Investigating Trust Factors in Human-Robot Shared Control: Implicit Gender Bias Around Robot Voice

Alexander Wong[1] and Gregory Dudek[2]

Abstract— **We extend a prior study of human-robot shared control (when a user is able to take over steering control of an autonomous agent) to investigate factors of robot behaviour that can influence trust. Our study compares agents with and without human voice indicators of uncertainty and evaluates differences in trust with inferred and introspective methods. We find that a human's trust in a robot can be influenced by verbal feedback from the robot agent. Specifically, people tend to lend more trust to agents whose voice is of the same gender as their own.**

I. INTRODUCTION

As electronic machinery and artificial intelligence methods are growing increasingly sophisticated and substituting human ability, our interactions with robots are becoming more frequent and demanding on both the robot and the user. How can we best design our machines to most effectively fulfill their role as a supplement or extension of human ability?

The demands on robotic machinery are evident as development of hardware and software accelerates to satisfy commercial needs, scientific inquiry, and creative interests. But the demands on the users (i.e. humans) are not all obvious; one might think that as robots gain more autonomous capabilities, human workload will decrease. However, we are more likely to see human workload redefined. Robots are designed to execute increasingly important tasks. In doing so, robots and their designers assume more responsibility. Consequently, users - the humans with whom robots interact with - will adopt more supervisory roles that call for the appropriate amount of trust in robots in order to collaborate effectively.

Trust is a key component of effective collaboration in any task. This experiment observes the importance of trust in human-robot interaction (HRI). Unfortunately, trust is a *feeling* (sometimes an unconscious one), and thus there is no objective or empirical way to measure it; at best, trust can be inferred and introspected. So, this study partially depends on people's self-reported feedback to measure trust, and partially infers trust with certain proven methods.

We utilize both self-reported and inferred methods of trust evaluation to investigate the factors of trust that can affect collaborative interactions between humans and robots; the class of interactions under investigation is of supervisor-worker (human-robot) relationships. We did this with a "Wizard of Oz" experiment in which participants supervised a simulated drone in several boundary tracking task scenarios. The trust factor (and independent variable) in the experiment is the presence or absence of an audio cue (triggered by a researcher - the "wizard") as an indicator of uncertainty; specifically, the three test cases are a baseline of no audio cue and two voice cues: one male, one female.

We hypothesize that a worker's verbal indicator of uncertainty will affect the trust a supervisor has in the worker. In the context of this study's experiment, this is manifested in the presence of a human voice audio indicator of uncertainty (in the simulated environment) that will affect the trust a supervisor has in the worker. Furthermore we are interested in whether the gender of the verbal feedback (i.e. the implied gender of the worker robot) has any influence on the supervisor's trust.

II. RELATED WORK

Xu and Dudek previously demonstrated the improvement trust-aware robots bring to efficient collaboration by applying a mathematical model: the Online Probabilistic Trust Inference Model (OPTIMo) [1]. This study aims to investigate factors extrinsic to Xu and Dudek's mathematical definition yet intrinsic to most human interaction: voice and gender.

In 2008, Walters et al. compared a robotic (neutral synthesized) voice to natural, gendered human voices and found an initial hesitation by humans when interacting with robotic-sounding robots [2]. However, the comfortable distance kept by humans from a robot was shown to decrease with subsequent encounters with the robot. A decade later, humans are interacting with artificial voice agents more frequently through personal devices and start home and assistant technologies. Are stigmas changing as robot personas become integrated with day-to-day life?

A 2012 experiment by Eyssel et al. indicated an in-group gender bias for psychological closeness [3]. This means that people tended towards feeling more positively towards a same-gendered robot (in this case gender being defined by the robot's voice). This tendency is not an deliberate preference held by individuals; rather, it is an implicit bias towards agents exhibiting certain characteristics.

Implicit gender bias is an automatic association humans make with another's gender and it is progressively surfacing as an important factor in all arenas such as in the workplace, legislature, media [4], and, we postulate, in HRI.

