Experimental Studies of Human Robot Interaction: Threats to Valid Interpretation from Methodological Constraints Associated with Experimental Manipulations

J Michael Innes [1,2][0000-0002-7232-6937] and Ben W Morrison [3,4][0000-0002-5026-4675]

1. Australian College of Applied Psychology
2. University of South Australia
3. Charles Sturt University
4. Macquarie University

Address for correspondence: Professor J M Innes, Australian College of Applied Psychology, 16 Coglin Street, Adelaide, South Australia, Australia 5000

Email: mike.innes@navitas.com

Telephone +61423322811

Running Heads: Experiments and robot interaction

Experimental Studies of Human Robot Interaction: Threats to Valid Interpretation from Methodological Constraints Associated with Experimental Manipulations

Abstract

The rapid development of Artificial Intelligence (AI) brings with it the increasing likelihood of ubiquitous interaction between humans and robots. A significant contribution to studying human-robot interactions (HRI) comes from experimental studies, whereby humans and robots interact in controlled conditions and researchers observe and measure the reactions of humans (and robots). The use of experiments to understand human interactions has long been a central source of information in the field of experimental social psychology. These studies have yielded numerous major insights into the causes and outcomes of interaction. The methodology of experiments, however, including the demands made upon human participants to behave in predictable ways and the impact of experimenters' expectancies upon results, has been a focus of much critical analysis. We examined a sample of 100 high impact HRI studies for evidence of potentially contaminating experimental artefacts and/or authors' awareness of such factors. In our conclusions we highlight several methodological issues that appeared frequently in our sample, which may impede generalisations from laboratory experiments to real-world settings. Ultimately, we suggest that researchers may need to reformulate the methodologies used to study the unique features of HRI, and offer a number of recommendations for researchers designing HRI experiments.

Key words: Human-Robot Interaction; Experiment Artefacts; Demand characteristics; Experimenter bias

Experimental Studies of Human Robot Interaction: Threats to Valid Interpretation from Methodological Constraints Associated with Experimental Manipulations

1. Introduction

The rapid development of Artificial Intelligence (AI) brings with it increasing deployment of robots into places where human beings have previously been the sole or principal occupant. Interaction with robotic systems in the office, factory, home, schools, laboratories, and hospitals will become ubiquitous [80]. Clearly, we expect that humans will interact and cooperate with these systems.

This interaction goes beyond traditional human-computer interaction, whereby humans are aided by the efficiency of the processing systems of the computer. The interaction will take the form of physical and mental interaction with a responsive "other", which will add to the outcome of the interaction over and above the capacity of humans in the system. Indeed, the robot may make significantly different inputs, changing the direction of the course of the interaction. The quality of such interaction, however, will depend upon the ability of humans to adjust to interaction with non-human artefacts, and the ability of designers to manipulate the robot to behave in a manner that will be acceptable to the human.

Commentators on the development of AI point out that the design of the artificial systems starts as a human endeavour, embedding within it human values, biases, preconceptions, and intentions [8]. The initial interaction of the human with such systems, therefore, involves the interaction of a human with another human, mediated by the embodiment in the robot. However, the evolution of the system will eventually result in an emergent organism/machine participating in the process. The question arises: how can we explore the nature of such emergent and developing interactions so that they may be beneficial and productive and continue to be acceptable to the human participant? The nature of the participants may also require methodologists to formulate new procedures to enable understanding of the processes that emerge and lead to positive and negative outcomes of the interactions. In the field of AI, a particularly relevant example comes from the study of the chess-playing program

'AlphaZero'. The program, which learns to play chess given only the rules of the game and guided by reinforcement-based learning, has astounded many observers by demonstrating human-like tactics in playing the game. For instance, the program plays sacrifices and employs very unusual moves, with a dynamism more typically associated with grandmasters than knowledge-based programs [74]. Further, the program has contributed to the development of the game itself, giving grandmasters new insights. Thus, the emergent system has now developed beyond the initial program.

The rules of chess, however, are transparent, simple, and defined in a program relatively easily. Social interaction, on the other hand, has no clear rules and depends heavily on very subtle cues and responses. Therefore, in the study of social interaction between robots and humans the question arises: how do designers acquire information about social interaction that enables them to program the relevant cues and responses to allow the interactions to develop?

1.1 The Use of Experiments in Studies of Human-Robot Interaction

The exploration of human-robot interaction (HRI), and the manipulation of factors that may affect the quality of interaction, have been advanced via the use of experimental methods. For present purposes, we define such experiments as a simulated interaction between a human being and a robot with the controlled manipulation of factors that are theoretically linked to the quality and value of the interaction. HRI researchers measure outcomes by the extent to which the interaction brings benefits, including the satisfaction of the human being resulting from the interaction, together with possible unintended and/or negative consequences. Experimentation may also entail the control of participants to the interactions, through the random allocation of participants to experimental conditions (i.e., the randomised controlled experiment). However, this is not a necessary requirement, with the possibility of manipulation of variables in less controlled field conditions where there is an inability or non-desirability to allocate participants at random.

