NISTIR 8345

Workshop Report: Test Methods and Metrics for Effective HRI in Real-World Human-Robot Teams, ACM/IEEE Human-Robot Interaction Conference, 2020 (Virtual)

Shelly Bagchi
Jeremy A. Marvel
Megan Zimmerman
Murat Aksu
Brian Antonishek
Heni Ben Amor
Terry Fong
Ross Mead
Yue Wang

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8345

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8345

Workshop Report: Test Methods and Metrics for Effective HRI in Real-World Human-Robot Teams, ACM/IEEE Human-Robot Interaction Conference, 2020 (Virtual)

Shelly Bagchi, Jeremy A. Marvel, Megan Zimmerman, Murat Aksu, Brian Antonishek
Intelligent Systems Division
Engineering Laboratory

Heni Ben Amor
Arizona State University
Terry Fong
NASA Ames Research Center
Ross Mead
Semio AI, Inc.
Yue Wang
Clemson University

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8345

January 2021

U.S. Department of Commerce
Wynn Coggins, Acting Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology
Interagency or Internal Report 8345
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8345, 16 pages (January 2021)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8345

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

Abstract

This report details the second annual, full-day workshop exploring the metrology necessary for repeatably and independently assessing the performance of robotic systems in real-world human-robot interaction (HRI) scenarios. This workshop continues the motion toward bridging the gaps between the theory and applications of HRI, enabling reproducible studies in HRI, and accelerating the adoption of cutting edge technologies as the industry state-of-practice. The second annual workshop, ‘Test Methods and Metrics for Effective HRI’, seeks to identify test methods and metrics for the holistic assessment and assurance of HRI performance in practical applications. The workshop was held on March 23, 2020, as part of the virtual ACM/IEEE International Conference on Human-Robot Interaction.

The focus is on identifying the key performance indicators of seemingly disparate sectors, including manufacturing, social, medical, and service robot solutions among others, and to foster the community based on the principles of transparency, repeatability and reproducibility, and establishing trust. The goal is to aid in the advancement of HRI technologies through the development of experimental design, test methods, and metrics for assessing HRI and interface designs.

Keywords

Robotics; Human-Robot Interaction; Collaborative Robotics; Human-Robot Teaming; Metrics; Test Methods.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

Table of Contents

List of Tables

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

Disclaimer

Certain trade names and company products are mentioned in the text or identified in certain illustrations. In no case does such an identification imply recommendation or endorsement by the NIST, nor does it imply that the products are necessarily the best available for the purpose.

The opinions expressed in this Workshop Report are those of the workshop participants and are not the official opinions of NIST. The summaries of the presentations have been reviewed by the speakers and the summaries reflect the speaker's main points.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

1. Introduction

Despite large advances in robot interfaces and user-centric robot designs, the need for effective HRI continues to present challenges for the field of robotics. As new technologies are integrated into human-robot teams in a myriad of application domains, ranging from smart manufacturing to home automation to space exploration, exposures to and expectations of robots will continue to grow rapidly.

A key barrier to achieving effective human-robot teaming in a multitude of domains is that there are few consistent test methods and metrics for assessing HRI effectiveness. The desire for repeatable and consistent evaluations of HRI methodologies drives the need for validated metrology. Such evaluations are critical for advancing the underlying theory of HRI, as well as establishing traceable mechanisms for vendors and consumers of HRI technologies to assess and assure functionality.

