From Characterising Three Years of HRI to Methodology and Reporting Recommendations

Paul Baxter, James Kennedy, Emmanuel Senft, Séverin Lemaignan, Tony Belpaeme
Centre for Robotics and Neural Systems, The Cognition Institute
Plymouth University, Plymouth, U.K.
Email: {paul.baxter,...,tony.belpaeme}@plymouth.ac.uk

***Abstract*—Human-Robot Interaction (HRI) research requires the integration and cooperation of multiple disciplines, technical and social, in order to make progress. In many cases using different motivations, each of these disciplines bring with them different assumptions and methodologies. We assess recent trends in the field of HRI by examining publications in the HRI conference over the past three years (over 100 full papers), and characterise them according to 14 categories. We focus primarily on aspects of methodology. From this, a series of practical recommendations based on rigorous guidelines from other research fields that have not yet become common practice in HRI are proposed. Furthermore, we explore the primary implications of the observed recent trends for the field more generally, in terms of both methodology and research directions. We propose that the interdisciplinary nature of HRI must be maintained, but that a common methodological approach provides a much needed frame of reference to facilitate rigorous future progress.**

***Index Terms*—Challenges; Human-Robot Interaction; Methodology; Recommendations; Research Methods**

I. Introduction

Human-Robot Interaction as a research field lies at the confluence of multiple disciplines, each with their own goals, assumptions, methodologies and techniques (figure 1). As a result, it provides a rich environment for a variety of research questions and empirical investigations. However, this inherent strength brings with it shortfalls in terms of mismatches between disciplines that should be accounted for. In this paper, we provide an overview of the current state of the field of Human-Robot Interaction through the prism of the ACM/IEEE HRI conference, and on this basis provide a set of guiding principles and technical recommendations that will help to consolidate the progress made thus far, and provide a platform for future contributions. In doing so, we seek to promote introspection in the community to provoke discussion, propagate best practice through our characterisations, and provide a guide to newcomers to the study of HRI – an important aspect given the multidisciplinary nature of the field.

We provide two levels of analysis, from researcher-level to field- and community-level. At the researcher-level, we identify good practice from both within and without the field, and formulate practical recommendations that can be readily applied to ongoing and future research. At the field-level, we consider the broader themes resulting from the inherently interdisciplinary nature of HRI, and how these relate to the methodological and technical challenges faced by researchers. In doing so, we seek to highlight common ground and future directions to provoke discussion in the field and ultimately improve the impact of HRI in terms of both research and applications.

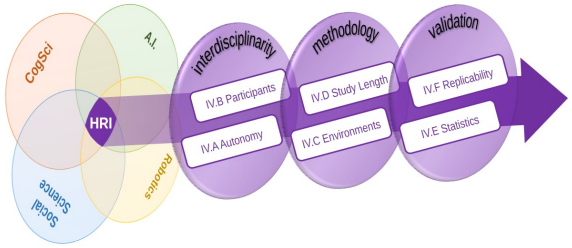

Fig. 1. HRI as a field seeks to integrate knowledge and techniques from multiple disciplines (also including design, psychology, etc), but has its own unique challenges, a number of which we characterise in this paper – numbers correspond to sections in this paper.

We have summarised data from papers presented at the last three HRI conferences (2013, 2014, 2015) to provide recent trends in application and methodology at the primary conference in the field. A total of 101 papers were analysed, with each individual paper classified across 14 categories according to the methods and approaches used within them. This process provides insights into current approaches and emerging trends in the field of HRI.

II. Motivations and Scope

As noted above, each discipline brought into HRI brings with it sets of assumptions and motivations. They may also bring different goals, which may or may not conflict. At the highest level, for instance, we may make a distinction between studies that are theoretically motivated *vs.* application oriented, and between those that are robot centred *vs.* human centred. For example, the use of modelling in cognitive science (where there are increasing numbers of models 'embodied' on robotic platforms) is typically intended to provide an exploration or account of some human-centred phenomenon [1] rather than explicitly seek to improve the robotic agents themselves – although this is on occasion a useful consequence. For robots intended for therapy, e.g. [2], the focus of development is necessarily therapeutic efficacy (i.e. human centred and application oriented) rather than models of robot cognition. In contrast, research to develop physically safe robots to interact with people are more robot-centred and application

oriented emphasising technical contributions, e.g. [3], whereas developmental robotics as applied in human-robot interaction contexts are more robot-centred but theoretically oriented, e.g. [4].

While the presence of this plurality of motivations is not at issue, these differing founding assumptions and intended applications require the use of differing hypotheses and consequently different appropriate methodologies to address them. This is apparent for example when reconsidering the examples from cognitive modelling and therapy: in the former, explicit characterisation of the way a human and robot behave (and possibly how they generate their behaviours) would be necessary, whereas in the latter, a focus would typically be on human behaviour metrics. Whilst such differences do not necessarily result in tension, they can give rise to differing and mismatched expectations between those with different disciplinary backgrounds (as may be expressed in a peer review process for example), typically where the results from one domain are applied to another.

