Towards methodological principles for user studies in Human-Robot Interaction

Kathleen Belhassein, Guilhem Buisan, Aurélie Clodic, Rachid Alami

To cite this version:

Kathleen Belhassein, Guilhem Buisan, Aurélie Clodic, Rachid Alami. Towards methodological principles for user studies in Human-Robot Interaction. Test Methods and Metrics for Effective HRI in Collaborative Human-Robot Teams Workshop, ACM/IEEE International Conference on Human-Robot Interaction, Mar 2019, Daegu, South Korea. hal-02282600

HAL Id: hal-02282600

https://hal.laas.fr/hal-02282600

Submitted on 10 Sep 2019

Towards methodological principles for user studies in Human-Robot Interaction

Belhassein Kathleen
CNRS, LAAS, CLLE
Univ. de Toulouse, LAAS
Toulouse, France
kathleen.belhassein@laas.fr

Buisan Guilhem
CNRS, LAAS
Univ. de Toulouse, INSA, LAAS
Toulouse, France
guilhem.buisan@laas.fr

Clodic Aurélie
CNRS, LAAS
Univ. de Toulouse, LAAS
Toulouse, France
aurelie.clodic@laas.fr

Alami Rachid
CNRS, LAAS
Univ. de Toulouse, LAAS
Toulouse, France
rachid.alami@laas.fr

Abstract—**Robots are becoming more and more present in our everyday life at work or for entertainment. This raised the need for a principled and well-understood interaction between humans and robots which led to the creation of the Human-Robot Interaction (HRI) research field. In order to assess the models, designs and algorithms elaborated in these researches, evaluation techniques have to be employed in the context of user studies. In this work, we aim to adapt from other research fields some principles that might be used in HRI user studies. More precisely we discuss around some frequent issues concerning recruited users, evaluation methods and replication of the studies, and how some methodological practices could circumvent them. We will finally discuss more generally on HRI studies, claiming that they need methods and assessment techniques specific to their particularities.**

Index Terms—**human-robot interaction, evaluation methods, user studies**

I. Introduction

By the emergence of social robots, Human-Robot Interaction (HRI) is a field in full expansion and can be defined as the study of ”understanding, designing, and evaluating robotic systems for use by or with humans [1]”. HRI has been created by the meeting between humans and robotic systems, as people are increasingly exposed to robots in their daily lives (toys, aerial robots, or assistive and service robots). Therefore, it is necessary to consider the human and its particularities for the design of robots and their evaluation. Thus, this field is at the border between several disciplines among which robotics and social sciences. This collaboration in which roboticists, engineers, psychologists and ergonomists work together in order to devise acceptable and useful robots for human partners raises important issues regarding experimental and evaluation methods. Indeed, the various fields involved in HRI studies do not meet the same methodologies, requirement and limitations. This paper aims to identify some of the important issues faced by HRI studies, and to account for the complexity and specificity of this interdisciplinary field. Specifically, we will focus on HRI assessment methods, due to the actual need to build solid and reliable tools to evaluate these robotic systems.

In the sequel, we will first present some particular aspects of users in HRI that must be considered when designing a user study in Sec II. Then we will focus more on methodological issues from evaluation methods used in HRI in Sec. III. Sec. IV will present the replication crisis that happens in fields like psychology or HCI and that is even more critical in HRI, and propose potential solutions. Finally, in Sec. V, we will take all these methodological issues to provide some guidelines, in order to make HRI user studies more rigorous while considering the particular limits of this disciplinary field.

II. Users in HRI studies

When conducting an user study, we obviously recruit today’s users, with their past experiences and expectations with robotics. Kuhnert et al. showed the existence of a *gap* between the user’s attitude towards existing robots and their expectations about the ideal everyday social robot [2]. Thus, when evaluating a human robot interaction it is important to evaluate the three aspects defined by Desmet and Hekkert: the instrumental interaction (interaction for expected purpose: always evaluated in current user study), the non-instrumental interaction (interacting for other than main purpose: often non evaluated) and the non-physical interaction (expectations: often under evaluated) [3]. Indeed, non-physical interaction refers to all preconceptions of the robot by the user, they could come from past experiences or imagination. Since today’s users have almost no past experiences with robot, those anticipations come mainly from imagination and fantasies and can widely vary from one to another. Thus, it is really important to evaluate the mindset of the user before the study. Some tools used in psychology can be useful in this context. For example, the implicit association test [4], used to measure automatic and implicit associations like prejudices, or priming paradigms could be used to control the expectations and beliefs of subjects towards robots.