The full-day workshop continued to address the issues surrounding the development of test methods and metrics for evaluating HRI performance across the multitude of application domains, including industrial, social, medical, field and service robotics. The morning session focused on establishing the diversity of approaches for addressing HRI metrology in different robotic domains, and featured talks by invited speakers and contributing authors. The afternoon session looked at the underlying issues of traceability, objective repeatability, and transparency in HRI metrology by means of late-breaking research poster presentations and open discussions with the HRI community. The need for establishing consistent standards for evaluating HRI in real-world applications inspired the creation of this workshop series. The 2020 workshop in particular focused on how the interfaces, technologies, and underlying theories impact the effective collaboration of human-robot teams. Specific goals included the following:

- To develop and encourage the use of consistent test methods and metrics in evaluating HRI technologies, producing quality data sets of pragmatic applications, and validating human subject studies for HRI;
- To establish benchmarks and baselines along a spectrum of key performance indicators for assessing and comparing novel HRI systems and applications;
- To support a discussion about best practices in metrology and what features should be measured as the underlying theory of HRI advances;
- To encourage the creation and sharing of high-quality, consistently-formatted datasets for HRI research; and
- To promote the development of reproducible, metrics-oriented studies that seek to understand and model the human element of HRI teams.

The workshop was held via BlueJeans on March 23, 2020, as part of the virtual ACM/IEEE International Conference on Human-Robot Interaction, and was well-attended, with a peak audience of about 35 members.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

1.1 Fields of Interest

This workshop continues to serve as a springboard for establishing a formalized and standardized HRI research community. Specific targeted interest groups include stakeholders in industrial, collaborative, medical, and service robotics:

- Researchers of novel HRI theories, applications, technologies, and systems;
- Researchers developing frameworks and models of real-world, human-robot teams;
- Researchers generating quality HRI datasets, or interested in consuming such datasets;
- Researchers studying the social impacts and acceptance of human-robot teaming;
- Researchers investigating HRI metrics, benchmarks, and other practices aimed at repeatable performance studies;
- End-users and consumers of HRI technologies;
- Manufacturers and integrators of HRI technologies; and
- Standards communities interested in performance metrics for human-robot systems.

1.2 Schedule and Format

Table 1 contains the workshop schedule. The structure followed that of last year's workshop. The first half of the day focused on the technical aspects of metrology for effective, real-world HRI and technical presentations of contributed research topics. The second half of the day focused on international efforts that explore repeatability, reproducibility, traceability, and the impacts of demographics, culture, and study design on the results of HRI research. Unfortunately, due to the last-minute switch to a virtual conference, the schedule had to be abbreviated to allow for attendees in multiple time zones. This also removed the planned breakout sessions, which the organizers felt would not be as effective virtually.

1.3 Discussion Topics

Presentations by contributing authors focused on the documentation of the test methods, metrics, and datasets used in their respective studies. Keynote and invited speakers were selected from a targeted list of HRI researchers across a broad spectrum of application domains. Poster session participants were selected from contributors reporting late-breaking evaluations and their preliminary results.

Discussions highlighted the various approaches, requirements, and opportunities of the research community toward assessing HRI performance, enabling advances in HRI research, and establishing trust in HRI technologies. Specific topics of discussion included:

- Reproducible and repeatable studies with quantifiable test methods and metrics;

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

Table 1. Schedule for the 2020 Test Methods and Metrics for Effective HRI in Real-World Human-Robot Teams Workshop

Time (ET)	Topic	Presenter
10:00	Welcome and Introduction	Dr. Jeremy Marvel (NIST)
10:10	Contributed Presentation	Miruna-Adriana Clinciu (Heriot-Watt University)
10:30	Contributed Presentation	Frank Förster (Queen Mary University)
10:50	Contributed Presentation	Kourosh Darvish (Istituto Italiano di Tecnologia)
11:10	Contributed Presentation	Rob Semmens (Naval Postgraduate School)
11:30	Invited Speaker	Dr. Yue (Sophie) Wang (Clemson University)
12:00	Break	
12:10	Contributed Presentation	Andrey Rudenko (Örebro University)
12:30	Contributed Presentation	Yigit Topoglu (Drexel University)
12:50	Contributed Presentation	David St-Onge (Ecole de Technologie Supérieure)
13:10	Contributed Presentation	Chittaranjan Swaminathan (Örebro University)
13:30	Closing Remarks	
13:40	Overflow Discussion	

- Human-robot collaboration and teaming test methods;
- Human dataset content transferability and traceability;
- HRI metrics (e.g., situation and cultural awareness);
- Human-machine interface metrics; and
- Industry-specific metrology requirements.