[1] Alex Wong is a Cognitive Science graduate of McGill University and an alumnus of the Mobile Robotics Lab (MRL) of the McGill Research Centre for Intelligence Machines (CIM), Montreal, QC, H3A 0E9, Canada. awong@cim.mcgill.ca

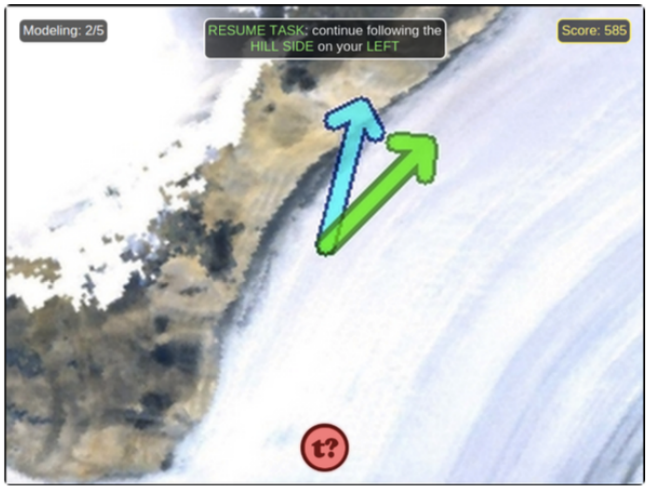
[2] Gregory Dudek is a Professor of Computer Science at McGill University and the director of MRL, CIM, Montreal, QC, H3A 0E9, Canada. dudek@cim.mcgill.ca

III. METHODS

This study is a continuation of the work done by Xu and Dudek, therefore the experiment design and procedures follow many of the same specifications. This section details the important features and modifications needed to understand this study. More details on vision-based boundary tracking, trust modelling, and efficient collaboration can be found in [1], [5], [6].

A. Infrastructure and Interface

The experiment composes of a graphical user interface (GUI), audio instructions, and a gamepad used by test subjects to exert supervision in three ways: manually taking over steering, training through steering, and by providing critiques.



RESUME TASK: continue following the HILL SIDE on your LEFT

Fig. 1. Graphical User Interface (GUI) [1]

The GUI (Fig. 1) features a simulated boundary tracking robotic agent (an aerial drone) in a mixture of environments. It shows experiment subjects a video feed of the agent with indicators for both the human's and agent's heading, with the agent's heading depicted throughout and the human's steering commands only depicted during periods of manual intervention. The persistence of the agent's heading is intended as an aid to the supervisor as to when yield steering control back to the agent. The GUI also displayes the agent's present task, a session score metric, occasional prompts to the supervising human to provide trust critiques, and, a study progress indicator.

Each scenario has a set plan of tasks that are only made known to the supervisor; the drone agent is naive. These tasks, such as "follow the coastline" or "turn left at the highway", are conveyed through the GUI and dictated using a non-neutral synthetic speech engine. Though this speech was synthetic, it is clearly a female voice. All participants wear an audio headset during the experiment to receive these instructions, as well as the audio cues.

Participants use the gamepad to steer (using an analog joystick which allow moderately fine controls) and three buttons to convey increases, decreases, or persistence of trust to the agent. These are intended to be salient representations of a supervisor's trust-state, and are symbolized as `t+`, `t-`, and `t=`. The video feed depicts a top-down aerial view of the agent over a given terrain: a two-dimensional image. The drone agent flies at a constant velocity and altitude; hence, the supervisor's steering controls are only to head leftward or rightward (parallel to the plane of the terrain).

B. Modifications: The Voice Agent

The main modification to this study has been inclusion of audio cues that indicate agent's states of uncertainty. Introducing the voice agent: a modified conservative agent that emits the words "I'm uncertain" in moments of oscillating heading in conjunction with making a mistake. There are many ways to define an autonomous agent's uncertainty, which means our definition of is far from robust; however, we accepted it as a sufficient condition to correlate the independent and dependent measures since it is an intuitive visual representation of uncertain behaviour.

Initial implementations intended to fully automate voice agents' audio cues (i.e. only trigger under certain conditions within the interaction sessions). However, complex situational contexts of fairly naive drone supervisors steering and training agents along different environments were inhibitory to successfully implementing fully automated voice agents. The compromised solution was to use a "Wizard of Oz" experiment technique where, unbeknownst to human participants, the audio cues were triggered by the researcher ("wizard") conducting the experiment. As explained in the trial procedure section, participants were led to assume the motivation for the voice cue was internal to the experiment interface. Consequently, deception became a necessary component to the psychological design of the experiment.

Unlike agents in previous iterations of experiments using this infrastructure, the voice agents were all initialized with the same hyper-parameters. This means their boundary tracking, learning, and trust-inference capabilities were consistent for all interactions. Hence, it is implied that any variation of behaviour and performance of voice agents is purely the result of supervisory inputs of the human participants and their trust in each agent.