The use of experiments to explore HRI does not take place in a vacuum. HRI researchers can learn a great deal from the experiences of researchers in other fields of inquiry. The use of experiments has a long history in the field of experimental social psychology. Many texts and articles

describe the need for careful design and the subtle creation and control of variables in artificial conditions in order to establish valid outcomes (e.g., [78]). The goal for social psychological researchers has been the creation of social features hypothetically likened to promote or inhibit social behaviour in interactional settings, and observing and measuring outcomes in controlled conditions. This goal presumably shares many attributes with the goals of HRI researchers.

The creation of the laboratory experiment has been likened to an art form [81]; experience and tacit, intuitive expertise is required to be able to create the settings in which humans may relate to others. More specifically, we may liken it to a form of drama where human participants are exposed to social events to which they may not have been previously exposed. The choice of independent variables to be manipulated is usually dependent upon theory and the choice of dependent measures is also associated with purposively selected techniques. The relevance of these issues, and their bases upon both an intuitive understanding of the subtleties and complexities of human behaviour and the relevant theoretically important variables, has several implications for understanding experimental results.

At the most fundamental level, the design of experiments with human participants requires the structuring of interactions and the selection of variables based upon a tacit, and therefore often intuitive, understanding of human behaviour. An experimenter "knows" what it is to be human and therefore can create social settings that fit in with the expectations of participants of what is expected in a social interaction (c.f. [13] for an extensive analysis of this concept for an understanding of social behaviour). Interaction with a robot, however, may involve a lack of understanding of what the robot understands in the setting. While a human being, at least initially, programs the robot's behaviour, the human will influence what the robot is likely to do. With continued association in an interaction, features of interaction may emerge that are not predicted from the initial program, as has occurred with the development of AlphaZero. Therefore, there may be the emergence of novel expectations and a failure to meet expectations on the part of the human participant, which the experimenter must assess and understand. The uncertainty of the initial behaviour programmed into the interaction only adds to the complexities of the interaction that will emerge.

At another level, the role of tacit knowledge in researchers' decisions to study certain phenomena and how to study them has long been considered in the literature on the sociology and the philosophy of science [7, 14-16, 27, and 58 among many others]. Arguably, an understanding of these fundamental processes in the conduct of scientific research is a pre-requisite in the design of HRI research studies.

1.2 Artefacts in the Design of Experiments

Researchers in the field of experimental social psychology have explored the factors, which, inherent to the creation of an experimental situation, threaten the validity of the relationship between the manipulated variable and the observed outcomes. Experimentation in the social and behavioural sciences is not the same as experimentation in the physical sciences. In human social interactions, participants bring expectations, intentions, and biases that will invariably affect the situation. They are not passive objects; they are actors and agents [43].

Fundamental to an understanding of the significant constraints on the interpretability of experiments is the fact that the experiment is, itself, an example of a human interaction. In human social interactions, participants bring to the situation expectations, intentions, and biases, which affect the outcomes of the interaction. This is part and parcel of human interaction and its complexity. There is a process of exchange between the participants, with change in one party followed by change and accommodation in the other. In the experiment, the actors bring to the setting biases and expectations formed in previous social encounters, which are extrinsic to any of the formal, theoretical factors that the experimenter is attempting to observe and measure.

The challenge to the experimenter is to attempt to minimise the effects of these background factors and explicate the link between the manipulated variable and the observed outcome of the interaction. In addition, experimentation requires the creation of artefacts that simulate the "real" conditions of the world, and these artefacts may introduce biases into the experimental setting, which affect outcomes in ways incommensurate with the original hypotheses.

Consideration of the influence of such methodological factors has been associated with many prominent figures in the field, in particular Donald Campbell and his co-workers (c.f. [9, 17, 75]). Indeed, a particularly cogent exposition of the importance of viewing the intrinsic validity of experimental investigations within the broader context of the philosophy of science and psychology can be found in [10]. Campbell's critical point is that experimental artefacts may affect not only the internal validity (the existence of a cause and effect relationship between variables) and external validity (the degree to which the results of one experiment may generalise to another) of an experiment, but critically, its construct validity (the degree to which the variables being studied have meaning in themselves, without additional and extraneous meanings).

Demand characteristics [53, 54] represent a noteworthy example of an experimental artefact. Here, the creation of a formal experimental interaction sets up expectations in the participants about what are acceptable and expected ways to behave. Participants know that they are being observed, and perhaps judged, and therefore will monitor their behaviour to suit what they see as the appropriate rules of the setting. These implicit judgements of the participants may not align with the expectations of the experimenter. Associated with this factor is the sense of apprehension of being evaluated in an experiment [70], with the participant believing that above all other aspects of the experiment there is an attempt to assess the degree to which he or she behaves in a valued, honest, or socially appropriate manner. Also, if the participant has volunteered for participation, then the motivation to volunteer may bring with it additional behavioural characteristics and dispositions, which will result in different outcomes had the person not been a volunteer [68]. Even the act of assessment of behaviour, in the form of a preliminary test of beliefs, attitudes, or expectations, has been shown to influence the later trajectory of behaviour over time by raising expectations about the objective of the experiment [40].