2. Invited Talks

Due to the switch to a virtual workshop, the number of invited talks was decreased to fit a shorter schedule. Dr. Yue (Sophie) Wang from Clemson University gave an invited talk at the virtual workshop.

Dr. Wang is the Warren H. Owen - Duke Energy Assistant Professor of Engineering and the Director of the I2R Laboratory in the Mechanical Engineering Department at Clemson University. Her research interests are in cooperative control and decision-making for human-robot collaboration systems, symbolic robot motion planning with a human-in-the-loop, cyber-physical systems, and multi-agent systems. Her research has been supported by NSF, AFOSR, AFRL, ARO, ARC, NASA EPSCoR, and Clemson University. Dr. Wang is a senior member of IEEE, and member of ASME and AIAA. She serves as the Chair of the IEEE Control System Society Technical Committee on Manufacturing Automation and Robotic Control. Her work has been featured in ASEE First Bell and State News.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

Dr. Wang's talk was titled ”Trust: A metric for human-robot collaboration”. She gave an overview of several projects ongoing at the I2R lab at Clemson, including collaborative assembly for manufacturing and autonomous driving. The topic of trust was discussed in context of two main types of applications where it is most important: collaborative tasks requiring both human and robot expertise, and dangerous or adversarial situations. For example, in the first situation, improper trust can result in disproportionate autonomy allocation and thus a higher mental workload on the human in addition to decreased task performance.

Several types of trust models were developed by Dr. Wang's lab to aid these scenarios. These fall into two broad categories: subjective trust for human-centered design and objective trust for a human-like, unbiased design. Some models developed by the I2R lab include the time-series trust model [1], the robot-to-human trust model [2], the mutual trust model [3], and the Dynamic Bayesian Network trust model [4]. Dr. Wang gave an overview of these trust models and their applications, as well as the metrics used to evaluate them. Details can be found in the referenced papers.

3. Abstracts of Accepted Presentations

Abstracts for the accepted presentations follow. Extended versions are available on the workshop website, https://hri-methods-metrics.github.io/.

Let's Evaluate Explanations!

Miruna-Adriana Clinciu, Heriot-Watt University, Edinburgh, UK

Transparency is an important factor for robots, autonomous systems, and artificial intelligence, if they are to be adopted into our lives and society at large. Explanations are one way to provide such transparency and natural language explanations are a clear and intuitive way to do this, helping users to understand what a robot or AI is doing and why. In this abstract, we highlight the importance of defining what makes a good explanation. Furthermore, we discuss evaluation methods for explanations by leveraging existing natural language generation evaluation metrics.

Towards Measuring Motor Resonance in Real-World HRI

Frank Förster, Queen Mary University of London and University of Hertfordshire, UK

Motor resonance (MR) has, in principle, considerable potential as a measure for assessing the quality of an ongoing interaction. However, the ways in which it is currently measured is impractical for applied scenarios, and none of the established measures can be calculated in or close to real time. We describe ongoing efforts to assess whether MR can be obtained in ecologically more plausible scenarios and discuss issues in need of clarification and the required methodological steps for moving towards real-time detection and measurement

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

Toward Common Metrics for Humanoid Robot Telexistence Evaluation

Kourosh Darvish, Istituto Italiano di Tecnologia, Genova, Italy

Telexistence technologies allow humans to sense that they exist at a distance regardless of time or space constraints; they make the human ubiquitous with interaction capability with the remote environment. Immersive real-time sensation with a humanoid robot telexistence system considers different technologies for perceiving the whole-body human motion or commands as well as providing feedback to the human from the robot. A fundamental element for such sensation leverages the intuitive and natural design of telexistence interfaces. Such a design allows higher user engagement, situation awareness, and lower user training, workload, and stress level; therefore enhancing the user experience. Common technologies that provide information regarding the user are motion capture systems, wearable sensors, RGB-D data, optical tracking systems, microphones, treadmill, and joypads. Concerning bilateral telexistence systems, intuitive interfaces such as various haptic devices, Virtual Reality (VR) headset, and speaker have been developed to provide force or tactile feedback, vision, and sound feedback to the human. In this manuscript, we aim at providing a brief overview of common methods to evaluate the telexistence systems and interfaces incorporated with humanoid robots. These methods will be used for evaluation of the humanoid robot, iCub, in telexistence scenarios.