We maintain that this richness is essential for the HRI community, and that it should be preserved. There is a benefit in closer collaboration and the cross-fertilisation of knowledge and methods. One potential means could be to provide a set of benchmarks and target tasks to facilitate comparison between approaches (as with the DARPA or RoboCup@home challenges): a danger of doing so however is the alienation of those parts of the community not engaged in these technical challenges, and the eventual treatment of these benchmarks as ends in their own right, rather than means as originally intended. Therefore, we rather suggest that the provision of a framework to set out common standards and best practice in methodology and reporting centred on the main challenges in the field would encourage and facilitate collaboration and the cross-application of results without bias towards/against any of the disciplines that feed into HRI. To this end, our intention in this paper is to examine and characterise the approaches used in recent HRI conference publications, the challenges that these give rise to, and hence to derive a set of recommendations that can serve as the basis for this common framework.

A reflection of the make up of the conference papers analysed, our perspective in this paper is primarily experimental, irrespective of the actual theme that may have been applied to the paper (e.g. studies, technical advances, design, etc). That is to say, we focus here on the running and reporting of empirical studies rather than theoretical, design or technical contributions in their own right, although we must acknowledge the importance of each of these. Equally, we note that qualitative and ethnographic approaches are fundamentally useful, even if this is not reflected directly in the papers covered in the present review; indeed, the methodological points we discuss below are largely relevant to these approaches in HRI.

In conducting our review exercise in this paper, there are a number of facets of HRI as a field that shaped our decision to focus on recent conference proceedings, with the HRI conference as a particularly important venue, as previously suggested [5]. Since the field is fast paced, with new technological and theoretical developments rapidly shaping the experiments that are run, conference papers provide the most readily and rapidly available results in the peer-reviewed domain, constrasted against the inherently slower publication turn-around of typical journal articles. Our decision to restrict our search to the past three years is similarly intended to explore recent trends given a relatively volatile field.

TABLE I

OVERVIEW OF PAPER AND STUDY TYPES COVERED BY YEAR. NUMBER IN BRACKETS INDICATES FOR EACH CATEGORY THE PERCENTAGE OF PAPERS THAT YEAR. *NHST: Null-Hypothesis Significance Testing*. A 'UNIVERSITY SAMPLE' IS A STUDY WHICH TOOK A SAMPLE OF STUDENTS OR RESEARCH STAFF FROM A UNIVERSITY OR RESEARCH INSTITUTION.

	2013	**2014**	**2015**	**Total**
Number of papers	26	32	43	**101**
With study	25	31	40	**96**
NHST	24 *(96%)*	30 *(97%)*	36 *(90%)*	**90**
University sample	14 *(56%)*	13 *(42%)*	18 *(45%)*	**39**
Lab study	19 *(76%)*	23 *(74%)*	30 *(73%)*	**72**
$>$1 session study	0 *(0%)*	1 *(3%)*	4 *(10%)*	**5**
Uses WoZ	3 *(12%)*	11 *(35%)*	11 *(28%)*	**25**

Through classifying the papers according to the chosen categories, we have identified a number of features of HRI methodology and reporting that warrant consideration, which we have coalesced into six challenges (figure 1 & section III). These challenges are not restricted to any particular disciplinary perspective, but are generally applicable, whilst remaining specific enough to result in practical and actionable recommendations. The aim in doing so is to structure our recommendations so as to provide the foundation for a common frame of reference within which HRI studies with all disciplinary flavours can push the field forward.

III. METHOD

In order to explore the state of the field of HRI, three years of published papers for the Human-Robot Interaction conference were analysed (table I). All 101 full papers from the 2013, 2014 and 2015 proceedings were collated for analysis on the 14 categories shown in table II. All categories were assessed by manually reading the papers and storing the values in a spreadsheet (available at http://goo.gl/PfK1IC).

The categories we chose were ones that were common to all experimental papers, which encompasses the vast majority of papers examined (96 out of 101). They were chosen due to their generality to experimental methodology, being aspects that would be reasonably expected of any study conducted in the field of Human-Robot Interaction. We thus include robot-specific aspects (e.g. nature of control) as well as the standard human-related factors (number of participants, etc), and we suggest that we have included all relevant factors of this nature.

To collect this data, certain definitions were required. Firstly, a lab study is considered to be one in which the participants would have to leave their environment and come to the evaluation location, whereas a non-lab (or 'wild') study is one in which the experimenters go to the participants' environment. Secondly, levels of robot autonomy are described in detail in section IV-A.

TABLE II
OUTLINE OF THE 14 CATEGORIES USED TO CLASSIFY EACH OF THE PAPERS CONSIDERED. *NHST: Null-Hypothesis Significance Testing*.

