Investigating Trust Factors in Human-Robot Shared Control

Undergraduate Honours Thesis 2018
Principle Investigator: Alexander Wong
McGill University Cognitive Science
Montreal, Canada
Email: awong@cim.mcgill.ca

Supervisor: Gregory Dudek
McGill University Dept. Computer Science
Centre for Intelligent Machines - Mobile Robotics Lab
Montreal, Canada
Email: dudek@cim.mcgill.ca

***Abstract*—We extend a prior study of human-robot shared control (when a user is able to take over control of an autonomous agent) to investigate the 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 but this is dependent on the human's gender identity and their perception of the robot's voice's gender. Evaluation of multiple of the study's trust metrics reveal a tendency for people 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, human interactions with robots are becoming more frequent and demanding on both the robot and the user. Keep in mind 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, but also a car, drone, phone, or watch.

The demands on robotic machinery are evident as development of hardware and software is accelerating to satisfy commercial needs, scientific inquiry, and creative interest. But the demands on the users (ie. humans) are not all obvious; one might think that as robots gain more autonomous capabilities, human workload will decrease. However, the reality is that as robots gain intelligence, there will be more interactions between humans and robots. More importantly, robots will be designed to execute increasingly important tasks. In doing so, robots and their designers will need to assume more responsibility. And 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 in effective collaboration in any task. In this report we study the importance of trust in human-robot interaction. Unfortunately, trust is a feeling (aware or unaware) and thus there is no objective or empirical way to measure it; at best, trust can be inferred and introspected. So, in part, we depend on people's self-reported feedback to measure it. However, there are also proven models with which trust can be inferred.

In this paper, we utilize both 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 (ie. the implied gender of the worker robot) has any influence on the supervisor's trust.

II. Methods

This study is a continuation of the work done by Xu and Dudek, so 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]–[3].

A. Infrastructure and Interface

The original infrastructure of the experiment composed 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.

The GUI (Fig. 1) featured a simulated boundary tracking robotic agent (an aerial drone) in a mixture of environments. It showed 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

Fig. 1. Graphical User Interface (GUI) [3]

intervention. The persistence of the agent's heading was intended as an aid to the supervisor as to when yield steering control back to the agent. The GUI also displayed 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 had a set plan of tasks that were only made known to the supervisor; the drone agent was uninformed. These tasks, such as "follow the coastline" or "turn left at the highway" were conveyed through the GUI and dictated using a synthetic speech engine. Though this speech was synthetic, it was clearly a female voice. All participants wore an audio headset during the experiment to receive these instructions, as well as the audio cues.

Participants used the gamepad to steer (using an analog joystick which allowed moderately fine controls) and 3 buttons to convey increases, decreases, or persistence of trust to the agent. These were intended to be salient representations of a supervisor's trust-state, and were symbolized as `t+`, `t−`, and `t=`. The video feed was a top-down aerial view of the agent over a given terrain: a 2-dimensional image. And, the drone agent was simulated to fly at a constant velocity and altitude; hence, the supervisor's steering controls were 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 (ie. only trigger under certain conditions within the interaction sessions). However, complex situational contexts of fairly nave 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.

The researcher followed strict specifications as to when to trigger an audio cue. [list here or in trial procedure?]

Unlike agents in previous iterations of experiments using this infrastructure, the voice agents were all tuned with the same hyper-parameters. This means their boundary tracking, learning, and trust-inference capabilities were consistent for all interactions. It is hence 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.

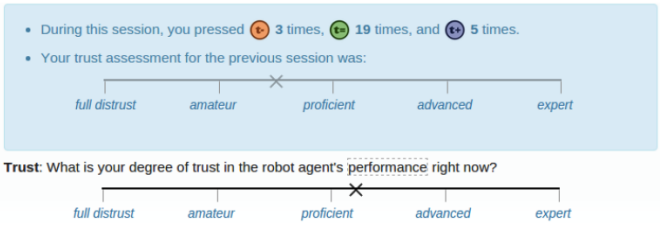


Fig. 2. Post-Session Trust Feedback Scale [3]

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 highlighted 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 [3]." 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 use 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.