Insights into Expertise and Tactics in Human-Robot Teaming

Rob Semmens, Naval Postgraduate School, Monterey, California

As robot capabilities rapidly evolve, the dynamics of human-robot teams will change. Autonomous, intelligent technologies will come to serve in roles that more closely resemble those of teammates, as opposed to tools. This will require humans to adapt and remain agile in developing novel strategies and tactics for employing these systems in complex, real-world scenarios. Building on previous work that presented a novel data set collected from teams of humans and robots playing capture the flag, the current research aims to identify measures capable of predicting successful teaming that lead to a positive, winning outcomes in the game. Video and text log analysis were used to describe gameplay and identify specific successful tactics. In conjunction with the experience levels of the participants, a number of measures of communication with autonomous robot teammates and robot efficiency were used to predict game performance. Only one metric was found to successfully predict game outcomes across all four games: level of robot involvement with offensive maneuvers. Several possible mechanisms for this observation are discussed, as well as multiple directions for future research directions leveraging this human-robot teaming platform.

Benchmarking Human Motion Prediction Methods

Andrey Rudenko, Örebro University, Sweden

Human motion plays a central role in human-robot interaction, especially for service robots,

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

personal assistants and human-robot teams. It includes motion trajectories in navigation experiments, hand reaching motions or full body poses for collaborative robotics action sequences, eye gaze directions, gestures, facial expressions, etc. Robots operating in close proximity to humans benefit from processing such motion cues using a model of human motion, and reason about the future to improve safety and efficiency of collaborations and service activities. Data plays a key role in building such models, as it is used for hyper-parameter estimation, motion patterns derivation or evaluation, and comparison of methods. There are several aspects for a good training and benchmarking dataset: accuracy of the provided ground truth, diversity of recorded scenarios, and provided additional cues are among the important ones. Following a rise of interest in human motion trajectories modeling and prediction, the insufficiency of existing datasets has triggered creation of new comprehensive benchmarks for outdoor and vehicle motion, driven by the progress in the automated driving domain. On the contrary, human motion recordings indoors and in pedestrian zones are still lagging behind. The drawbacks of commonly used datasets mainly come from two sources: (1) severe artifacts in ground truth estimation from noisy input (e.g., cameras or range scanners) and (2) recordings collected in non-challenging environments with homogeneous human motion. As an alternative to the traditional recordings of natural scenes, we propose a data collection procedure to generate a diverse and accurate human motion dataset in a controlled weakly scripted setup surpassing the limitations of existing datasets.

Integrating neuroimaging methods and neurohormonal measures for assessment and benchmarking of effective social bonding in HRI

Yigit Topoglu, Drexel University, Philadelphia, PA

One of the primary goals in the Human-Robot Interaction (HRI) field is to be able to design and develop robots that can engage effectively when interacting with humans. As the use of autonomous agents in everyday life has been growing, significant research has been focused on understanding the social bonding between the user and the agent. Social bonding increases the trustworthiness and reliability of the robot from the user's perspective, even enabling closer psychological and physical proximity. HRI research commonly employs the use of subjective and behavioral measures such as questionnaires and surveys to evaluate the user's degree of social bonding to the robot. While these measures are good at assessing the user's behavior after the moment of social affiliation, they lack the ability to capture the user's social bonding behavior in the moment of affiliation and can be dampened through the user's social inhibition. Neuroimaging and neurohormonal measures can provide unique neuroscientific information about the user's social bonding behavior at the moment of interaction, which can enrich the understanding of unique mechanisms behind the user's reactions. We argue that measuring these neural correlates alongside traditional behavioral and subjective measures can offer a complementary approach to current HRI methodology with the potential to capture information beyond traditional methods. We believe that future HRI theory development may benefit greatly from an integrated neuro-

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

scientific measurement approach.