Category	Classes
Stimuli	Colocated Robot / Non-Colocated Robot / Virtual Robot / Video / Photo / Text / None
Interactive	Yes / No
Robot type/model	Name / N/A
Use of Wizard-of-Oz	Autonomous / Perceptual WoZ / Cognitive WoZ / User Tele-operation / Experimenter Tele-operation / N/A
Occurences of 'wizard'	n / N/A
Study with people	Yes / No
University sample	Yes / No
Mean age participants	Mean / Unstated / Unclear
Conducted in lab setting	Yes / No
Participants per condition	Mean / Unclear / /N/A
Interaction duration (min)	Mean / Unstated / N/A
Interactions per week	n / N/A
Experiment length (weeks)	n / N/A
Use of NHST	Yes / No

The most common unit for each of the relevant categories is used, with translations made if necessary. For papers that present multiple studies, or pilot studies as well as a larger evaluation, only the larger evaluation using a robot, or last study was considered. For interaction durations, if a time range was provided, then the maximum of the range was recorded. Missing data, or cases in which the information was not clear, were annotated in the data collection exercise, with clarification notes appended.

IV. HRI CHARACTERISATION AND RECOMMENDATIONS

Examination of the collected data suggests six broad characteristics that encompass a wide range of non-discipline-specific aspects of HRI research. Roughly following the design process of a system and its subsequent evaluation and reporting, we can consider them to be comprised of (figure 1): robot autonomy and study participants (interdisciplinary aspects), environment and study length (methodological considerations), and statistics reporting and replicability (validation for the community). In the following subsections we provide summary information of the collected data in the 14 identified categories. We note that only 40 of the 96 ($\sim$42%) papers with studies contain all of the information in the 14 categories we examined.

A. Level of Robot Autonomy

We recorded whether or not there was any interaction between the robot (or other stimulus used in an evaluation) and the participants: i.e. those in which the behaviour of the robot is in some way influenced by the behaviour of the interacting human(s). Then, we define several categories in order to assess the levels of autonomy used in HRI studies, shown below, with the results reported in table III. These include a conceptual division in the use of Wizard-of-Oz (WoZ) techniques:

– *Autonomous:* The robot is fully autonomous; minor interventions are still possible, such as starting the system.

TABLE III
AUTONOMY LEVELS ACROSS ALL THREE YEARS OF HRI PUBLICATIONS OF STUDIES, INCLUDING THE IDENTIFICATION OF NUMBER OF *interactive* STUDIES. RELATIONSHIP BETWEEN LEVELS OF AUTONOMY

Autonomy Level	Interactive	Total
Autonomous	38 (*40%*)	46 (*48%*)
Perceptual WoZ	8 (*8%*)	9 (*9%*)
Cognitive WoZ	16 (*17%*)	16 (*17%*)
Participant tele-operation	12 (*13%*)	12 (*13%*)
Experimenter tele-operation	2 (*2%*)	2 (*2%*)
Not Applicable	0 (*0%*)	13 (*14%*)

– *Perceptual WoZ:* The wizard replaces a robotic function (typically a perception capability, such as speech recognition) that could be autonomous (algorithms or tools exist for that function and could have been applied in that context). The function is performed by a wizard for practical reasons (time, difficult technical deployment, computational constraints).

– *Cognitive WoZ:* The wizard replaces cognitive capabilities of the robot, such as deciding what speech to say, what gestures to use, or what actions to take. This can possibly lead the user into ascribing cognitive capabilities onto the robot that do not exist.

– *Participant Tele-operation:* The participant in the study tele-operates the robot as part of the study design, for instance to study shared autonomy.

– *Experimenter Tele-operation:* An experimenter tele-operates the robot as part of the study design, with no intent to deceive participants that the robot has autonomous capabilities (as in the case of WoZ).

– *Not Applicable:* Studies where the autonomy of the robot is not relevant to the procedure, e.g. no robot is present, there is no study, participants watch a video.

In many cases it was difficult to assess the level of autonomy of a robot used in an evaluation. Indeed, 5 papers from 26 utilising a WoZ omit the word 'wizard' altogether. This has previously been raised as an issue in HRI and clear reporting guidelines have already been put forward [6]. Greater adoption of these guidelines would clearly aid the field in understanding the context of the studies conducted.

Note that the level of autonomy, as per our definition, is not to be taken as a proxy for the system (or experiment) *complexity*: some of the systems labeled as autonomous implement simple, fully scripted interactions. On the contrary, some of the wizarded experiments do involve complex autonomous processing for certain parts.

Wizard-of-Oz, as a manipulation technique, is often an experimentally appropriate methodology. A case in point consists in using the robot as a puppet to uncover specific social human behaviours when confronted with a machine (which is typical for the *human centred, theory focused* research line introduced in section II).