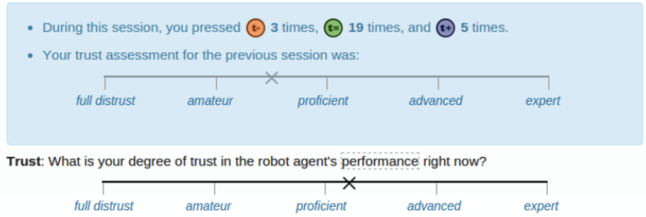
The only difference in voice agents was the gender of the voice recording: we implemented a male and a female agent. The choice of content in the audio was intentional so as to give the voice agent a human touch. Voice cues for male and female voice agents were recorded from a male identifying and a female identifying human rather than generated from a synthetic speech engine. Also, the inclusion of the "I'm" in "I'm uncertain" personified the agent by implying a sense of self-awareness.

The audio cues were recorded in a systematic method to reduce the chance of tonal variation, which otherwise may have subtly biased a human participant. In the recording process, voice actors read the entire sentence "I would share my solutions with you but I'm uncertain that they are correct." From the recording, the leading and trailing audio was cropped; only the desired audio cue remained while still honouring the context in which it was uttered.

C. *Trust Metrics and Dependent Measures*

Many details from each experiment trial, such as the agent's heading and pose, were logged on a frame-by-frame basis as part of the software design, but we selected four key metrics to collect to infer a participant's trust.

Two metrics count the supervisors' **acts of trust**: a "decision and behaviour of relying upon another individual's abilities [1]." One is each session's score, a measure of coverage progress that increments in larger steps in periods when participants did not override an agent's steering. We used this metric assuming that taking over steering is an act of distrust, and thus trust can be subjectively quantified as higher when the drone is given more autonomy. The other metric is the trust critiques of the agent given by the human supervisor. As described previously, trust critiques are salient reflections of a supervisor's trust state. Records of these critiques can be used to infer changes in trust, and can be summed (as `t+`, `t-`, and `t=` are recorded as `+1`, `-1`, and `+0` respectively) to reflect total trust along a session.



- During this session, you pressed t- **3** times, t= **19** times, and t+ **5** times.
- Your trust assessment for the previous session was:

Trust: What is your degree of trust in the robot agent's performance right now?

Fig. 2. Post-Session Trust Feedback Scale [1]

The two other metrics measure a supervisor's **degree of trust**: "a quantifiable subjective assessment towards another individual [1]." One is self-reported trust feedback where participants provide a trust assessment upon completion of every session. The format is a continuous scale from full distrust to expert (Fig. 2). While providing this feedback, supervisors were shown their trust critiques for the session and their response to the most recent trust feedback so that they can report on the session as a whole and in reference to their prior trust assessment. The last metric is a questionnaire of three trust attributes which participants complete after all three interactions with an agent. Here, participants used the same continuous scale to assess the agent on performance, adaptability, and collaboration.

D. *Trial Procedure*

Each experiment trial ran for approximately 40 minutes as follows.

First, test participants were given an introduction to the experiment via consent form that summarized the general motivations for the study. Next, participants were given verbal (from the investigating researcher), printed (on the screen), and visually aided (with images and diagrams) instructions about how to execute their role as supervisor. We included some pointers to disambiguate any areas that might be misinterpreted to avoid adding noise to the results. Here are some key points that were emphasized to participants:

- Think of trust as "how much do I trust this drone to do its job" and not as "how much do I trust this drone to follow my instructions"
- Each agent learns from your steering interventions, but that learning is reset when a new agent is introduced
- Feel free to give trust critiques as often or infrequently as you like, but keep in mind that if you do not give a critique for five seconds, the GUI will prompt you for one
- Supervise each drone independent of your experience with another and void meta-gaming each map

Participants then completed three practice sessions for the purpose of familiarizing each participant with the controls, interface, and drone agent's behaviour. Following this, participants completed a questionnaire about demographic information, and experience and bias a priori. Next, participants were briefed with further details regarding the upcoming experiment phase. Emphasized during this was that they would be supervising three different agents, which, among other things, were implied to be parameterized to communicate uncertainty differently. This was intentionally deceptive to mask the independent variable of the study, but important to ensure participants had the appropriate interpretation of the audio cues they would be receiving.

