NISTIR 8417

Workshop Report: Novel and Emerging Test Methods and Metrics for Effective HRI, ACM/IEEE Conference on Human-Robot Interaction, 2021

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

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8417

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

NISTIR 8417

Workshop Report: Novel and Emerging Test Methods and Metrics for Effective HRI, ACM/IEEE Conference on Human-Robot Interaction, 2021

Shelly Bagchi, Jeremy A. Marvel, Megan Zimmerman,
Murat Aksu, Brian Antonishek, Xiang Li
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.8417

February 2022

U.S. Department of Commerce
Gina M. Raimondo, 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

National Institute of Standards and Technology
Interagency or Internal Report 8417
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8417, 15 pages (February 2022)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8417

Abstract

This report details the third annual, full-day workshop exploring the state-of-practice in 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 aims of shortening the lead time between the theory and applications of HRI, enabling reproducible studies, and accelerating the adoption of cutting-edge technologies as the industry state-of-practice. The workshop was held on March 12, 2021, as part of the virtual ACM/IEEE International Conference on Human-Robot Interaction.

This third installment of the annual workshop, 'Test Methods and Metrics for Effective HRI,' seeks to identify novel and emerging test methods and metrics for the holistic assessment and assurance of HRI performance. The focus is on identifying innovative metrics and test methods for the evaluation of HRI performance, and to advance the growth of the HRI community based on the principles of collaboration, data sharing, and repeatability. The goal of this workshop is to aid in the advancement of HRI technologies through the development of experimental design, test methods, and metrics for assessing interaction and interface designs.

Keywords

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

Table of Contents

List of Tables

Disclaimer

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.
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.

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. 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 necessity for validated metrology is driven by the desire for repeatable and consistent evaluations of HRI methodologies.

The full-day workshop continued to address the issues surrounding the development of novel and innovative test methods and metrics for evaluating HRI performance across the multitude of application domains, including medical, field, manufacturing, and social robotics. It built on the conclusions and takeaways from two previous workshops on Test Methods and Metrics for HRI, both of which were very successful and well-attended [1, 2].

The workshop was intended to push the boundaries of testing and evaluating HRI, and to establish benchmarks and standards for advancing the state of the art of HRI performance. A key focus was on inter-disciplinary collaboration and multi-domain applicability of test methods and metrics. Specific goals included the following:

- to develop and encourage the use of **consistent metrology** for HRI, producing quality datasets of pragmatic applications, and validating human subject studies for HRI;
- to explore novel and emerging **metrology tools** that have broad applicability across HRI domains;
- 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 virtually on March 12, 2021, as part of the annual ACM/IEEE International Conference on Human-Robot Interaction, and was well-attended, with a peak audience of about 40 members.

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 in industrial, collaborative, medical, and service robotics include:

- 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; and

Table 1. Schedule for the 2021 Test Methods and Metrics for Effective HRI Workshop

Time (ET)	Topic	Presenter
12:00	Introduction	Dr. Jeremy Marvel (NIST)
12:15	Invited Talk	Andra Keay (Silicon Valley Robotics)
13:00	Short Break	
13:15	Breakout Discussions	
14:00	Contributing Authors	Dr. Conor McGinn (Trinity College Dublin)
14:10		Dr. Robin Murphy (Texas A&M)
14:20		Junxian Wang (Monash University)
14:30	Long Break	
15:00	Keynote	Dr. Christoph Bartneck (University of Canterbury)
16:00	Short Break	
16:15	HRI Standards Group Intro	
17:00	End	

- researchers studying the social impacts and acceptance of human-robot teaming.

1.2 Schedule and Format

Table 1 contains the workshop schedule. The structure focused on the innovations in metrology for effective, real-world HRI, and featured invited speakers and short presentations of contributed extended abstracts. 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.

A follow-up meeting in the same week focused on ongoing standardization efforts for HRI metrology via a meeting of the IEEE Robotics and Automation Society's *Study Group on Metrology for HRI*. The meeting was highly attended, with many repeat attendees from the workshop. Following that meeting, applications for full working groups were submitted to IEEE and received approval; future workshops will integrate this standards effort into their content. For more information, please see IEEE P3107[1] and P3108[2].

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 do-

[1] https://standards.ieee.org/ieee/3107/10709/
[2] https://standards.ieee.org/ieee/3108/10710/

mains. Poster session participants were selected from contributors reporting late-breaking evaluations and their preliminary results.

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

2. Invited Talks

Due to the virtual format and time zone considerations, two invited talks were included to fit the shorter schedule. An opening talk was given by Andra Keay from Silicon Valley Robotics, and an afternoon keynote was given by Dr. Christoph Bartneck from the University of Canterbury, New Zealand. Links to talk recordings are available at the workshop website, https://hri-methods-metrics.github.io.