For the assessment of effective social bonding in HRI, adding neurocognitive and neurobiological measures can provide a good comparison between social human-human and human-robot interactions. Neurocognitively, in human-human interaction, previous studies show that there is a relationship between increased electrical and hemodynamic brain activity and increased social bonding with the partner. Neurobiologically, some hormones such as oxytocin are related in regulating social behavior and social bond formation in partners regarding trust in human-human pairs. Using these methods together in HRI research similar to human-human interaction research together can enable the comparison between the person's behavior when affiliating with a human and with a robot. Understanding this comparison can indicate the factors that make human-robot bonding similar to human-human bonding.

Brain activity may also be useful as a social intent detector metric that can be used to gauge the social effects of human-robot interactions because brain areas regarding social cognitive processes activate based on the interaction. Therefore, monitoring brain activity can provide valuable information about the user's cognitive response in a social bonding scenario with a robot. Portable and wearable brain activity monitoring methods such as functional Near-Infrared Spectroscopy (fNIRS) and electroencephalogram (EEG) have improved significantly over the last decade in terms of hardware, software, and algorithms for effective mobile brain/body imaging. With recent advancements, these methods are now readily available for use in unencumbered, real-world dynamic environments such as monitoring the brain activity of participants walking outdoors and pilots in the cockpit flying an aircraft. While brain activity can explain the user's cognitive behavior, having only neurocognitive information may not be enough. By evaluating neurohormonal activity, we can examine the user's affective and social affiliative processes, willingness to trust and subconscious behavioral preferences. Examining both neurocognitive and neurohormonal measures jointly offer the potential to broaden the understanding of human bonding and social affiliation beyond using either approach individually.

In conclusion, considering the points mentioned above, examining neurocognitive and neurohormonal measures alongside traditional HRI behavioral and subjective measures can enhance the understanding of the user's social bonding behavior during interaction since these modalities provide valuable information about the neural reactions of the user to the robot. Grasping the user's neural mechanisms and rationale behind the user's social bonding behavior in the moment of affiliating with a robot can provide supportive information in determining the factors that make the interaction most effective, which is beneficial for social robot design. Including these metrics into future standards of HRI measurement will be essential for improving the description of mechanisms in theories of HRI.

Teaming with aerial systems: cognitive load assessment

David St-Onge, Ecole de Technologie Supérieure, Montreal, Canada

Multi-robot systems adoption is increasing for disaster response, industry, and transport

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

and logistics. Nevertheless, humans will remain indispensable to control and manage these fleets of robots, and particularly so in safety-critical applications. However, more sophisticated AI techniques create unintelligible robot control programs that are not necessarily human-centered. Furthermore, a human operators' cognitive capacities are challenged (and eventually exceeded) as the sizes of autonomous fleets grow. Our goal is to assess various means of measuring the operators cognitive load in an exploration task with six unmanned aerial vehicles (UAVs).

Quantitative Metrics for Execution-Based Evaluation of Human-Aware Global Motion Planning

Chittaranjan Srinivas Swaminathan, Örebro University, Sweden

How well a robot adheres to social norms is a crucial factor in the acceptance of robots. Making robots aware of motion patterns exhibited by people can be beneficial in this sense. Let us consider a robot operating in a tight corridor alongside people. A human-aware robot would plan paths along the flow of people in order to avoid forcing people to maneuver around the robot or vice versa.

Over the years several different representations have been developed to model dynamics in environments. For simplicity we call these Maps of Dynamics (MoDs). With respect to mapping the motion patterns of people, these methods can be broadly classified into three groups: trajectory mapping, velocity mapping, and mapping of spatial configuration changes.