Moreover, as many of the recruited users has none to very few past experiences with robots, the novelty (and wouaw) effect when interacting with robot during a one shot study is huge, and may not be representative of a robot long term use. To assess this *cumulative experience* [5], User Experience (UX) and Human-Computer Interaction (HCI) designers conduct *longitudinal studies* [6], gathering data on a long period of time. However, in the human-robot interaction field, this study may not be applicable as is. Indeed, the used material (robots)

This research was supported by the French National Research Agency (ANR) grant reference ANR-16-CE33-0017 (JointAction4HRI).

is expensive and often in limited quantities in laboratories. In order to diminish this effect, the evaluation should be part of a larger, more cognitively intense or time pressurizing task where the user must almost *forget* about everything concerning the robot except the part to be evaluated, which should be crucial in this task. This effect could also be reduced by making an habituation phase at the beginning of the experiment, in which the user can act more freely with the robot.

In some HRI studies, a questionnaire is administered before the interaction task in order to apprehend the degree of familiarity and knowledge of the participants about the robots. Even if some recruited people are from the same professional environment and are therefore already accustomed to robots, this measure of knowledge about robots is never used to remove participants from the user study. In addition to the bias that this recruitment may pose, the sample is therefore not representative of the general population but of a particular subpopulation. To ensure a better representation of the population, it would therefore be necessary to randomly sample and recruit outside of the professional circle, as in [7] or [8].

Finally, HRI studies have frequently few participants. For example, about 44% of user studies published in the proceedings of the conference HRI'17 involve fewer than 30 participants. However, the size of the sample is a prerequisite for obtaining sufficient statistical power to conclude on the results obtained, and to avoid type I errors (false positive) or type II (false negative). Beyond the statistical issues, it seems important to be modest about the conclusions drawn from studies involving too few participants, and therefore not to generalize to the population the results obtained.

The particular case of web studies

To respond to this difficulty and have a large number of participants, it is now frequent to be confronted to web studies (e.g. [9]), especially via the crowdsourcing web platform Amazon Mechanical Turk. Indeed, it is then much easier to recruit participants and have them take small tasks or questionnaires directly online. It seems however important to be vigilant on the study conclusions, since we are not in a context of human-robot interaction in its strictest sense; the participant is not confronted to the robot, does not share its physical space, and therefore will not have the same reactions that he would have during an interaction in the real world. Nevertheless, it could be an interesting tool for participatory design studies, i.e. all the studies that do not deal with a fully implemented robot that must be evaluated but rather that serve to explore the responses of users to certain specific behaviors and to collect their opinion.

III. Evaluation methods

The key concept in HCI is usability [10], which regroups effectiveness (ability to perform a task), efficiency (ability to perform the task without wasting resources) and satisfaction. More often than not, HRI user studies focus on user satisfaction (and *acceptability*) evaluated with questionnaires created or adapted specifically for this context of interaction [11] [9] [12] or borrowed from HCI [13]. Even if HRI studies join the field of Human-Computer Interaction and user experience design by the fact that they are both trying to improve the human use of interactive systems, it is a hugely different experience to interact either with a robot or a computer. We must not forget that in the case of HRI studies, we are talking about two agents that interact and no longer an agent that interacts with a product/an interface. Therefore, the methods and tools of HCI are not always suitable to be used during a situation of interaction between a human agent and a robotic agent [14]. First of all, people tend to attribute mental states and human traits to robots. This human tendency to anthropomorphism is not only dedicated to robots but also animals, objects or natural phenomenon. However, robots are more perceived as an agent endowed with likelike qualities than other technologies; for example, studies have shown physiological and brain responses of subjects to acts of violence perpetrated on robots, showing that we also feel empathy for robots [15]. In addition, the perceived risks to evolve in the same environment as a robot are obviously not the same than when we use computers or other technologies, and can induce negative emotions or feeling of insecurity [16]. Thus, these particularities in the subjective experience of interacting with a robot has to be considered in the build of experimental and evaluation tools.

A. Use of self-assessment methods

Although the simplest and most widespread evaluation method in HRI studies is the self-assessment method with questionnaires, it is necessary to understand that there may be a significant difference between what subjects self-report of their own experience and what they really experienced and felt, for example because of the social desirability bias [17]. In addition, there are very few questionnaires created for HRI studies that meet the validity and standardization criteria of these methods. Indeed to be validated, a questionnaire must be reliable and consistent, i.e. the results must be replicate in comparable situations; it must be valid, that is to say it actually measures what it is supposed to and not another dimension; and finally, it must be sensitive to change. More often than not, HRI questionnaires are simply evaluated on their internal consistency (all the items from one dimension are correlated to each other) with Cronbach's alpha. However, to valid that the questionnaire measures the construct of interest, it would be important to use on a pilot study another method of evaluation for this same construct and use correlation matrices between them [18].