Subsequently, participants participated in three sets of three experiment sessions. Each session featured one of three maps, and each set of three featured a different agent: one baseline agent with no audio cue, and two voice agents: one male, one female. The agent order was randomly selected for each participant.

The three maps were of a highway, a coastline, and a hybrid of the two, which were always presented in this order. The highway was characterized by having poorly defined boundaries but otherwise fairly straight paths for the agent to follow. The coastline was characterized by having very clearly defined boundaries but also very winding paths. The hybrid map presented a mixture of these macro-features. After each session, participants completed a trust feedback questionnaire; after all three sessions per agent, participants completed the longer agent-questionnaire inquiring about trust in the agent's performance, collaboration, and adaptability.

Throughout the experiment sessions, the "wizard" sat approximately four meters behind the human participant. From this position the "wizard" was close enough to interpret agent behaviour effectively, while being far enough so as to not be able to see which agent was active in the session and thus eliminating potential biases as they triggered the audio cues for each voice agent.

Finally, after all 12 sessions, each between 60-85 seconds long, participants completed a final questionnaire on the experiment's mental, physical, and temporal demand, any frustrations experienced, and any other feedback they had on the study.

IV. RESULTS AND ANALYSIS

The sample size was small totalling 14 participants (8 male, 6 female). Thus, results are inconclusive as to whether or not to validate the hypothesis. This is especially true given that comparison of agents using a continuous scale metric is subjective to the test subject, and also a relative comparison. Nevertheless, we did find interesting patterns in the data that warrant further investigation. Of the 14 experiment participants, one male's trial had to be withdrawn from the data set due to audio malfunction.

Participants were either current or former undergraduate students from a range of fields of study including physics, computer science, medical sciences, psychology, music performance, and more. Participants' ages ranged between 18 and 25, all reported having less than a year of robot programming experience, and on a scale from 0 to 10, an average of 5.5 for proficiency at driving cars, 6.3 for comfort level in a self-driving car, 3.7 for experience teleoperating robots, and 3.7 experience with a gamepad.

A. Aggregate of all Maps or Attributes

We analyzed the study data by comparing results visually and statistically with a Friedman χ^2 test. This test was selected due to the data being continuous repeated measurements. First we looked at the aggregate of all sessions across all maps (or across three attributes in the case of the agent-questionnaire). Of the four trust metrics, neither of the inferred measures of trust (sum of trust critiques per session and session score) exhibited any trends graphically nor revealed any significant statistical variation of trust between the three agents. As for the two introspected measures of trust (post-session trust feedback and post interaction questionnaire), we observed a noticeable difference in trust for the agent with a female voice compared to the baseline and male-voice agent. However, these differences were not statistically significant (trust feedback: χ^2=0.56, p$\leq$0.75, agent-questionnaire: χ^2=0.66, p$\leq$0.71).

B. Analysis by Map or Attribute

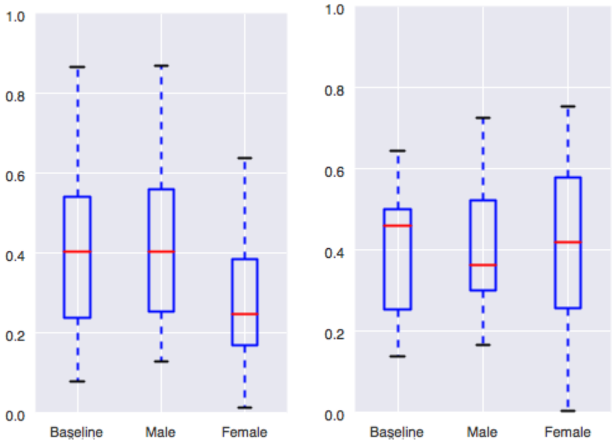
We then broke down the data by map (or attribute in the case of the agent-questionnaire) and found varying preferences to trust different agents depending on the map/attribute. On the highway map, participants consistently exhibited less trust in the baseline agent versus the voice agents. There was no pattern of trust on the coastline or hybrid maps.

Of the three trust attributes in the questionnaire, the female-voice agent was evaluated lowest for performance and adaptability, consistent with the aggregate results, but not significantly.

C. Filtering by Participant Gender

We next filtered the results by gender of the supervising participant. Here we found significant bias between both groups to preferentially trust agents of the same gender, but only by some metrics.