2.1 Andra Keay

Bio: Andra Keay is Managing Director and Founder of Silicon Valley Robotics (SVR), the leading non-profit robotics cluster, with more than 600 robotics startups and 50% of the global robotics investment activity. The process of commercializing research innovation into real world product is at the heart of Silicon Valley Robotics. Andra is an expert in aligning the multidisciplinary stakeholders required in a high tech innovation cluster and speaks regularly on the growing business of AI, robotics, and the ethical issues that emerge. Silicon Valley Robotics is a not-for-profit association of robotics companies with the mission of supporting innovation and commercialization of robotics technologies. Silicon Valley Robotics partners with many global organizations to create a blueprint for emerging technology innovation, to provide a landing pad for visiting companies and to make connections with robotics and AI startups and investors.

In her talk, Andra laid out the progression of robotics companies in Silicon Valley, from 60 companies in 2010 to 600+ in 2020. There was also a significant increase in funding in the domain - about 55 times higher over the ten years. This makes SVR about three times bigger than any other comparable organization internationally. Andra discussed

her experience bringing together robotics stakeholders with collaborators in related Silicon Valley industries and those in relevant application domains. She also highlighted areas of potential robotic expansion in remote work, which is of large interest during the pandemic; these areas included interacting with materials and machinery, and controlling mechanical equipment. Although there is more work to be done to make that a reality, bringing these areas to the attention of industry is one of Andra's goals.

Andra also spoke about the "Robotics 2.0" era we are presently transitioning to. Unfortunately, many larger industries and organizations are lagging behind and still only function in the Robotics 1.0 space: Dull, Dangerous, Dirty, and Dumb applications, where robotics are caged and used away from humans. Meanwhile, Robotics 2.0 focuses on four new factors: Smarter, Safer, Sensors, and Single-tasks. Silicon Valley startups are focusing on these new areas, where robots and humans interact, whether in industrial collaboration, social robotics, service robotics, etc. These areas are not yet profitable, but can be tracked by the rising investments in the field. Due to the massive growth in these areas, Andra emphasized the need for current metrics and standards to catch up to these new applications. Reviewing trends in the current space, the decreasing cost of hardware, high use of the Robot Operating System (ROS), and transition to new interface technologies (e.g., phones and tablets) were mentioned. Finally, for the future Robotics 3.0 era, Andra envisions systems moving from single-task to multitasking. She proposes 4Ms: Multitasking, Emotive, Morphing, and Multi-agent systems. By using new research grants as a metric, Andra has seen interest in these new technologies increasing over recent years.

The industry insights Andra provided were a different viewpoint for many of our audience members from academia. Her conclusions highlighted the importance of not only new research, but the need for standardization and cohesive metrics on safety and reliability before robotics companies can achieve more widespread success and adoption.

A follow-up question after the talk asked: What can the robotics community do to increase the public's trust in new products across emerging application domains? One suggestion Andra offered was that robot platforms should be individually identifiable to increase traceability, for example using registration numbers in addition to company branding, just as vehicles and aircrafts do. In addition, a robot registry was mentioned to make it easier to find information on a robot's oversight, capabilities, software, etc. to give some transparency. Finally, the introduction of local tech councils was mentioned, to give some regional control over the use of robotics in the area, e.g., delivery or service models in public places.

2.2 Christoph Bartneck

Bio: Dr. Christoph Bartneck is an associate professor and director of postgraduate studies at the HIT Lab NZ of the University of Canterbury. He has a background in Industrial Design and Human-Computer Interaction, and his projects and studies have been published in leading journals, newspapers, and conferences. His interests lie in the fields of Human-Computer Interaction, Science and Technology Studies, and Visual Design. More

specifically, he focuses on the effect of anthropomorphism on human-robot interaction. As secondary research interests, he works on bibliometric analyses, agent based social simulations, and the critical review of scientific processes and policies. In the field of Design Christoph investigates the history of product design, tessellations, and photography.

Dr. Bartneck was asked to present the workshop keynote due to his work on the seminal Godspeed Questionnaire Series [3]. He gave a talk entitled "There is method to the madness: Research methodology in HRI". He discussed how he developed the Godspeed Questionnaire due to the lack of a survey for capturing human perception of a robot. The need for this survey was shown by how the community began to use Godspeed in new research, and the original paper has had close to 2,000 citations since 2009. The survey has been used internationally and also been translated into 15+ languages.