In addition to the factors introduced to an experimental setting by the use of human beings as participants, the experiment brings with it an additional factor, namely the meta-factor of the experimenter being a participant, albeit one who is playing a different role, with different rules of engagement compared to the naïve participants. The experimenter is, at one level, an additional stimulus in the experimental setting [25, 37, 44, 48, 71] and that stimulus may have effects outside of

any effect of the theoretical variables being explored. The participants may also have beliefs and expectations about the experimenters and their roles, including some suspicion about what is being done and why [46].

There are also biases that result from the expectation of the experimenter. After all, the experiment is primarily being conducted to test a theory. While the aim of the experiment in formal philosophical terms is to falsify the theory [59], experimenters are usually trying to demonstrate that there exists a causal link between a manipulated variable and a behavioural outcome [49, 62-68]. These artefacts can combine to have complex effects upon the behaviour of participants, as experiments demonstrating the interaction of pre-tests and characteristics of volunteers have demonstrated (c.f. [72]). While not all studies have demonstrated biases resulting from experimenter effects (e.g., [1]), there are simply too many replications for any impartial observers to remain unconvinced [66, 67]. The artefacts outside of the formal theoretical factors studied in a behavioural experiment are simply too voluminous and have too much proven impact to be ignored.

A further factor may also be relevant to the development of systems of HRI. The experimenter may also have beliefs and expectations of what are the expected outcomes of the study based upon extra-scientific beliefs and attitudes, which lie outside of the formal, scientific activity of designing and conducting the experiment. Such intentions may influence, perhaps unconsciously and unintentionally, the ways in which the experiment is conducted and the outcomes created [29, 32, 35, 39, 41, 46, 56, 76). The influence of socio-political biases on the outcomes of social psychological research has recently become prominent in the field of social psychology [19, 21, 33] and these factors will increasingly need to be considered in any future research. Given the likelihood that large economic consequences for the workforce at all levels of society and the subsequent social and moral outcomes (e.g., [36]), the beliefs that investigators have about economic forces and social systems may play a role in the guidance and creation of social experiments.

1.3 The Decline of Interest in Artefacts of Design

The 1960s and the 1970s saw the development of methodology to understand the biasing influences of these variables. The evidence for the impact of such variables led some to predict the demise of the discipline of social psychology (e.g., [6]). There were many within the field who believed that the utility of the experimental method with the creation of strong socially demanding variables was waning (e.g., [28, 57, 82, 83]). Initially, in an attempt to escape the deleterious impact of experimenter bias, the field showed a renewed vigour in its approach to experimental design. An increased awareness of the problems led to greater care in the design of experiments. Researchers introduced a raft of strategies to mitigate the risk of experimenter bias, including the active blinding and/or exclusion of the experimenter from the setting, participant 'hypothesis checking', the use of multiple coders in scoring observations, and more recently, a greater emphasis on replication (a detailed account of these strategies can be found in section 4.1 Future Recommendations for best-practice in the design of HRI experiments).

Later years, however, witnessed the demise of concerns about the influence of artefacts. Various developments in the field of social psychology led researchers to think that the demands of the setting or the biases of the experimenter were being implicitly controlled in the methodology adopted. These included the move away from the creation of behavioural encounters between participants and the substitution of hypothetical interactions in paper-and pencil questionnaires or use of computer-based automatized interactions between participants and simulated interactors (c.f. [3]). The automation of experimental procedures, with computers providing stimulus exposure without the appearance of a human operator, was also expected to have helped to eliminate bias and expectancy effects.

Recent papers have demonstrated the demise of interest in the effects of experimenters' and generalised expectancies on interactions [34, 38] in laboratory settings [5]. However, we suggest that the threats to the validity of experiments have not gone away, but have merely been forgotten or ignored. Their existence has become normalised, and with this, their perceived importance has been diminished. Although it might be considered reasonable to 'live with' these threats with a healthy awareness and caution, the normalisation has instead led to complete absence of even a mention of

caution. A failure to recognize or to acknowledge research of an earlier era is, of course, not new; this phenomenon has been recognised frequently (e.g., [46, 78] and also then promptly forgotten. This is not a case, either, of "citation amnesia" (e.g., [61]). There is not a failure to cite particular articles in summarising previous research; it is a matter of ignoring, or being unaware of, domains of research activity from the past and may be a form of what O'Gorman [51] has referred to as "making the modern culture of amnesia".

One major outcome from a decline in interest in experimental artefacts has been the emergence of concerns about the replicability of studies in social psychology. Many studies have failed to replicate their findings, and the intrusion of experimental artefacts has been proposed as one explanatory factor. In one spectacular example, researchers [22] failed to replicate a long-standing result in the social psychological literature, namely the unconscious priming or stimulation of a stereotypic act [2]. The researchers concluded that experimenters' expectations were instrumental in explaining the original effect. This resulted in the consideration by many that the methods of social psychology were fundamentally flawed based upon the tacit knowledge of certain particularly talented experimenters.