When employed, Wizard-of-Oz necessitates special care: since the interaction becomes partially (or in some cases, entirely) a human-human interaction, mediated by a 'mechanical puppet', the researchers need to ensure replicability of the wizarded behaviours between participants, and be careful not to

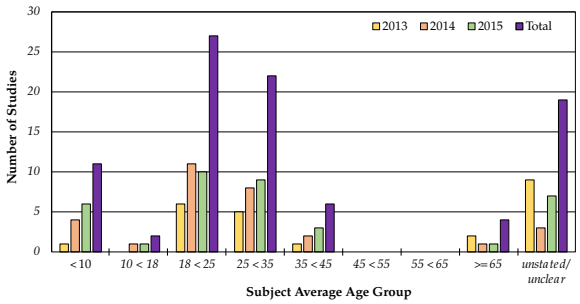

Fig. 2. Histogram of the average age of evaluation participants by age and total from the last 3 years of HRI conference publications. There is a clear peak for the age of student-based samples.

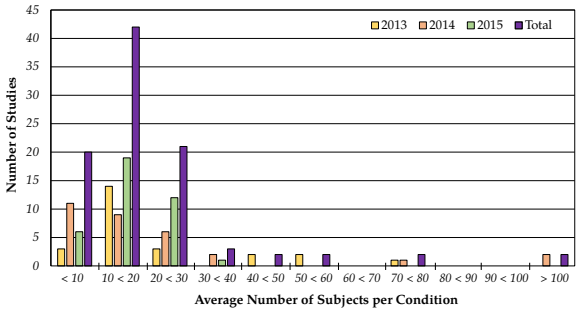

Fig. 3. Histogram of the average number of participants per condition of evaluations from the last 3 years of HRI conference publications. The majority of conditions have fewer than 20 subjects.

introduce human biases [7]. To avoid these pitfalls, a common practice entails the wizard strictly adhering to a pre-defined interaction script.

The level of autonomy of the robot may also alleviate these issues: the more autonomous the robot, the smaller the human intervention surface, and the less likely the introduction of discrepancies between participants, given that a human operator will adapt their own behaviour in the interaction.

According to our findings (table III), around 40% of studies presented at the HRI conference over the last three years have implemented an interaction with a mostly autonomous robot. While this is certainly not negligible, it also means that a majority of the research presented at the HRI conference does not involve interactive autonomous systems.

To address this underlying misunderstanding caused by the differing high-level research goals, and in line with our goal to establish a common framework, one recommendation would consist of explicitly commenting in academic publications on the level of autonomy of the system, set in the perspective of the longer-term scientific agenda.

B. Participant Populations

For ecological validity it is good practice to perform evaluations with samples that are representative of the population with which a system is intended for use (i.e. to avoid *sample bias*). Such practice allows for better generalisation to the 'real-world', which is particularly desirable given that a large quantity of HRI research is conducted in the context of applications which require practicable solutions (autism therapy, child education, elderly care, etc.). There will undeniably be a trade-off between striving for ecological validity and experimental control, but there are a number of steps which can be taken with regards to participant populations that would be of great benefit to the validity of research in the field.

There is a clear imbalance of ages being used in HRI studies (figure 2). When research was not conducted with children (aged less than 18), or the elderly (aged over 65), 87% of studies used samples which drew from university populations (where age is stated). It may be the case that the intended end-user of these findings would indeed be only students/academic staff, or findings are not required to generalise to the wider population, but this seems unlikely to be the case for all instances. Additionally, it is worth noting that of the papers analysed that involved subjects, 18 did not report the age of these participants (figure 2), which further reduces the extent to which conclusions can be drawn.

Such samples are often dubbed 'convenience' samples, and whilst it is indeed convenient to use students which are readily available to test a system, questions must be raised as to how much can be gleaned from any findings. This will vary from case-to-case, but in principle, we feel that convenience samples should be avoided, as they may give rise to sample biases. We should strive towards greater ecological validity to push the field forwards, and ensure that the conclusions do not over-generalise away from the specific characteristics of the participant group used.

In addition, a substantial portion of evaluations in the field gather data from sample sizes which would be considered small in terms of human studies (figure 3). In psychology there have been concerns over small sample sizes leading to underpowered studies, in turn creating an incoherent body of literature [8].

For HRI to avoid these same problems, larger and more representative samples are required. However, this is not so easy to put into practice due to the sheer amount of effort involved in obtaining not just a greater number of participants, but also more diverse ones to maintain the generality of conclusions (where this is appropriate). Indeed, in some cases (e.g. in therapeutic or medical domains), larger sample sizes may not be possible. In this case, the importance of reporting standards come to the fore.

C. Evaluation Environments

The environment in which an evaluation is run can have a great influence on the behaviour and responses of participants [9]. The majority of studies in HRI appear to be run in laboratories, with an average of M=75% (SD=1%) of experiments conducted in the lab over the last three years of HRI conference publications. It has been debated within psychology as to whether lab experiments provide external validity (the extent to which generalisation to other settings and samples is possible) [10], with the conclusion that experiments at least require 'experimental realism': the degree of authenticity with regards to the phenomenon under exploration.

However, there is clearly a motivation for HRI experiments to move out of the lab and into the field, or the 'wild', in order to gather results which have demonstrable applicability. With such a commitment to field studies, there comes a trade-off between control and ecological validity. Some of these issues have previously been discussed in the context of HRI [9]. On the one hand, there is significant effort required on the part of the experimenters to run studies outside the lab, which needs to be acknowledged. Naturally however, the level of effort does not in itself guarantee a good study. Indeed, there is the possibility of introducing a number of new confounds related to the environment itself: for example the potentially complex effects of children talking to each other about the robot whilst the experiment is taking place in a school study.