Finally, it is common to see questionnaires used in another language. However, if a questionnaire is created in a specific language it has to be used only in that language; as specified by [11] when the Godspeed questionnaire series were developed, only native speakers can understand the meaning of an item and its translation can easily modify it. It is therefore absolutely necessary when using a questionnaire from another language to use the back translation method (i.e. translate back into the original language the questionnaire previously translated into the target language). In addition, the questionnaire once translated must imperatively be validated following the same rules as an original questionnaire, to ensure that the translated version measure the exact same construct at the original, as it was the case for example for the French translation of the UX questionnaire AttrakDiff [19].

B. Other evaluation methods for acceptability

Including heart rates, brain or skeletal muscles electrical activity, blood pressure, respiratory frequency or even galvanic skin responses, there are a number of different physiological measures that can be used to evaluate the participant's physiological response in touch with a robot. These evaluation methods have the advantage of being able to prevent participants from consciously modifying their answers, as this is the case with self-assessment measures. [20] used skin conductance response, heart rate and corrugator muscle activity measured via surface electromyogram to evaluate the human responses to several motions of a CRS A460 robot in addition to self-assessment measure. The results reveal a strong arousal response when the robot used fast motions, illustrating that physiological measures can be useful indicators of the mental state of participants when they interact with a robot. However, the use of these techniques must always be coupled with at least one other evaluation method as suggested by [21], in order to avoid erreoneous interpretations and causal relationships.

In interaction situations and even more in cooperative situations, taking social signals into account may also be useful for assessing the human ability to accept to engage into a task with a robotic partner. Behavior observation techniques are for example used to measure shared gazes, in particular with video recording and eye tracking [22].

C. Evaluate efficiency and effectiveness in human-robot interaction

But what is the point of making a robot behavior satisfying but useless? Once more, tools exist in HCI to measure the other components of usability but are not always adapted to HRI studies. For example, freeze-probe techniques are frequently used in the case of evaluating the situation awareness (i.e. the perception, comprehension and projection of elements in the environment [23]). They consist of freezing the task in progress and administrating a questionnaire about the situation at the exact moment of the freeze point [24]. Such a device, mostly developed for use in the aviation or military domain, is a real challenge and almost impossible to apply in HRI, since we are in the case of a real-world physical interaction with a robotic agent and we can not just make the robot disappear.

Most of the existing HRI questionnaires deal more with the physical aspect of the robot and its acceptability by the user [11] [9] than its effectiveness and usefulness. In previous work, we tried to propose a preliminary version of a questionnaire specific to HRI studies and measuring the decision-making processes of the robot in a joint action context with a human [25], but this tool remains at the draft stage and deserves to be refined, coupled with other evaluation methods, used in other studies and finally validated according to the criteria of self-assessment methods.

Moreover, adding efficiency/effectiveness measurements in a study can be pretty simple given the technical abilities inherent to robots (e.g. timing the task completion, the number of user errors). Thus, we could consider using the robot as a tool for measuring its impact on the human performance, and therefore evaluate the efficiency of human-robot interaction. In addition, one of the most widely used methods of cognitive psychology is mental chronometry, which uses reaction times as a measure, and refers to the temporal study of information processing [26]. Reaction times are commonly used learning, imitation, perceptual interference, or attention, and are also used in some HRI studies [27] [28].

Behavioral measures, as described above, can also be adapted for evaluating how the behavior of the robotic agent can improve the human performance [29].

Whether it concerns the evaluation of acceptability or efficiency and effectiveness in human-robot interactions, a more frequent use of objective measures (e.g. physiological or task performance measures) could improve the validity of the results of HRI studies and their methodological rigor. Bethel & Murphy also consider that the use of a single method of evaluation is not sufficient, and recommend using three of them, including at least one objective method [30]. These different techniques must still be consistent with each other in their use in a user study, this recommendation is not always feasible but should be taken into account when establishing the methodological plan.