Artefacts may vary in presence and magnitude from laboratory to laboratory and the manipulation of variables, dependent upon the ability of the investigator to use their tacit understanding of social behaviour, may vary due to the ability of the experimenter. Therefore, the replication of studies, which on the surface manipulate the same variable, may not in fact do so because of unconscious tacit factors. The field with which we are currently concerned is the study of HRI. We ask the question; is the experimental study of such interactions failing to take account of the subtle and not so subtle artefacts discussed here? We examine a sub-sample of recent papers published in three high profile research outlets to determine: (1) the prevalence of experimental artefacts in this field of research; and (2) the frequency with which authors acknowledge the existence and potential impacts of such factors.

2. Method

2.1 Literature search strategy

To investigate the presence of experimental artefacts in the existing literature, a sample of recently published work was collected from two prominent journals in the field, which both demonstrated relatively high impact; the *Journal of Human-Robot Interaction (now ACM Transactions on Human-Robot Interaction)* and the *International Journal of Social Robotics*. Further, in recognising the importance of conference proceedings in the Information Technology fields, peer-reviewed proceedings from the *Annual ACM/IEEE International Conference on Human Robot Interaction*, widely recognised as the flagship conference in the HRI scientific community, were included in the collection. In using this strategy, our intention was to capture a 'snapshot' of recent impactful HRI research, rather than conduct an exhaustive and comprehensive collection process that would be more appropriate for a systematic review.

The terms 'human', 'robot', 'interaction', and 'experiment' were selected and used for the literature search, which was limited to the previous 10 years of publication. Articles were selected initially only on the basis of the title of the paper, which indicated that the content was concerned with the study of human and robot interaction with a likelihood that observations were made of the results of actual interaction, and that these were done under experimental conditions with a comparison of conditions executed under different instructions.

This initial search yielded an extensive listing of results, and so the authors randomly selected a sample from this population of papers for analysis. The authors selected 30 studies from the Journal of Human-Robot Interaction, 40 from the International Journal of Social Robotics, and 30 from the four previous (2016-2019) International HRI conferences. The differences in the size of the samples drawn from the two journals roughly reflects the different lengths of time the outlets have existed and the frequency of publication.

2.2 Procedure

Two authors examined the articles independently. The authors examined the method sections of each paper to determine the presence of potential experimental artefacts in the procedures

employed. The authors also examined the bibliographies to ascertain whether they contained any reference to the literature on experimenter bias and or demand characteristics. Note was also taken of more general reference to literature in social psychology that might indicate a general awareness of factors within an experiment, which might affect the behaviour of participants.

In an attempt to counter similar expectancy effects to those discussed by the authors in the current paper, an independent researcher trained in experimental methodology was recruited to conduct a review of a sub-sample of the papers collected (five from each source). The researcher had no awareness of our specific research questions, nor did she have previous experience in HRI research. To eliminate expectancy effects, we only instructed this reviewer to critically review the studies.

3. Results

The data from both authors and the independent reviewer were collated and common issues relating to experimental artefacts are discussed here. From the limited sample of papers assigned to her for review, the independent reviewer identified the same experimental artefacts (or potential for them) identified by the two authors. The results of our examination reveal a pervasive pattern of potential demand characteristics and bias in the experimental methods employed in HRI studies.

In many of the cases, we found explicit reference in the methods section to the physical presence of an experimenter or observer during the behavioural sequences and/or to the use of recording devices, which were not concealed from participants. Further, we found no mention of the potential impact of overt observation on participants' behaviour.

There was also clear evidence that in most instances the coding of data a human operator was performing the coding and interpreting the results (in some instances, without formal coding procedures or checks for inter-rater reliability). In no case was there any indication that procedures were adopted to blind the observer to the condition from which the data were being extracted, so the possibility of bias, intentional or otherwise, cannot be eliminated.

Many of the studies adopted a *Wizard of Oz* technique (i.e., whereby participants interact with an interface without knowing that the responses are generated by a concealed human operator rather than a machine) in studying participants response to a robot. In most cases, the robot was controlled by an experimenter who possessed knowledge of the study's aims and hypotheses. However, we did not find any substantial discussion of how such a technique may introduce experimenter expectancies, which may threaten the validity of the results. In such instances where deception was used, some studies have used post-experimental questionnaires, which may alert participants to the use of deception (e.g., "Did you believe the robot was acting autonomously; [26]).

In none of the papers were there any references to the literature on experimenter effects, demand characteristics or other identified artefacts. So there is no evidence from the citations that any of the authors were aware of the possibility extraneous and confounding variables on the data. Any conclusions drawn for the experiments, therefore, could not consider alternative interpretations of why the human, or robot, behaved as it did, in line with factors that draw on expectations and behaviours outside of the experimental setting itself.