The two other metrics measure a supervisor's degree of trust: "a quantifiable subjective assessment towards another individual [3]." 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 are 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 use 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 [ref. in appendix] 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 don't 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 4 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.

III. 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 McGill 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 from a scale from 0 to 10, an average of 5.5 for proficiency at driving cars, 6.3 for comfort

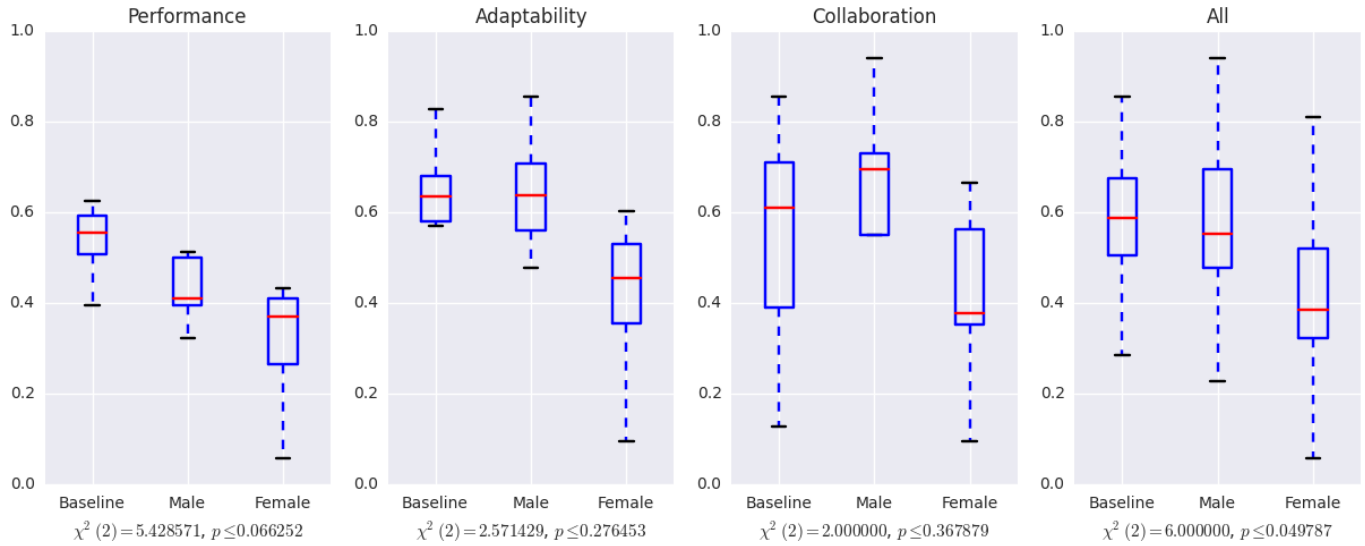
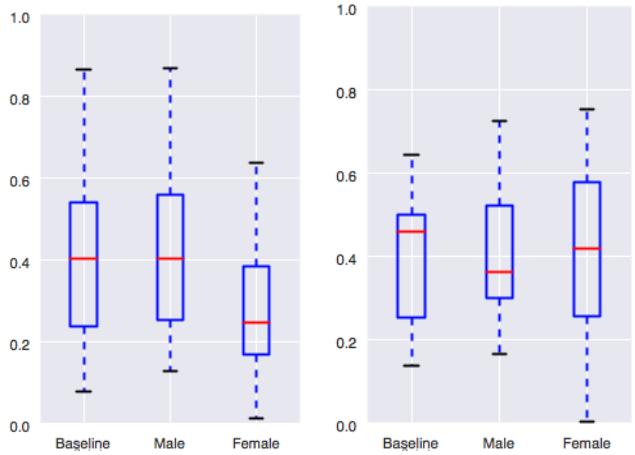


Fig. 3. Agent-Questionnaire Filtered for Male Subjects

level in a self-driving car, and 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 by comparing plots and statistically with a Friedman χ^2 test 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).