IV. The replication crisis in HRI

Replication of results is an important concept for any discipline following the scientific approach; it is a question of repeating a study to determine if the results are reproducible and therefore reliable. Psychology and more generally social and medical sciences have known since the 2000s what is called the replication crisis: according to [31] involving 1500 researchers, more than half of them have failed to reproduce the results of their own studies. This phenomenon seems to be also very important in the field of HRI. Indeed, if the replication crisis begins to appear in HCI [32] mainly because of closed source code used in the experiments, HRI presents other issues. First, the code used in robotic studies is often much more complex and heavy to use. Indeed, to evaluate a high-level component (e.g. a task co-planning algorithm) a whole component stack is needed (e.g. low-level motor controllers, path finding, trajectory following, motion planning, localization, face detection, speech synthesis, speech recognition). If some of those components are standards and widely available, others can be state of the art, unstable or even tweaked by the experimenters to match their own needs, making it more difficult to replicate the experiment if those components are not precisely described or not available.

Moreover, the material used can also be changed (e.g. by 3D printing a gripper to fit the object manipulation task needed), and components tuned to work accordingly. Those small changes on the hardware and on the software needs to be reported, else the experiment replication is impossible.

Finally, many user studies in HRI used Wizard of Oz technique, because robotic systems are often not robust enough to act autonomously in an environment with humans [33]. The use of WoZ technique makes very difficult the exact replication of the situation due to the fact that the complete scenario of the study is not available. In the case of HRI evaluation, we can also ask ourselves if humans evaluate the robot or the human controlling the robot.

V. Discussion and Conclusions

A. Recommendations for designing an user study

By taking all those issues and solutions coming from different fields we might provide some checkpoints when designing a HRI user study:

1) The *more users* the better. It improves the statistic analysis of the study and could erase some bias.
2) *Widen* the recruitment. Your colleagues know a lot about technology: randomly recruit people in your bakery, in the supermarket, using flyers...
3) *Be rigorous* with your protocol. There are so many unwanted uncontrolled variables. Don't be one!
4) Let the user *accustom* to the robot and its behavior before starting the experiment. In order to diminish the "wouaw" effect and make measures closer to a long term use.
5) Make sure your experiment is physically and psychologically *safe*. Apart from hurting an user, you risk to prevaricate your measures if the experimenter is stressed about something going wrong.
6) Objectively *measure if your robot is useful* in what you are making it do. A robot can be really satisfying, but it will be quickly forgotten if it does nothing.
7) Use the *right tools* for your measurements. Widely used and standardized tools will give more credibility to your study. Questionnaires are not the end.
8) Make theoretically solid and valid *tools* specific to HRI and publish them if they don't exist. It will benefit the whole community.
9) *Give as much details as possible* about your study. Gives the source code, hardware schematics or references and the description of the environment in order to make others able to reproduce your experiment.
10) When doing a *Wizard of Oz be rigorous*. Emulate only a small, non-evaluated, component of your robot. Write the rules you follow down, respect them and publish them along with your study.

B. General thoughts on HRI

Doing a user study is only one part of a HRI researcher work. Indeed, as stated before, the HRI community is at the crossroads of several other disciplines [34], but presents its own issues. Psychology focuses more on finding principles underlying human behavior than on designing agent behavior. UX designers are more inclined to make products satisfying and enjoyable than increasing their usefulness in everyday life. HCI designers rarely deal with autonomous, moving agents capable of physically act upon their surroundings. Roboticists aims more at designing robots replacing humans in tedious, repetitive or dangerous tasks than to make robots work in concert with humans. This is why we think HRI must be considered as an unique field with its own challenges and habits.

1) Of the importance of the user study: We have seen too many papers refused in conferences for not presenting an user study, even when the paper stated clearly that it was a technical contribution. Besides, we have seen too many papers accepted despite having a really poor user study. To solve this, we propose to mark the distinction between technical proposition papers, describing algorithms and computing methods for HRI and evaluation papers, presenting user studies. On one hand, this will allow authors to develop more both technical aspects and evaluation methodology, solving one problem of the replication crisis. On the other hand, the reviews will be easier and more homogeneous. For example, when reviewing a paper containing an astonishing, theoretically solid, new co-planning algorithm but evaluating it with a poorly designed user study, a roboticist according more attention to the algorithm would positively review this paper, whereas a psychologist would review it poorly because of the user study.