Four studies in the *Journal of Human-Robot Interaction* were more explicitly concerned with social psychological effects and the bibliographies for those papers are considered in more detail here. These papers, while sharing the same restrictions of the majority, did have more extensive bibliographies that explored issues within social psychology. This was the result of their exploring matters directly related to social psychological theories, which were considered pertinent to the interactions observed in the studies. But the papers still omitted consideration of the influence of systemic social artefacts in the experimental settings.

In one case, the possibility of inter-group relationships was considered [20]. This paper had, naturally, extensive reference to the literature on group relations, but no explicit reference to any bias factors. This possible oversight may be accounted for by the fact that a study was reported in the literature on inter-group relationships, which appeared to have discounted the role of demand characteristics in the interpretation of the results [4, 77]. However, the value of the paper, with the particular method used to account for demand characteristics, may be somewhat diluted due to the

very transparent manipulation, which could equally be accounted for by the introduction of an additional demand upon the participants to comply with experimental instructions. In a field such as inter-group relationships, extended to the reaction of robots as a social category, the epistemic effect of externally based expectations and biases may be fundamental and must be explicitly addressed.

In another case, the topic of research was the exploration of obedience to instructions provided by a robot versus a human [25], which explicitly linked with the famous and controversial research by Milgram [47]. While this link with the earlier research is appropriate, there is no acknowledgement of the recent developments in the interpretation of the Milgram results, many of which do turn on subtle features of the interaction with the human actors in the Milgram case and which need to be addressed in the robot studies (c.f. [31, 55, 60]).

In a third case, the study addressed factors of cognitive dissonance [41]. Here again, the enormous literature on dissonance theory, which has developed over more than 60 years after the theory was first promulgated [23], would seem to provide an adequate defence against any threats arising from demand characteristics (c.f. [18]). However, while dissonance theory has been perhaps the most powerful theory to come from the field of social psychology [30], major methodological criticisms from as far before as Chapanis and Chapanis [11] have never been adequately acknowledged [6] and the critical interpretation of dissonance experiments in terms of a demand characteristic such as evaluation apprehension [69, 70] still stands. Participants in any dissonance arousing setting may be responding to an expectation that their honesty is being assessed, and therefore, their morality is being evaluated. This motivation may lead them to behave in a manner consistent with the expectations of dissonance theory. This criticism has never been adequately countered. Modern researchers need to be aware of the limitations of methodology even if the theories that have been tested have continued on undistracted. Methods still can undermine the validity of theory.

The fourth example paper explores the potential of touching a robot in a “low accessibility” site of the anatomy to increase physiological arousal [42]. Touching a robot’s buttocks, breasts or thighs did increase arousal in the “offending” human participant. Such an experimental design,

however, carries with it virtually all of the demand characteristics that would lead a human participant to expect that such arousal was what the experimenters were expecting to find. Even if the participants believed that this was what was expected and they then tried not to be aroused, as a reaction to their belief, that inhibition would be likely to increase arousal yet further. No mention of such an artefactual outcome is considered.

4. Discussion

The analysis of a small but essentially representative set of papers in HRI reveals a pervasive pattern of experimental artefacts and bias in HRI experimental research. Additionally, there is seemingly a lack of consideration of the extensive literature on the effect of experimenter expectancies and participant expectations upon the outcomes of experimental studies.

The papers in the corpus examined reveal consistent failures to take account of potential bias in the manipulation of variables and in the measurement of outcomes. This leads us to consider what may the effect of such biases in the understanding or interaction with aliens, namely robots. At the most fundamental level, artefacts and associated tacit and intuitive (perhaps unconscious) influences upon the design of experiments, can bias the data derived from experiments which will, in turn, lead to biases in the theories and programs which are built into the robots that will emerge in the interactions with humans in the workplace of the future. Algorithm bias (e.g., [12, 50, 52, 79]) can result from the creation of variables in experimental studies that create variables with poor construct validity, as well as from the biases that may exist in the beliefs and actions of the creators of the algorithms using the data.

The emerging and developing interactions will in turn change and if the biases are not understood in the initial data, then they are unlikely to be detected and understood subsequently. As the algorithms change the more so the interactions arising from them will change. The discipline of experimental social psychology has demonstrated that human social behaviour can only be understood incompletely through the avenue of experiments, even with extensive knowledge about threats to validity. It is unlikely that the understanding of HRI will advance if there is ignorance of these threats.

4.1 Future recommendations for best-practice in the design of HRI experiments

The current exercise was akin to constructing an archaeological trench, taking a limited slice of the available data to get a layered view of the extant literature. While future researchers may find value in a comprehensive systematic 'excavation', the current 'snapshot' analysis has already revealed clear trends in in the conduct of HRI researchers similar to trends already discovered in the social psychological sciences. While in the extreme these trends reveal a need to reformulate the methodologies used to investigate HRI, as a starting point, we make several recommendations for HRI researchers embarking on the design of experimental studies:

- HRI researchers should consider the use of deception or active concealment in their instructions to participants. Participants who have no knowledge of the experimenter's hypothesis are less susceptible to demand characteristics. In cases where deception is used, researchers should take care when implementing post-experimental questionnaires, as participants may 'rationalise' their responses due to the introduction of demand characteristics (e.g., questions that may alert the participant to the possibility of deception are problematic).
- In line with the above, HRI researchers should employ a 'hypothesis check' with their participants during the debrief stage of the study. Here participants are asked to describe what they believe the aims and/or hypotheses of the study were. Where participants correctly deduce this information, the results of the study should be interpreted and reported with caution. A useful measure is Rubin's [73] Perceived Awareness of the Research Hypothesis (PARH) questionnaire, which should ideally be administered by a researcher not involved in the primary conduct of the experiment. The PARH is a 4-item scale that asks participants to rate the extent to which they believe that they are aware of the researchers' hypotheses. Significant correlations between participants' mean PARH scores with experimental effects may indicate the presence of demand characteristics. As we know that the degree of bias will differ across individual participants and studies, researchers should aim to

identify those particular participants who are problematic for a particular study, rather than making a global ruling of contamination. However, such measures are not without problem. It is possible that participants who had not previously discerned the purpose of the research will do so post-hoc, and untruthfully report a level of awareness.

- Where possible, HRI research participants should not be in the same room as an experimenter during data collection; or, if this is impossible, the experimenter in the room should be unaware of the condition the participant is in and/or the hypotheses that are being studied (i.e., double-blind conditions). Further, to avoid the possibility of sending subtle cues regarding their expectations, experimenters should not be tasked with delivering procedural instructions to participants, particularly in person. Nor should they be tasked with 'performing' the role of the robot in designs employing a Wizard of Oz technique.
- HRI researchers should try to conceal the nature of experimental manipulations from their participants. One example for doing so is the use of between-subjects over within-subjects designs where possible. However, researchers should note that while this design carries a range of additional benefits (e.g., avoidance of practice-effects), in using it the research forgoes a degree of statistical power in the analysis of results.
- If any coding is required that involves interpretation of the data, researchers should employ at least a second independent coder. This coder should be unaware of the research question(s) and should code the data based on a set of predefined criteria. Inter-rater reliability should be reported (to account for chance agreement, Cohen's Kappa [24] is recommended).
- Where possible, HRI researchers should avoid the use of pre-experimental measures, which may prime the participants to the issues under investigation, or worse, the aims of the experiment.

- Finally, HRI should place higher value on replications, establishing the effects through a different experiment (thus generalising it to other methods), replicating the same experiment in a different lab, and/or employing various research methods (i.e., a triangulation method).

Compliance with Ethical Standards:

This was not a funded study. The authors declare that they have no conflict of interest.

5. References

1. Barber, T.X., & Silver, M.J. (1968). Fact, fiction and the experimenter-bias effect. *Psychological Bulletin*, 70, 1-29.
2. Bargh, J.A., Chen, M., & Burrows, L. (1996). Automaticity of social behaviour: Direct effect of trait construct and stereotype activation in action. *Journal of Personality and Social Psychology*, 71, 230-244.
3. Baumeister, R.F., Vohs, K.D., & Funder, D.C. (2007). Psychology as the science of self-reports and finger movements: What happened to actual behavior? *Perspectives on Psychological Science*, 2, 396-403.
4. Berkowitz, N.H. (1994). Evidence that subjects' expectancies confound intergroup bias in Tajfel's minimal group paradigm. *Personality and Social Psychology Bulletin*, 20, 184-185.
5. Bless, H., & Burger, A.M. (2016). A closer look at social psychologists' silver bullet: Inevitable and evitable side effects of the experimental approach. *Perspectives on Psychological Science*, 11 (2), 296-308.
6. Brannigan, A. (2004). *The rise and fall of social psychology: The use and misuse of the experimental method.* New York: de Gruyn.
7. Brewer, W.F. (2012). The theory ladenness of the mental processes used in the scientific enterprise: evidence from cognitive psychology and the history of science. In R.W. Proctor & E.J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes*. New York: Oxford university Press.
8. Broad, E. (2018). *Made by humans: The AI condition*. Melbourne: Melbourne University Press.
9. Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312.

10. Campbell, D.T. (1994). The social psychology of scientific validity: An epistemological perspective and a personalized history. In W.R. Shadish & S. Fuller (Eds.) *The social psychology of science*. New York: Guilford. Pp.124-161.
11. Chapanis, N.P., & Chapanis, A. (1964). Cognitive dissonance: Five years later. *Psychological Bulletin*, 61, 1-22.
12. Clegg, B. (2017). *Big data: How the information revolution is transforming our lives.* London: Icon.
13. Collins, H. (2019). *Forms of life: The method and meaning of sociology*. Cambridge, Mass.: MIT Press.
14. Collins, H. (2018). *Artifictional intelligence.: Against humanity's surrender to computers*. Cambridge: Polity.
15. Collins, H. (2010). *Tacit and explicit knowledge*. Chicago: University of Chicago Press.
16. Collins, H., & Evans, R, (2007). *Rethinking expertise*. Chicago: University of Chicago Press.
17. Cook, T.D. & Campbell, D.T. (1978). *Quasi-experimentation: Design and analysis issues for field* settings. Chicago: Rand-McNally.
18. Cooper, J. (2007). *Cognitive dissonance: 50 years of a classic theory*. New York: Sage.
19. Crawford, J.T., & Jussim, L. (Eds.). (2018). *The politics of social psychology*, New York: Routledge.
20. Deliginais, C., Stanton, C., McGarty, C., & Stevens, C.J. (2017). The impact of intergroup trust and approach behaviour towards a humanoid robot. *Journal of Human-Robot Interaction*, 6, 4-20.
21. Duarte, J., Crawford, J., Stern, C., Haidt,J., Jussim, L. & Tetlock, P. (2015). Political diversity will improve social psychological research. *Behavioral and Brain Sciences*, 38, 1-58.
22. Doyen, S., Klein, O., Pichon, C-L., & Cleermans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *Plos One*, 7(1), article e 29081. Doi: 10.1371/journal.pone.0029081
23. Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

24. Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. 1st Edition, John Wiley & Sons, London
25. Friedman, N. (1967). *The social nature of psychological research*. New York: Basic Books.
26. Geiskkovitch, D.Y., Cormier, D., Seo, S.H., & Young, J.E. (2016). Please continue, we need more data: an exploration of obedience to robots. *Journal of Human-Robot Interaction*, 5, 82-99.
27. Gerard, H. (1999). A social psychologist examines his past and looks to the future. In A. Rodrigues & Levine, R. (Eds.), *Reflections on 100 years of experimental social psychology*. New York: Basic Books. (Pp. 47-81).
28. Grinnell, F. (2009). *Everyday practice of science: where intuition and passion meet objectivity and logic*. New York: Oxford University Press.
29. Hart, C.W. (1947). Some factors affecting the organization and prosecution of given research projects. *American Sociological Review*, 12, 514-519.
30. Harmon-Jones, E., & Mills, J. (1999*). Cognitive dissonance: Progress on a pivotal theory in social psychology.* Washington, DC: American Psychological Association.
31. Haslam, S.A., Reicher, S.D., & Millard, K. (2015). Shock treatment: Using immersive digital realism to restage and re-examine Milgram's 'Obedience to Authority' research. *Plos One,* DOI:10.1371/journal.pone.0109015.
32. Hudson, L. (1972). *The cult of the mind*. London: Jonathan Cape.
33. Inbar, Y. & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspectives on Psychological science*, 7 (5), 496-503.
34. Innes, J. M. (2005). Decline of fact in artefact: Loss of control in social psychological studies. *Australian Journal of Psychology Supplement*, 57, 89.
35. Innes, J.M., & Fraser, C. (1971). Experimenter bias and other possible biases in psychological research. *European Journal of Social Psychology,* 1,297-310.
36. Innes, J.M. & Morrison, B. W. (2017). Projecting the future impact of advanced technologies on the profession: Will a robot take my job? *Australian Psychological Society InPsych*, 39 (2), 34-35.

37. Kintz, B.L., Delprato, D.J., Mettee, D.R., Persun, C.E., & Schappe, R.H. (1965). The experimenter effect. *Psychological Bulletin*, 63, 223-232.
38. Klein, O., Doyen, S., Leys, C., daGama, P.A., Miller, s., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioural experiments. *Perspectives on Psychological Science*, 7(6), 572-584.
39. Knapp, R.H. (1963). Demographic, cultural and personality attributes of scientists. In C.W. Taylor & F. Barron (Eds.). *Scientific creativity*. New York: Wiley.
40. Lana, R.E. (1999). Pretest sensitization. In R. Rosenthal & R.L. Rosnow (Eds). *Artifact in behavioural research*. New York: Academic Press. Pp. 121-146.
41. Levin, D.T., Harriott, C., Paul, N.A., Zhang, T., & Adams, J.A. (2013). Cognitive dissonance as a measure of reactions to human-robot interaction. *Journal of Human-Robot Interaction*,2, 1-17.
42. Li, J., Ju, W., & Reeves, B. (2017). Touching a mechanical body: Tactile contact with body parts of a humanoid robot is physiologically arousing. *Journal of Human-Robot Interaction*,6, 118-130.
43. McAdams, D.P. (2015). *The art and science of personality development*. New York: Guilford.
44. MacCoun, R.J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, 49, 259-287.
45. McGuigan, F.J. (1963). The experimenter: A neglected stimulus object. *Psychological Bulletin*, 60, 421-428.
46. McGuire, W.J. (1969). Suspiciousness of experimenter's intent. In R. Rosenthal & R.L. Rosnow (Eds). *Artifact in behavioural research*. New York: Academic Press. Pp. 13-60.
47. Milgram, S. (1974). *Obedience to authority*. New York: McGraw-Hill.
48. Miller, A.G. (1972). *The social psychology of psychological research*. New York: Free Press.
49. Miller, N., & Pollock, V.E. (1994). Meta-analysis and some science-compromising problems of social psychology. In W.R. Shadish & S. Fuller (Eds.) *The social psychology of science*. New York: Guilford. Pp. 230-261.
50. Muller, J.Z. (2018). *The tyranny of metrics*. Princeton: Princeton University Press.