As with the participants themselves (section IV-B), we suggest that ecological validity should be the main concern: is the experimental environment suitable given the experimental hypotheses? Secondly, we would suggest that since some types of confound are difficult to control for, a minimal requirement should be to report those confounds most likely to have an effect on the hypotheses.

D. Length of Empirical Studies

Novelty has often been raised as a potentially confounding or influencing factor for HRI studies [11], [12]. There is commonly a call for more long-term studies, or a statement of the desire for long-term investigation in the 'future work' section of HRI research papers. Table I shows that from 96 studies in the last 3 years, only 5 have consisted of more than one session interacting with a robot (one in 2014 and four in 2015). Whilst it is recognised that many longer-term studies may be published in different venues (be they journals or other conferences), these figures still raise questions about how we should consider the length of empirical studies.

There are of course many situations in which researchers may either wish to explicitly exploit a novelty effect, or a novelty effect is simply not relevant for the hypotheses in question. However, given a general desire to see HRI systems applicable to, and deployed in, the real world (e.g. as consumer systems), the issue of how human interactant behaviour will change over time as the novelty effect wears off remains an open question, whether this novelty effect applies at the level of the individual with expectations shaped by the anthropomorphic features of the robot (one person interacting repeatedly with a single robot system) or at the societal level (as social robots become commonplace in the public domain). For example, at the individual level, there are some suggestions that once the novelty effect is overcome, the robot behaviour will need to be more than just believable at a shallow level and beyond the role played by the robot embodiment, thus raising the necessity for deeper models of cognition and human behaviour [13].

What then constitutes long-term HRI? We would suggest that this is linked to the overcoming of the novelty effect, which in turn is related to the robot, its behaviour, and the interaction context, as elements influencing the extent to which novel behaviours are preferred over familiar ones [14]. This non-standard concept of the novelty factor may prove problematic in terms of comparing different studies. However, one way of addressing this could be to develop and use reliable behavioural metrics (based on gaze and linguistic behaviours for example) for the characterisation of familiarity.

E. The Approach to Statistics

Null-Hypothesis Significance Testing (NHST) is the de-facto standard for evaluating the importance of results. In this process, one checks the hypothesis that the data distribution (comprising sample size, mean and standard deviation for normally distributed data for example) obtained from an intervention condition does not differ from the distribution from a control condition (the null hypothesis): if this hypothesis can be rejected (i.e. a p-value less than or equal to some threshold, typically 0.05), the result may be considered 'significant'. On the face of it, this provides a useful means of characterising the 'success' (or not) of a method or intervention. This state of affairs is reflected in the HRI papers in our sample: $\sim$95% (90 out of 96 studies, see table I) of the papers employ NHST and report p-values to support the conclusions.

However, in recent years there has been increasing criticism of the importance conferred onto this means of statistical analysis in multiple fields of research[1], e.g. [15]. Indeed, the problematic nature of NHST has been acted upon by certain psychology journals, which have effectively banned the use of it to rest the main results of manuscripts on, e.g. [16]. This reflects three main concerns (and others): the arbitrary threshold for significance, replication sensitivity, and lack of effect size information.

Firstly, significance is typically held at a p-value of 0.05 or less (or 0.01 in the biological sciences). This is an arbitrary threshold (1 in 20 chance) that persists for historical continuity rather than theoretical or empirical merits. Determining the utility and/or importance (and this is often how significance is treated) of the result based on such an arbitrary threshold seems flawed from the perspective of the scientific method. Secondly, empirical results have suggested, and simulation studies have shown, that the p-value is highly volatile in experiment replications, with a variation in an initially significant p-value in the range [0.00008,0.44], 80% of the time [17]. p-values are thus unreliable in the face of replication. Thirdly, p-values do not incorporate any information about effect sizes: a highly statistically significant result from the perspective of NHST does not relate to the size of the observed experimental effect, and thus can not be used alone to assess the importance/impact of the result.

Descriptive statistics is sensibly recommended as the first stage of data analysis: we suggest that an increased emphasis on this should form part of standard reporting practice to circumvent some of the issues raised above. As an extension to this, we thus recommend that a minimal requirement for reporting mean-based data from multiple conditions should be

[1]Note that NHST is rigourous and mathematically valid, and thus not intrinsically problematic - the issue is rather the interpretation of the result, and the meaning derived from it in experimental contexts.

the provision by authors of Confidence Intervals (CI's) [17], [18], where the 95% CI is typically used[2]. Whereas p-values vary to a great extent, CI's have been shown to be more reliable, with an 83% chance that replication will give a mean within the CI of the original experiment [19]. CI's also inherently provide information about the effect size, thus providing an additional benefit over the reporting of p-values alone.