Dr. Bartneck framed his talk by dividing HRI studies into two camps: those trying to solve a problem vs. those trying to understand the world. He related this back to differences between scientists, engineers, and designers, all of whom collaborate within HRI. He outlined some of the issues he sees within HRI engineering, the main premise of which is that the success criteria is within the human partner, and often relies on instinctive reactions that are difficult to distill into scientific comparisons. Additionally, results often are numbers that have no context to evaluate them by, because the study may have used customized metrics that have no existing results for comparison. Finally, the reluctance of the community to undertake replication studies was mentioned as a barrier to progress.

Some solutions suggested during the talk included limiting the complexity of the variables under study. Due to the innate complexity of human studies, Dr. Bartneck cautioned against going beyond a 2x2 study, as valid conclusions are very difficult to come to past that point. Another rule of thumb he emphasized was to write the introduction and methods section of your paper before running the study. This helps nail down the methodology before having to interact with participants, and helps pre-register the study so that you are not tweaking results to meet your desired conclusions in the future (e.g. 'p-hacking'[3]). In addition, Dr. Bartneck encouraged publishing of negative results as a benefit to the community.

The next section of the talk covered aspects of study design, such as experiments in the lab vs. in the wild, or the use of qualitative vs. quantitative metrics. On the second point, Dr. Bartneck was very emphatic about the strength of conclusions made from quantitative data, as it is much higher in statistical validity than that of qualitative metrics. Autonomous vs. Wizard-of-Oz (WoZ) robot control was the next decision discussed [11]. Unfortunately, due to the complexities of robot technologies, Dr. Bartneck acknowledged that it is often impractical if not impossible to use a fully autonomous robot. However, some transparency about the WoZ process is needed to keep the public's impression of robot capabilities in check. Other topics mentioned included physical robots vs. virtual avatars, and crowdsourced human studies. Ultimately, the conclusion was that touch is an essential part of a robot's interaction with humans and the world, and it is hard to simulate or imagine accurately.

[3] https://scienceinthenewsroom.org/resources/statistical-p-hacking-explained/

The last section of the talk focused on statistical significance of results and the 'p-value problem', where results tend to become significant as the sample size grows. Dr. Bartneck emphasized considering whether experimental results are scientifically interesting, rather than merely numerically significant, and recommended reading a 1995 text by Chatfield, *Problem Solving: A Statistician's Guide* [4]. Some extreme examples of 'p-hacking' were given where far-fetched correlations were made with statistical significance, while ultimately the research question did not make sense logically or practically. On the other hand, Dr. Bartneck also referenced the research trend he calls the 'file drawer effect': where negative results or unsuccessful studies are not generally accepted for publication, but shelved away and not disseminated. This results in bias among published studies, as well as the types of research being attempted, which is also linked to the current replication crisis in HRI research. Although this is a difficult problem to overcome without systemic change from the publication venues and peer-review process, Dr. Bartneck encouraged the audience to keep these topics in consideration and keep submitting these types of research as views change.

Dr. Bartneck's talk was very well-attended and generated a very interesting discussion, which can be found in the workshop recordings. We appreciate his contribution and the unique viewpoints he presented to the workshop audience.

3. Abstracts of Accepted Presentations

Abbreviated abstracts for the accepted presentations follow; these are contributions from external authors and are included verbatim. Extended versions are available on the workshop website, https://hri-methods-metrics.github.io.

Towards the Development of Test Methods for Collaborative Cleaning with a UV Robot

Conor McGinn (Trinity College Dublin)

The use of ultraviolet germicidal irradiation (UVGI) as a means of room sterilization has been increasing in recent years. To protect staff and patients from exposure to potentially harmful UV rays, it has become standard practice to evacuate rooms prior to use. The aim of this work was to explore how a UV robot might be deployed safely alongside people. For UV robots to operate with people in the room, it is required that appropriate personal protective equipment (PPE) is worn, the UV output of the device is regulated, and tests are performed to verify that background radiation levels do not exceed occupational safety thresholds. In this paper, we outline three conditions that must be met for safe collaborative cleaning with a UV robot. In doing this, we provide a framework for performing a new type of robot-enabled room disinfection.

Conducting Human-Robot Integration Research during Three Real World Disasters

Camille Peres, Ranjana Mehta, and Robin R. Murphy (Texas A&M University)

Unmanned ground, aerial, and marine robots have been used at disasters, public safety incidents, and other non-routine (off-normal) events since 2001, when they were used during the response to the 9/11 World Trade Center disaster. One of the major benefits of using unmanned robotic systems is inherently that the human does not have to be in the dangerous areas of the disaster, however, it does not remove the need for the human. Therefore it is useful to have human-robot interaction data from actual disasters. However, it is difficult to get access to collect research in the field with the actual responders during an event; and, once embedded with responders, there are few established protocols for collecting quantitative data as part of a disaster response with robots. As part of an effort to identify the unique needs and risks that disaster robot operators experience, the Center for Robot-Assisted Search And Rescue (CRASAR) collaborated with researchers from Occupational Health and Industrial Engineering to imbed participant observers and have them document the experience of small unmanned aerial systems (sUAS) pilots during disaster response for three disasters: Hurricane Harvey in 2017, the Kilauea volcanic eruption in 2018, and Hurricane Michael in 2018.