51. O'Gorman, F. (2017). *Forgetfulness: Making the modern culture of amnesia*. London: Bloomsbury.
52. O'Neil, C. (2016). *Weapons of math destruction*. London: Penguin.
53. Orne, M.T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics. *American Psychologist*, 17, 776-783.
54. Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R.L. Rosnow (Eds). *Artifact in behavioural research.* New York: Academic Press. Pp. 143-179.
55. Passinin. S., & Morselli, D. (2010). The obedience-disobedience dynamic and the role of responsibility. *Journal of Community and Applied Social Psychology*, 20, 1-14.
56. Pastore, N. (1949). *The nature-nurture controversy*. New York: King's Crown Press.
57. Pepitone, A. (1999). Historical sketches and critical commentary about social psychology in the golden age. In A. Rodrigues & Levine, R. (Eds.), *Reflections on 100 years of experimental social psychology*. New York: Basic Books. (Pp. 170-199).
58. Polanyi. M. (1958). *Personal knowledge*. London: Routledge.
59. Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.
60. Reicher, S.D., Haslam. S.A., & Smith, J.R. (2012) Working toward the experimenter: Reconceptualizing obedience within the Milgram paradigm as identification-based followership. *Perspectives on Psychological Science*, 7, 315-324.
61. Robinson, K.A., and S.N. Goodman, S.N. (2011). A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Annals of Internal Medicine*, 154, p. 50.
62. Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Scientist*, 51, 268-283.
63. Rosenthal, R. (1966). *Experimenter effects in behavioural research*. New York: Appleton-Century-Crofts.

64. Rosenthal, R. (1969). Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R.L. Rosnow (Eds). *Artifact in behavioural research.* New York: Academic Press. Pp. 181-277.
65. Rosenthal, R. (1976). *Experimenter effects in behavioural research (enl.ed.)* New York: Irvington.
66. Rosenthal, R. (1994). On being one's own case study: Experimenter effects in behavioural research-30 years later. In W.R. Shadish & S. Fuller (Eds.) *The social psychology of science*. New York: Guilford. Pp. 214-229.
67. Rosenthal, R., & Rosnow, R.L. (1969). The volunteer subject. In R. Rosenthal & R.L. Rosnow (Eds). *Artifact in behavioural research*. New York: Academic Press. Pp. 61-120.
68. Rosenthal, R., & Rubin, D.B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377-386.
69. Rosenberg, M. J. (1965). When dissonance fails: On eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology*, 1, 18-42.
70. Rosenberg, M.J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R.L. Rosnow (Eds). *Artifact in behavioural research*. New York: Academic Press. Pp. 280-350.
71. Rosenzweig, R. (1933). The experimental situation as a psychological problem. *Psychological Review*, 40, 337-354.
72. Rosnow, R.L., & Suls, J.M. (1970). Reactive effects of pretesting in attitude research. *Journal of Personality and social Psychology*, 15, 338-343.
73. Rubin, M. (2016). The Perceived Awareness of the Research Hypothesis Scale: Assessing the influence of demand characteristics. Figshare. doi: 10.6084/m9.figshare.4315778
74. Sadler, M., & Regan, N. (2019). *Game changer.: AlphaZero's ground breaking chess strategies and the promise of AI*. Alkmaar: The Netherlands.
75. Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalised causal inference*. Boston: Houghton-Mifflin.

76. Sherwood, J.J., & Nataupsky, M. (1968). Predicting the conclusions of negro-white intelligence research from biographical characteristics of the investigator. *Journal of Personality and Social Psychology*, 8, 53-58.
77. St. Claire, L., & Turner, J.C. (1982). The role of demand characteristics in the social categorization paradigm. *European Journal of Social Psychology*, 12, 307-314.
78. Toomela, A., & Valsiner, J. (Eds.). (2010). *Methodological thinking in psychology: 60 years gone astray?* Charlotte, NC: Information Age Publishing.
79. Wagner, A.J., Borenstein, J., & Howard, A. Overtrust in the robotic age: A contemporary ethical challenge. *Communications of the ACM* 61(9), 22-24.
80. Walsh, T. (2018). *2062.* Melbourne: LaTrobe University Press.
81. Wilson, T.D., Aronson, E. & Carlsmith, K. (2010). The art of laboratory experimentation. In S.T. Fiske, D.T. Gilbert & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1, pp. 51-81). Hoboken: John Wiley & Sons.
82. Zajonc, R.B. (1999). One hundred years of rationality assumptions in social psychology. In A. Rodrigues & Levine, R. (Eds.), *Reflections on 100 years of experimental social psychology*. New York: Basic Books. (Pp. 200-214).
83. Zimbardo, P.G. (1999). Experimental social psychology: Behaviorism with minds and matters. In A. Rodrigues & Levine, R. (Eds.), *Reflections on 100 years of experimental social psychology*. New York: Basic Books. (Pp. 135-157).