Global motion planning is the phase of mobile robot motion planning that happens before the robot starts moving. It constitutes a general plan (i.e., a sequence of motions to execute) that a robot should use to reach the goal pose from its current pose. Thus, MoD may positively impact global motion planning through enabling the robot to plan paths that comply with the implicit traffic rules and are along less congested regions. This is expected to lead to less intrusive motions around people, therefore leading to less reactive replanning and also effectively reducing the 'freezing robot' behavior. In consequence, an MoD-aware planner is expected to lead to time-efficient motions: less time is wasted in reactive replanning and execution of motions. To the best of our knowledge, there are currently no studies of the actual effects of executing dynamics-aware motion plans in the literature.

In our previous work, we developed a motion planner that uses MoD information and showed that it can be tuned to follow or avoid observed flows, depending on application requirements. However, the evaluation scenarios were simplified to demonstrate advantages of the MoD-aware planner. We used a synthetic environment and only evaluated the generated paths themselves, and not the effects of executing the plans in actual flows of people.

We believe that in more complex scenarios, to assess the impact of MoDs on quality of paths, the traditional metrics (path length, curvature, roughens) are not sufficient and it will be necessary to test the actual execution of MoD-aware paths. It is non-trivial to assess

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345

whether a plan generated using one type of MoD is better than a plan generated using another. Recall that our motivation for employing MoDs is so that the robot can avoid extra maneuvering around approaching persons, and hence, save time. Therefore, we would like to know: will a robot save time if it employs MoDs in motion planning?

4. Documentation and Future Plans

Peer-reviewed, full-paper submissions by contributing authors will be submitted to a special issue of the Transactions on Human-Robot Interaction, tentatively scheduled for publication in March of 2021. Additionally, this workshop report will be made publicly available for the use of the research community.

This workshop is the second in a series of workshops leading toward formalized HRI performance standards. Early workshops are intended to target community and consensus building, and encourage the establishment of a culture of repeatable and reproducible, metrology-based research in HRI.

A third workshop is planned for the 2021 ACM/IEEE International Conference on Human-Robot Interaction, and will specifically address the action items from this year's workshop. Identified needs include:

- Further guidelines for reproducible and repeatable studies with quantifiable test methods and metrics;
- Human dataset creation and transferability of such content;
- A central repository for hosting such datasets as well as software tools for HRI; and
- Standards of practice for HRI, particularly for conducting human studies.

The IEEE Robotics and Automation Society (RAS) is hosting and supporting standardization efforts related to the last item. An IEEE Study Group for development of metrology standards for HRI has been initiated since the 2020 workshop. A meeting of the study group is scheduled to take place at the 2021 workshop and will gather input from the HRI community in creating a roadmap for future standards in the field.

References

[1] Saeidi H, Wang Y (2019) Incorporating trust and self-confidence analysis in the guidance and control of (semi)autonomous mobile robotic systems. *IEEE Robotics and Automation Letters* 4(2):239–246. https://doi.org/10.1109/LRA.2018.2886406

[2] Mizanoor Rahman SM, Liao Z, Jiang L, Wang Y (2016) A regret-based autonomy allocation scheme for human-robot shared vision systems in collaborative assembly in manufacturing. *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, , pp 897–902. https://doi.org/10.1109/COASE.2016.7743497

[3] Rahman SM, Wang Y (2018) Mutual trust-based subtask allocation for human–robot collaboration in flexible lightweight assembly in manufacturing. *Mechatronics* 54:94 – 109. https://doi.org/10.1016/j.mechatronics.2018.07.007. Available at http://www.sciencedirect.com/science/article/pii/S0957415818301211

[4] Wang Y, Humphrey LR, Liao Z, Zheng H (2018) Trust-based multi-robot symbolic motion planning with a human-in-the-loop. *ACM Trans Interact Intell Syst* 8(4). https://doi.org/10.1145/3213013

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8345