A further approach that could be brought to bear on this problem is statistical modelling. While this is on occasion seen to merely be an alternative means of performing a statistical analysis, we suggest that it should rather be seen as a change of perspective. Rather than forming just another statistical test of significance, the purpose is to gain an incrementally better view of the phenomena under investigation. In the Bayesian modelling perspective for example, there is an emphasis on the accumulation of data, of integrating new observations with existing knowledge. Previous results help to form *priors* for example, which shapes the way new data is viewed. In this perspective, the role of experimental methodology takes on a more central importance – it becomes the means by which data may be consistently integrated into ever more reliable priors. Our focus on guidelines to form a common methodological frame of reference thus feeds into these efforts.

F. Replicability

Replication (conducting the same experiment anew) and *reproduction* (re-running analyses on the original data to validate results) are instrumental in weaving a solid and trustworthy scientific fabric. Concerns have been voiced over the replicability of results in the sciences [20]. A recent large-scale replication of 100 psychology studies resulted in only 36% of studies having significant results, while originally 97 of the 100 studies reported significance ($p < 0.05$). A looser, subjective definition of replication found that only 39% of results could be deemed as successfully replicated [21]. While no published evidence exists on the replication of HRI studies, it is likely that replication will be of a similar level, due to the many methodological parallels between HRI studies and psychology studies.

A first obstacle is the lack of replicability: HRI studies are often challenging to replicate due to the nature of robotic hardware, the experimental setup, and the particular platform, environment and participants used. Access to specific robotic hardware is often restricted, especially if hardware is rare, expensive or difficult to access – e.g. androids or bespoke platforms. In addition, publications often do not have a detailed methods section facilitating replication, and software is, despite increased attention for open source initiatives, not widely shared in the HRI community.

On the other hand, increasing the reproducibility of our studies is likely less of a challenge. It mainly calls for sharing datasets and/or results and the means of analysing them (e.g. data processing scripts). Whenever the datasets can not be made anonymous, privacy concerns are likely to arise: those may be alleviated with agreed consent from the participants that "their data may be used for academic purposes" and through adequate sharing methods within the community. Note that we observe in recent years a clear trend toward ensuring datasets are available for papers to be considered for publication (case in point, taken from the author guidelines of PLOSOne: *"PLOS will not consider submissions from which the conclusions are based on proprietary data")*. We can only encourage the HRI community to actively embrace this practice.

A second obstacle however is the lack of incentive to replicate or reproduce studies. Academic reward systems and the current reviewing culture favour novelty over replication. This not only leads to a lack of validation of results and claims, but leads the field to chase the novel and exciting, rather than confirming or –perhaps even more importantly– refuting claims. As [21] eloquently points out, "Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both".

A possible solution might be to create a new outlet for replication studies: if a journal or conference would welcome brief publications on successful or unsuccessful replications, this would demonstrate that replication is valued and would incentivise the consolidation of HRI insights.

V. Discussion

Our identification of six characteristics of HRI studies, supported by recent conference publication trends, and our subsequent exploration, have led to the proposal of six recommendations. These are both specific *researcher-level* recommendations that can be readily and practically applied to ongoing empirical work and the reporting of these, and also more general *field-level* recommendations that apply to the level of the field rather than individual researchers (see table IV for a summary).

A. Interdisciplinary Methods and Tools

Given the diversity of discipline-specific motivations and goals (section II), there are a number of sources that emphasise the importance of a common or shared mission if interdisciplinary efforts are to succeed, e.g. [22]. At the level of the field, we caution against specifying a mission statement that is too specific in terms of application or method. Such an effort would be likely to provide exclusions from the field, which we would suggest is (at least currently) unnecessary. From this perspective the current (brief) mission statement listed on the HRI community website provides a suitably general outline of the field: "*HRI is the multidisciplinary study of human-robot interaction*". At the level of individual research contributions however (e.g. a single study, series of experiments, or project), we believe such a statement to be a necessity for clarity of hypothesis, coherency, and appropriateness of the methods and metrics employed to investigate them.

However, with such a broad mission statement as used by the HRI community website, there need to be structures in place to ensure coherence in the field and to promote cross-disciplinary

[2]The use of 95% is a similarly arbitrary threshold as the 0.05 threshold for NHST p-values. However, CI's only provide a descriptive perspective, and not a metric of significance in themselves, thus avoiding the threshold problem.

TABLE IV
A SUMMARY OF THE RECOMMENDATIONS, WITH OPERATIONAL SUGGESTIONS AT RESEARCHER-LEVEL (ON THE LEFT) AND AT THE FIELD-LEVEL (ON THE RIGHT), WHERE APPROPRIATE.