This paper documents those events and specifically our efforts to collect quantitative data on sources of pilots stress and fatigue; assess fatigue; and identify how fatigue changes over the course of a mission for sUAS pilots and the response teams. We describe lessons learned regarding methods for how to best collect data in this dynamic field setting. We specifically provide descriptive information for each of the three events regarding the response team and the unmanned vehicles they were using; attributes of each disaster; methods used during that disaster; and adaptations made to the methods based on the pilots' and researchers' experiences during the disaster. We conclude with recommendations based on the experiences of the researchers and sUAS pilots.

Metrics for Evaluating Social Conformity of Crowd Navigation Algorithms

Junxian Wang, Wesley P. Chan, Pamela Carreno-Medrano, Akansel Cosgun, and Elizabeth Croft (Monash University)

Autonomous robot navigation in populated environments has been an active research field in the past decade [8, 9]. A key challenge is to design algorithms that allow robots to navigate safely and socially in such environments. Recently, advancements in computation hardware and machine learning algorithms have enabled a series of deep learning based methods to emerge [5–7, 10]. These works have diversified definitions of what constitutes "social behavior", as well as evaluation criteria for social conformity of the resulting robot navigation behavior. Given this lack of consistency in social evaluation metrics used for crowd navigation algorithms, it is not surprising to find that publications that aim to provide socially conforming crowd navigation algorithms usually present evaluations that are mainly focused on efficiency, and lack details on the social conformity of the trained results.

Furthermore, the lack of a set of well defined standardized metrics make it difficult, if not impossible, to compare performances of algorithms published in different works. To fill this gap, this work proposes a set of metrics intended to be used for evaluating and comparing different crowd navigation algorithms from a social conformity aspect. The proposed metrics are applied to a collection of state-of-the-art crowd navigation algorithms.

4. Documentation and Future Plans

Contributing authors were encouraged to provide full-paper submissions to a special issue of the Transactions on Human-Robot Interaction, tentatively scheduled for publication in March of 2022. Additionally, this workshop report will be made publicly available for the use of the research community.

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

A fourth workshop is planned for the 2022 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. Two IEEE Working Groups for development of metrology standards for HRI have been initiated. These groups will pick up from the previous Study Group and will hold initial meetings in late 2021.

References

[1] Shelly Bagchi, Murat Aksu, Megan Zimmerman, Jeremy Marvel, Brian Antonishek, Heni Ben Amor, Terry Fong, Ross Mead, and Yue Wang. Workshop report: Test methods and metrics for effective HRI in collaborative human-robot teams, 2019. *NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology*, 2020.

[2] Shelly Bagchi, Jeremy Marvel, Megan Zimmerman, Murat Aksu, Brian Antonishek, Heni Ben Amor, Terry Fong, Ross Mead, and Yue Wang. Workshop report: Test

methods and metrics for effective HRI in real-world human-robot teams, 2020 (virtual). *NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology*, 2021.

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[4] Chris Chatfield. *Problem solving: a statistician's guide*. CRC Press, 1995.

[5] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6015–6022. IEEE, 2019.

[6] Changan Chen, Sha Hu, Payam Nikdel, Greg Mori, and Manolis Savva. Relational graph learning for crowd navigation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10007–10013. IEEE, 2020.

[7] Yu Fan Chen, Miao Liu, Michael Everett, and Jonathan P How. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 285–292. IEEE, 2017.

[8] Akansel Cosgun, Emrah Akin Sisbot, and Henrik Iskov Christensen. Anticipatory robot path planning in human environments. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 562–569. IEEE, 2016.

[9] Yuqing Du, Nicholas J Hetherington, Chu Lip Oon, Wesley P Chan, Camilo Perez Quintero, Elizabeth Croft, and HF Machiel Van der Loos. Group surfing: A pedestrian-based approach to sidewalk robot navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6518–6524. IEEE, 2019.

[10] Michael Everett, Yu Fan Chen, and Jonathan P How. Motion planning among dynamic, decision-making agents with deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3052–3059. IEEE, 2018.

[11] Laurel D Riek. Wizard of oz studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1):119–136, 2012.