)
25. Rossi, S., Conti, D., Garramone, F., Santangelo, G., Staffa, M., Varrasi, S., Di Nuovo, A.: The Role of Personality Factors and Empathy in the Acceptance and Performance of a Social Robot for Psychometric Evaluations. Robotics **9**(2), 39 (2020)
26. Sprunk, C., Röwekämper, J., Parent, G., Spinello, L., Tipaldi, G.D., Burgard, W., Jalobeanu, M.: An Experimental Protocol for Benchmarking Robotic Indoor Navigation. In: Experimental Robotics. pp. 487–504. Springer (2016)
27. Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T.M., Dixon, C., Tielman, M.L.: Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. pp. 3–12 (2020)
28. Triantafyllou, P., Mnyusiwalla, H., Sotiropoulos, P., Roa, M.A., Russell, D., Deacon, G.: A Benchmarking Framework for Systematic Evaluation of Robotic Pick-and-Place Systems in an Industrial Grocery Setting. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 6692–6698. IEEE (2019)
29. Weisz, J., Huang, Y., Lier, F., Sethumadhavan, S., Allen, P.: Robobench: Towards Sustainable Robotics System Benchmarking. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 3383–3389. IEEE (2016)