R1: State the motivation, context and long-term goal of the research	
State the end-goal of the research (e.g. therapy, cognitive modelling, etc)	Provision and curation of collaborative, open tools to facilitate shared understanding and best-practice
R2: Clarify the level of robot autonomy	
The level of robot autonomy and/or 'wizarding' should be specifically and clearly stated; wizarded robot behaviours should be avoided as a benchmark condition.	
R3: Use of ecologically valid subject groups and experiment environments	
Based on the experimental hypotheses, assess the appropriate subject group; recognise the constraints that the use of a single subject group imposes on the study conclusions	
R4: Relate the notion of long-term interactions to overcoming the novelty effect	
Introduce metrics for familiarity of the study subjects with the robot as a means of characterising the novelty effect	
R5: Use descriptive statistics	
As a minimal requirement, report 95% Confidence Interval for metrics of each condition; emphasise the build up of evidence over arbitrary significance judgements.	Enforce reporting standards in conference and journal publications
R6: Support replication and reproduction	
Ensure detailed methodology; provide source code whenever possible; publish datasets and/or intermediary results, along with the tooling to analyse them (when applicable)	Provision of a peer-reviewed publication venue specifically for independent experimental replications; provide guidelines and infrastructure to share datasets

collaboration while preventing fragmentation. We suggest above (section II) that the imposition of common benchmark tasks could introduce unwanted biases in the long-term, and introduce technical barriers to entry for certain sections of the community. Our proposal to formulate a common framework for methodological and reporting considerations forms the beginning of an alternative approach. In the same way that a characterisation methodology such as conversational analysis can provide a common and formal basis for comparison of qualitative observations between studies, so can such a common framework do the same for the multiple disciplines within HRI. The recommendations we propose (summarised in table IV) are pitched at two levels to encourage a coordinated effort at achieving this: standards for individual researchers to follow, but also suggested changes in field-level infrastructure that can bring about the wider cultural change desired to facilitate the efforts of individuals. Indeed, such efforts are apparent in other fields, for example in health research (equator-network.org).

Regarding this field-level infrastructure, the provision of a number of tools for collaboration and shared understanding would be of use in addressing some of the issues that arise from a vibrantly interdisciplinary field. One such tool is a community FAQ. Such a resource could contain technical advice/resources, reporting recommendations, explanations of key jargon, best practices, etc. covering all HRI disciplines (quantitative and qualitative, technical and social). This would contribute to bridging cross-cultural "language" issues by having one entry-point that researchers (and newcomers to HRI research in particular) could use as a reference.

However, as with any introduction of new standards and/or recommendations, there is a need to minimise the 'barrier to entry' to maximise uptake within the community. The more specific researcher-level recommendations we make are pitched to minimise this barrier, whilst providing significant benefits. Our recommendations for collaborative tools and field-level infrastructure (publication support for peer-reviewed replication studies for example) on the other hand will require more significant personal investment, although if such tools are mandated as part of article submission processes (for example), the motivation to conform is likely to prove sufficient to overcome any initial inertia.

B. Facilitating Long-Term HRI

One feature raised from recent studies is a notably small number of longer-term studies (section IV-D). Since novelty effects are typically present in shorter-term evaluations, and given the as yet under-appreciated role that robot morphology design plays in shaping interaction expectations, it is difficult to assess from current evidence what long-term phenomena arise in genuinely long-term interactions between humans and robots. In this case, there is a strong drive to increase the autonomous competencies of the robots that are able to support these studies. However, our paper review exercise has shown, commensurate with the interdisciplinary nature of the field of HRI, that levels of autonomy in robotic systems are currently only limited (section IV-A). This clearly represents a significant challenge for the community: with the requirement for autonomous behaviour comes a need for more elaborated models of appropriate robot behaviour generation in response to social and environmental cues. Efforts in this area are becoming increasingly prevalent in the fields of AI and Cognitive Science, with a multitude of cognitive architectures being developed [23], although these have as yet only a limited impact in HRI.

This requirement for deeper levels of cognitive model is not in our view restricted to the more robot-centred strands of HRI; we suggest it is also a central requirement for the human-centred perspectives. There is a need to formalise in some way the knowledge of human behaviour and adaptation (including psychology, cultural studies, and neuroscience to

varying degrees) to enable application to HRI, whether it is in the form of a robotic system, or as a means of analysing human behaviour (whether it be reaction times or learning outcomes) in an experimental setting.

C. Discipline Dependencies

From the outset of this paper, we have emphasised that HRI lies at the convergence of multiple disciplines; we have also suggested that it would be beneficial to maintain this plurality of approaches. However, we must then also acknowledge that these different disciplines have differing dependencies and goals (section II).

For example, technical developments have the power to advance the field. Given the central role of robots in HRI (in all senses of the phrase), this is uncontroversial. However, there are mutual constraints on these developments. For example, as we have shown (section IV-A), robot wizarding is partially employed to overcome various technical challenges, which results in a limited capacity to engage in long-term studies (section IV-D). Whereas technically-oriented papers may typically appear in other publication venues, the more recent introduction of the technical theme in the HRI conference reflects an acknowledgement of this dependency on technical issues. Nevertheless, it may be worth raising the expectations of the technical content of all HRI contributions as part of the review process, in the same way that methodological issues are currently rigorously assessed.