The aggregate of three attributes in the agent-questionnaire revealed the most convincing evidence of in-group bias. Male participants were less trusting of female voice agents[1] while female participants were more trusting of the female voice agent[2]. In both cases, the baseline and male voice agent were almost the same - only trust in the female voice agent varied. These findings were consistent with male participant assessments for adaptability and collaboration, but not performance (Fig. 5). Meanwhile, they were consistent with female participant assessments for performance and collaboration, but not adaptability (Fig. 4). These results were significant for the aggregate data filtered by gender (male: χ^2=6.000, p$\leq$0.049, female: χ^2=8.111, p$\leq$0.017), but only significant among attributes for collaboration as assessed by female participants (χ^2=8.333, p$\leq$0.016).



(a) Male Participants (b) Female Participants

Fig. 3. Trust Feedback Filtered by Participant Gender

Trust feedback results in Fig. 3 show - albeit insignificantly - male participant's tendency to distrust female voice agents in comparison to both baseline and male voice agents (χ^2=5.073, p$\chi^2$0.079), but an equivalent distrust in the male voice agent is not reciprocated among female participants. While the prior finding is not statistically significant, its p-score is close to less than 0.05 and therefore worthy of further investigation as will be elaborated on in the future directions section of the conclusion.

We found no pattern of trust as implied by session scores or trust critiques, even when filtered by gender.

D. First Agent Only

In anticipation of possible introduction of bias in trust salience after sequential interactions with the same maps, we conducted the same analysis but limited to only the first agent participants first interacted with. This considerably shrunk the already small dataset to four sets of three sessions per baseline and female agents, and five sets of three sessions per female male agent. However, the data had extremely high error margins and we found no noteworthy patterns.

V. DISCUSSION

The results suggest that the presence of a voiced audio cue from a worker does affect a supervisor's trust, but this effect

[1] $\mu, \sigma_{baseline}$=0.58, 0.18, μ, σ_{male}=0.58, 0.18, μ, σ_{female}=0.41, 0.20
[2] $\mu, \sigma_{baseline}$=0.56, 0.18, μ, σ_{male}=0.59, 0.16, μ, σ_{female}=0.63, 0.24