2) Be modest and critical: Too many papers try to conclude general HRI principles from small scale one shot evaluations. A nice example is about Hall proxemics theory [35]. The proxemics theory has been made for explaining humans relations to space and positioning when interacting with each other. This theory presents several limits, the distances depends largely on the cultural influences [36], age [37] and on the context (e.g. crowded environments). But several papers in HRI tended to implement this concept as is and drew general conclusions from it (e.g. a robot should not come closer than 1.2 meters to a human), totally disregarding the type of the robot and the context. Does a Cozmo robot also need to stay at 1.2 meters? What about people making Cozmo roll on their arms? Does a PR2 have to move backward when crossing a human in a corridor narrower than 1.2 meters? Thus, we recommend to stay modest when publishing results and conclusions. Making an effort to control bias and keeping the context clear is crucial for valid researches since effects can come from many, uncontrolled causes. We also recommend reviewers and readers to be critical about presented results, and to beware of the punchlines and global, stated conclusions.

3) See HRI as a special distinct field with its specific methods and limits: As HRI has its own challenges that cannot be tackled using results found in other research fields, we think HRI must have its own approaches. To do so, researchers have to set up *standards*. Those standards should be methods allowing researchers to know what you can and cannot do when submitting work to HRI community. By doing so, researchers would also draw limits to the discipline. Is an user study of the ergonomic of a joystick controlling the speed of a rover a part of HRI? Does a human aware task planner assuming a total controllability of the human have a part in HRI? Researchers also need to write down a set of widely accepted results. Those results must be theoretically funded and validated by multiple experiments. We also propose the creation of a HRI dedicated ethical committee where usual constituting members (e.g. psychologists) should be supported by safety critical and embedded software engineers and researchers. Such a committee will be able to treat both the psychological risks (usual in psychology experiments) and the physical risks due to the presence of an autonomous agent potentially capable of injuring a participant.

By analyzing today's issues in HRI user studies and the corresponding challenges in other disciplines and how they tackle them, we tried to draw some guidelines in order to make better user studies in HRI. However, as the HRI field is so large, the user studies that can be done with robots can be really different from one another and we are aware that those guidelines might not be strictly respected for all of them. We still think that at least knowing what is done in other fields and having in mind those principles will help researchers to conduct more reliable and significant user studies.

References

[1] M. A. Goodrich and A. C. Schultz, “Human-robot interaction: A survey,” *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2008.

[2] B. Kuhnert, M. Ragni, and F. Lindner, “The gap between human's attitude towards robots in general and human's expectation of an ideal everyday life robot,” in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 1102–1107.

[3] P. Desmet and P. Hekkert, “Framework of product experience,” *International Journal of Design*, vol. 1, no. 1, pp. 13–23, 2007.

[4] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, “Measuring individual differences in implicit cognition: The implicit association test.” *Journal of Personality and Social Psychology*, vol. 74, no. 6, pp. 1464–1480, 1998.

[5] V. Roto, J. Hoonhout, E. Law, and J. Hoonhout, “User experience white paper,” 2011.

[6] J. Lazar, *Research methods in human computer interaction*, 2nd ed. Elsevier, 2017.

[7] N. Sadoughi, A. Pereira, R. Jain, I. Leite, and J. F. Lehman, “Creating prosodic synchrony for a robot co-player in a speech-controlled game for children,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. ACM Press, 2017, pp. 91–99.

[8] C. Kidd and C. Breazeal, “Robots at home: Understanding long-term human-robot interaction,” in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 3230–3235.

[9] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, “The robotic social attributes scale (RoSAS): Development and validation,” in *2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI17)*. ACM Press, 2017, pp. 254–262.

[10] J. Grudin, “Utility and usability: research issues and development contexts,” *Interacting with Computers*, vol. 4, no. 2, pp. 209–217, 1992.

[11] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likability, perceived safety of robots.” *International Journal of Social Robotics*, vol. 1, pp. 71–81, 2009.

[12] M. Heerink, B. Krose, V. Evers, and B. Wielinga, “Measuring acceptance of an assistive social robot: a suggested toolkit,” in *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009, pp. 528–533.

[13] M. Hassenzahl, M. Burmester, and F. Koller, “AttrakDiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualitt,” in *Mensch & Computer 2003*, G. Szwillus and J. Ziegler, Eds. Vieweg+Teubner Verlag, 2003, vol. 57, pp. 187–196.

[14] J. E. Young, J. Sung, A. Voida, E. Sharlin, T. Igarashi, H. I. Christensen, and R. E. Grinter, “Evaluating human-robot interaction: Focusing on the holistic interaction experience,” *International Journal of Social Robotics*, vol. 3, no. 1, pp. 53–67, 2011.

[15] A. M. Rosenthal-von der Putten, F. P. Schulte, S. C. Eimler, L. Hoffmann, S. Sobieraj, S. Maderwald, N. C. Kramer, and M. Brand, “Neural correlates of empathy towards robots,” in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 215–216.