There of course remain further open questions in the field that will require multi-disciplinary consideration. One notable example of this is the role that robot behaviour and morphology relate to one another with respect to human perceptions and reactions. Such theoretical and design questions are clearly fundamental to overall progress in the field, including to applications. We suggest that the resolution to these issues, and others, will require the application of empirical investigation to characterise and explore the phenomena: i.e. conducting studies to collect data to subsequently inform further refinement. Our focus in this paper on providing a common frame of reference through methodological guidelines is precisely aimed at providing support for such multi- and cross-disciplinary efforts: our recommendations (table IV) provide the basis of this frame of reference.

VI. Conclusion

What we advocate for the field of HRI is the maintenance of the plurality of discipline-specific motivations, rather than the imposition of a single set. Nevertheless, a common framework should be provided to facilitate the interaction of these differing approaches such that the non-unitary field as a whole can move forward. In other words: to maintain HRI as a collaborative field between disciplines, rather than their unification into a new single field. In this paper, we have examined recent trends in HRI publications to define challenges that face this interdisciplinary approach, and derived both practical and more general methodological recommendations that we suggest will provide the start of a much needed common frame of reference that will consolidate the progress made thus far, and provide a platform for future contributions.

Acknowledgements

This work is partially funded by the EU FP7 project DREAM (grant 611391, http://dream2020.eu), and the EU H2020 project L2TOR (grant number 688014, http://www.l2tor.eu).

References

[1] J. L. McClelland, "The Place of Modeling in Cognitive Science," *Topics in Cognitive Science*, vol. 1, no. 1, pp. 11–38, Jan. 2009.

[2] A. Tapus, M. J. Mataric, and B. Scassellati, "The Grand Challenges in Socially Assistive Robotics," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 35–42, 2007.

[3] B. D. Argall and A. G. Billard, "A survey of Tactile Human-Robot Interactions," *Robotics and Autonomous Systems*, vol. 58, no. 10, pp. 1159–1176, 2010.

[4] Y. Nagai and K. J. Rohlfing, "Computational analysis of motionese toward scaffolding robot action learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 44–54, 2009.

[5] C. Bartneck, "The end of the beginning: a reflection on the first five years of the HRI conference," *Scientometrics*, vol. 86, no. 2, pp. 487–504, 2011.

[6] L. D. Riek, "Wizard of Oz studies in HRI: A systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, 2012.

[7] I. Howley, T. Kanda, K. Hayashi, and C. Rosé, "Effects of social presence and social role on help-seeking and learning," in *9th ACM/IEEE International Conference on Human-Robot Interaction (HRI'14)*. ACM, 2014, pp. 415–422.

[8] S. E. Maxwell, "The persistence of underpowered studies in psychological research: Causes, consequences, and remedies," *Psychological Methods*, vol. 9, no. 2, p. 147, 2004.

[9] R. Ros *et al.*, "Child-robot interaction in the wild: Advice to the aspiring experimenter," in *13th International Conference on Multimodal Interfaces (ICMI'11)*. ACM, 2011, pp. 335–342.

[10] L. Berkowitz and E. Donnerstein, "External validity is more than skin deep: Some answers to criticisms of laboratory experiments," *American Psychologist*, vol. 37, no. 3, p. 245, 1982.

[11] R. Gockley *et al.*, "Designing Robots for Long-Term Social Interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*. IEEE, 2005, pp. 1338–1343.

[12] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *Human-Computer Interaction*, vol. 19, no. 1, pp. 61–84, 2004.

[13] S. Lemaignan, J. Fink, P. Dillenbourg, and C. Braboszcz, "The Cognitive Correlates of Anthropomorphism," in *Proceedings of the Workshop: A bridge between robotics and neuroscience at the Human-Robot Interaction Conference*, Bielefeld, Germany, 2014.

[14] I. Leite, C. Martinho, and A. Paiva, "Social Robots for Long-Term Interaction: A Survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[15] R. Nuzzo, "Statistical errors," *Nature*, vol. 506, no. 7487, pp. 150–152, 2014.

[16] D. Trafimow and M. Marks, "Editorial," *Basic and Applied Social Psychology*, vol. 37, no. 1, pp. 1–2, 2015.

[17] G. Cumming, "Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better," *Perspectives on Psychological Science*, vol. 3, no. 4, pp. 286–300, Jul. 2008.

[18] D. H. Johnson, "The Insignificance of Statistical Significance Testing," *The Journal of Wildlife Management*, pp. 763–772, 1999.

[19] G. Cumming, J. Williams, and F. Fidler, "Replication and researchers' understanding of confidence intervals and standard error bars," *Understanding Statistics*, vol. 3, no. 4, pp. 299–311, 2004.

[20] R. D. Peng, "Reproducible Research in Computational Science," *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.

[21] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, 2015.

[22] R. R. Brown, A. Deletic, and T. H. Wong, "How to Catalyse Collaboration," *Nature*, vol. 525, pp. 315–317, 2015.

[23] D. Vernon, *Cognitive Systems - A Primer*. MIT Press, 2014.