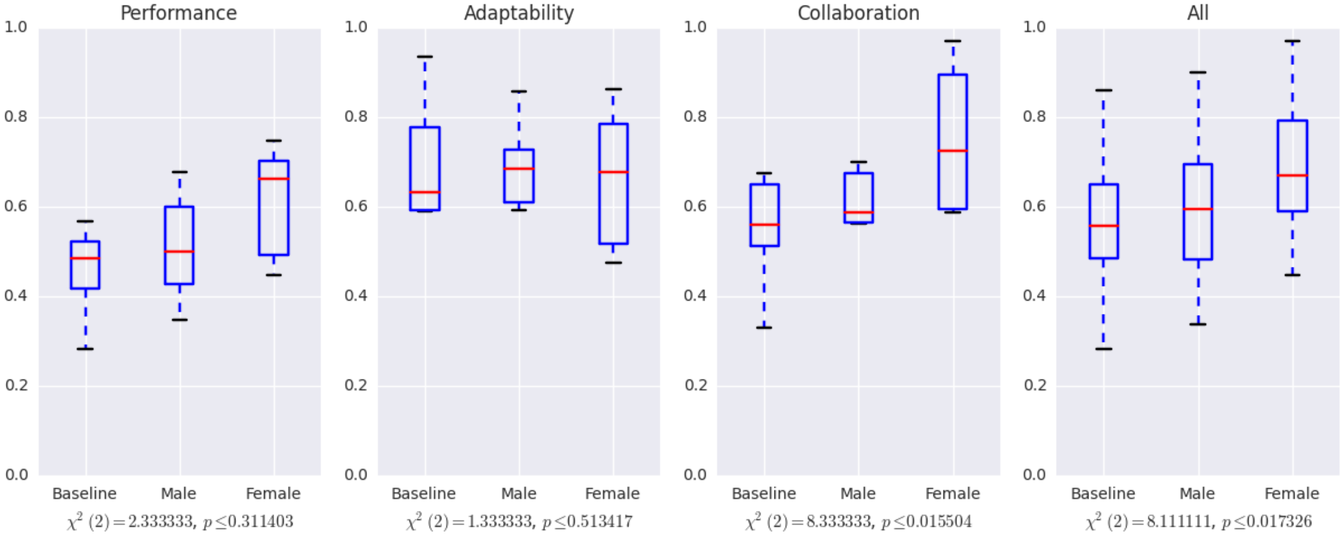


Fig. 4. Agent-Questionnaire Filtered for Female Subjects

is limited by the genders of the supervisor and the worker's voice, as there seems to be an in-group bias to trust. However, this alone does not verify this study's main hypothesis. Our results looked at two categories of data: measurable acts of trust and introspected degrees of trust. Findings from the latter category would be more convincing if they were corroborated by the former since they were found to be correlated in previous work by Xu and Dudek [1] where their trust model was able to predict degree of trust based on acts of trust alone. A key difference in the studies is this study did not utilize the trust modelling from previous work; voice agents were minimally designed, modified from conservative agents to be as close to a pure boundary tracker as possible. So, considering the significant or close to significant findings of this study of fairly small sample size, we recommend further, more controlled, investigation.

We found difference in trust, when filtered by gender of supervisor, to be characterized in two ways. If baseline trust was similar to in-group trust, it suggests distrust for the out-group specifically. Comparatively, if baseline trust was similar to out-group trust, it suggests increased trust for the in-group only. This is an important distinction because male participants seemed to exhibit mostly the first case (out-group distrust) while female participants seemed to exhibit mostly the second case (increased in-group trust). As described in the results and analysis section, these are the significant findings of this study.

We collected feedback from participants through the post-experiment debriefing questionnaire and conversations held after trials. From these, it is clear that individuals' various innate trust in intelligent machinery, differences in experience with autonomous robotics, gamepad competence, and personal expectations play a role in their trust awareness in a supervisor-worker relationship, especially with a non-human worker agent. Much of the additional feedback on the study made sense, as they form a valid basis for trust in human-human interactions.

Common feedback from participants was that increased familiarity with each agent (by having more interaction time) would indirectly affect trust as a supervisor learns an agent's strengths and limitations. This aligns with the observations made by Walter et al. [2].

The most salient feedback given from some participants contradicted the expectations from our hypothesis: they would prefer less human-likeness in a worker robot, and therefore trust it less for having human-like features. While our hypothesis does not specify whether voice agents would positively of negatively affect trust, this proposed aversion to human likeness contradicts the Walter et al. findings. It does not, however, preclude the role of implicit gender bias in HRI.

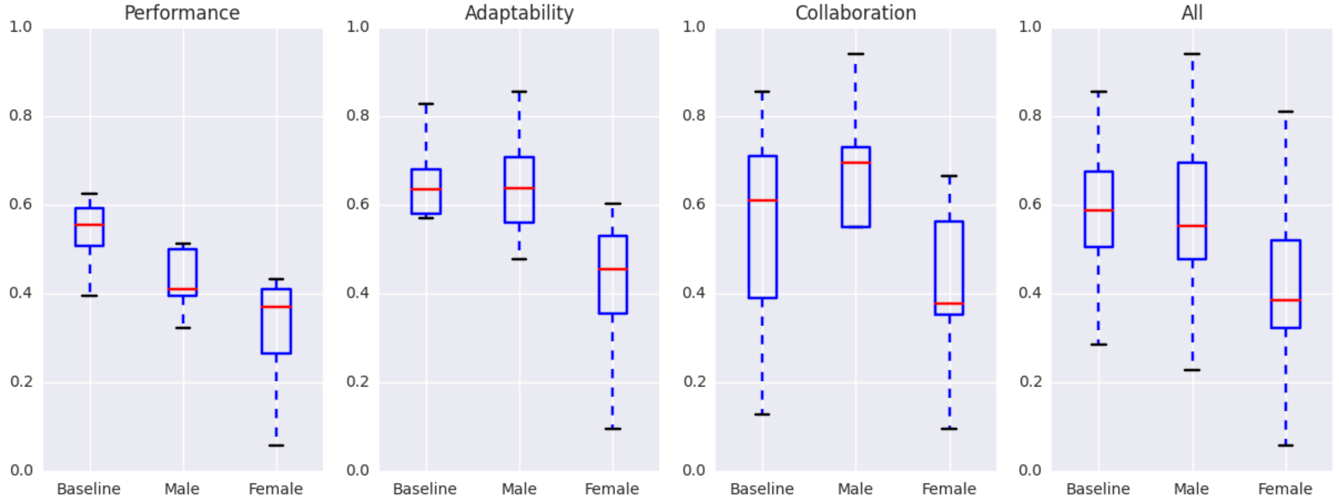
From participants' feedback and the observed tendency for in-group bias found in the results, we can extrapolate the following conclusions:

- Human supervisors of robot workers are likely to have a preference to trust - and therefore collaborate effectively - with a robot agent that has a voice (and possibly personality) of the same gender expression.
- Trust is likely to be dependent on familiarity with a given robot's ability (as it would also be between humans).
- Users/supervisors of robots should be given ample options to adjust parameters of trust factors (such as voice gender), in order to personalize and optimize their collaborative experience.

VI. CONCLUSION

A. Limitations

The most complicated limitation of this study is the broad subjectivity of the data measurement due to implicit biases of individual participants. Despite concerted efforts to provide clear and plentiful instructions to participants, there are still many ways that individuals can form their own approach to the task of supervision. This creativity is promising as it shows active interest among people to collaborate with



$\chi^2(2) = 5.428571,\ p \leq 0.066252$ $\chi^2(2) = 2.571429,\ p \leq 0.276453$ $\chi^2(2) = 2.000000,\ p \leq 0.367879$ $\chi^2(2) = 6.000000,\ p \leq 0.049787$

Fig. 5. Agent-Questionnaire Filtered for Male Subjects

robotic agents. However, it does pose obstacles to the scientific process.

There is room for further statistical analysis of this study's results. Other than the Friedman χ^2 test, there are certainly more statistical methods that can be applied, such as the Kemeny-Young scheme, using stepwise regression to calculate covariance and statistical significance of the relatedness of the data. This is particularly true in light that many of the recorded metrics were not fully evaluated, such as the temporal distribution of trust critiques versus audio cues. Our study also observes a small sample size, limiting clarity and definitiveness in data analysis.

B. *Future Directions*

The challenges of conducting this study have involved many unpredictable or uncontrollable loose ends of controlling the supervision task. There are so many angles to dissect issues within HRI, and this is promising for the field because they pose questions with answers that can and should be pursued by integrating many fields beyond computer science and electrical engineering, such as psychology, behavioural economics, and sociology, to name a few.

Future work should at least attempt to refine and re-conduct this study to verify the interesting observations have been made about in-group gender bias. This study infrastructure logs far more data than was discussed in this report, and another study with a larger sample size, more rigorous statistical analysis, and more time to invest could take advantage of this.

The "I'm uncertain" audio cue was designed partially to be able to compare to other audio or non-audio cues. This is just a sliver of all the potential trust factors worthy of scientific inquiry, and many of them could be tested with this interface.

One final note is to remind readers that a robot is any intelligent machine capable of acting on or interacting with the physical world. It could be the semi-humanoid C-3P0 archetype, as well as a car, drone, phone, or watch. So for every mention of HRI in this report, the same applies to human-computer interaction, human-machine interaction, and any other synonymous term.

ACKNOWLEDGMENT

We would like to thank all those who participated in the study. Also, much appreciation is deserved by members of the McGill Research Centre for Intelligent Machine's Mobile Robotics Lab, for guidance and friendliness throughout. A particularly special thanks to J. K. A. and J. W. for their voice-acting talents. Most importantly, the principal investigator, Alex Wong, is grateful to his supervisor, Professor Dudek, for inspiring him to pursue his interests in robotics.

REFERENCES

[1] A. Xu and G. Dudek, "Maintaining efficient collaboration with trust-seeking robots," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 3312–3319.

[2] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. Te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE, 2008, pp. 707–712.

[3] F. Eyssel, L. De Ruiter, D. Kuchenbrandt, S. Bobinger, and F. Hegel, "If you sound like me, you must be more human: On the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. IEEE, 2012, pp. 125–126.

[4] C. Pelletier, J. Ashta, A. Jang, Y. Hitti, and I. Moreno, "Biasly ai." [Online]. Available: https://sites.google.com/view/biasly/home

[5] A. Xu and G. Dudek, "A vision-based boundary following framework for aerial vehicles," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'10)*, 2010, pp. 81–86.

[6] A. Xu and G. Dudek, "Towards modeling real-time trust in asymmetric human-robot collaborations," in *Int. Symposium on Robotics Research (ISRR)*, 2013.