[16] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nehaniv, E. A. Sisbot, R. Alami, and T. Simon, “How may i serve you?: a robot companion approaching a seated person in a helping context,” in *Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human Robot Interaction*. ACM Press, 2006, pp. 172–179.

[17] R. J. Fisher, “Social desirability bias and the validity of indirect questioning,” *Journal of Consumer Research*, vol. 20, no. 2, pp. 303–315, 1993.

[18] S. Tsang, C. Royse, and A. Terkawi, “Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine,” *Saudi Journal of Anaesthesia*, vol. 11, no. 5, pp. 80–89, 2017.

[19] C. Lallemand, V. Koenig, G. Gronier, and R. Martin, “Cration et validation dune version franaise du questionnaire AttrakDiff pour lvaluation de lexprience utilisateur des systmes interactifs,” *Revue Europenne de Psychologie Applique/European Review of Applied Psychology*, vol. 65, no. 5, pp. 239–252, 2015.

[20] D. Kulic and E. Croft, “Physiological and subjective responses to articulated robot motion,” *Robotica*, vol. 25, no. 1, pp. 13–27, 2007.

[21] C. L. Bethel, K. Salomon, J. L. Burke, and R. R. Murphy, “Psychophysiological experimental design for use in human-robot interaction studies,” in *The 2007 International Symposium on Collaborative Technologies and Systems IEEE*, 2007.

[22] M. Gharbi, P.-V. Paubel, A. Clodic, O. Carreras, R. Alami, and J.-M. Cellier, “Toward a better understanding of the communication cues involved in a human-robot object transfer,” in *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*. IEEE, 2015, pp. 319–324.

[23] J. Riley, L. Strater, S. Chappell, E. Connors, and M. Endsley, “Situation awareness in human-robot interaction: Challenges and user interface requirements,” in *Human-Robot Interactions in Future Military Operations*, 2010, pp. 171–192.

[24] P. Salmon, N. Stanton, G. Walker, and D. Green, “Situation awareness measurement: A review of applicability for c4i environments,” *Applied Ergonomics*, vol. 37, no. 2, pp. 225–238, 2006.

[25] S. Devin, C. Vrignaud, K. Belhassein, A. Clodic, O. Carreras, and R. Alami, “Evaluating the pertinence of robot decisions in a human-robot joint action context: The PeRDITA questionnaire,” in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 144–151.

[26] M. Posner, *Chronometric explorations of the mind*. Oxford University Press, 1986.

[27] F. Bunlon, J.-P. Gazeau, F. Colloud, P. J. Marshall, and C. A. Bouquet, “Joint action with a virtual robotic vs. human agent,” *Cognitive Systems Research*, vol. 52, pp. 816–827, 2018.

[28] J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey, and J. Ventre-Dominey, “I reach faster when i see you look: Gaze effects in humanhuman and humanrobot face-to-face cooperation,” *Frontiers in Neurorobotics*, vol. 6, 2012.

[29] M. Staudte and M. W. Crocker, “Investigating joint attention mechanisms through spoken humanrobot interaction,” *Cognition*, vol. 120, no. 2, pp. 268–291, 2011.

[30] C. L. Bethel and R. R. Murphy, “Review of human studies methods in HRI and recommendations,” *International Journal of Social Robotics*, vol. 2, no. 4, pp. 347–359, 2010.

[31] M. Baker, “1,500 scientists lift the lid on reproducibility,” *Nature*, vol. 533, no. 7604, pp. 452–454, 2016.

[32] F. Echtler and M. Huler, “Open source, open science, and the replication crisis in HCI,” in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*. ACM Press, 2018, pp. 1–8.

[33] L. Riek, “Wizard of oz studies in HRI: A systematic review and new reporting guidelines,” *Journal of Human-Robot Interaction*, pp. 119–136, 2012.

[34] B. Mutlu. (2015) Open letter to the HRI community | HRI 2015.

[35] E. T. Hall, *The hidden dimension: man’s use of space in public and private*. The Bodley Head Ltd, London, 1966.

[36] A. Sorokowska *et al.*, “Preferred interpersonal distances: A global comparison,” *Journal of Cross-Cultural Psychology*, vol. 48, no. 4, pp. 577–592, 2017.

[37] J. R. Aiello and T. De Carlo Aiello, “The development of personal space: Proxemic behavior of children 6 through 16,” *Human Ecology*, vol. 2, no. 3, pp. 177–189, 1974.