(a) Male Participants (b) Female Participants

Fig. 4. Trust Feedback Filtered by Participant Gender

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. 3). Meanwhile, they were consistent with female participant assessments for performance and collaboration, but not adaptability (Fig. 5). 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).

Trust feedback results in Fig. 4 show - albeit insignificantly - male participant's tendency to distrust female voice agents in

[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

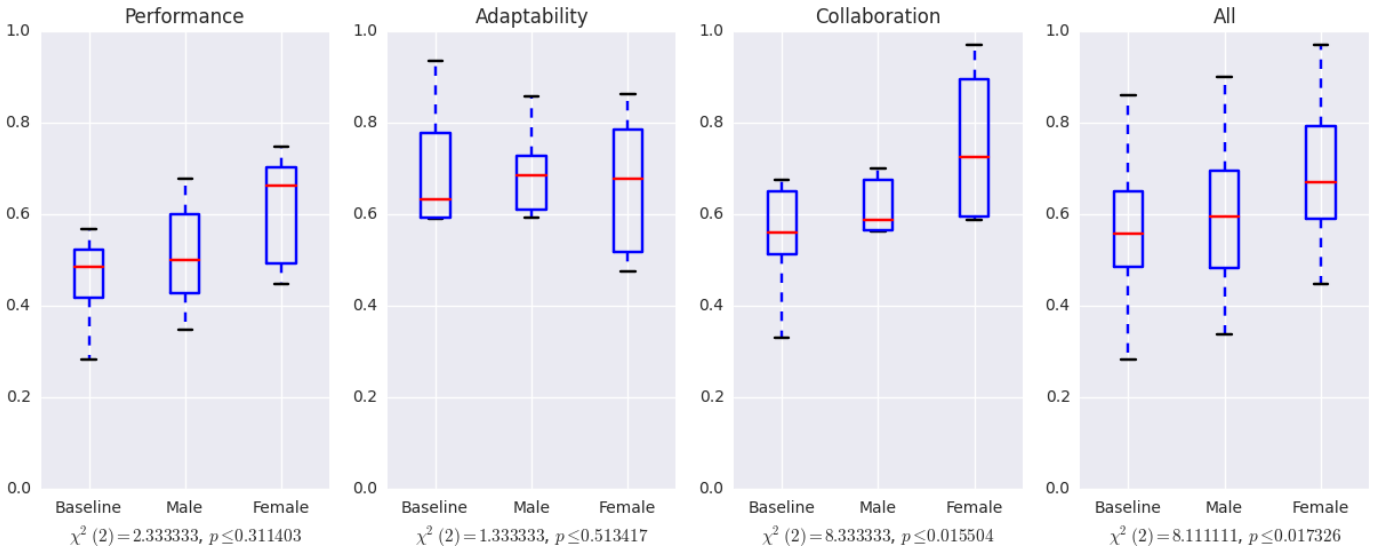


Fig. 5. Agent-Questionnaire Filtered for Female Subjects

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.

IV. Discussion

The results suggest that the presence of a voiced audio cue from a worker does affect a supervisor's trust, but this effect 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 correlate in previous work by Xu and Dudek [ref. maintain efficient collaboration] where their trust model was able to predict degree of trust based on acts of trust along. 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.

Analysis by map provided no significant insight - tendency to trust one agent over another varied from map to map. The highway is a special case where all trust metrics show the average trust for the baseline agent to be lowest among the three agents. We discuss three possible unintended causes for this. First, we see Fig. 6 shows a lower sum of trust critiques for the baseline agent. By the Friedman χ^2 test, this is extremely statistically significant (χ^2=54.504, p$\chi^2$0.000), but there is also a huge error margin. It is like that this study's small sample is too great a limitation for conducting reliable statistical analysis. An alternative proposition is that the highway is a participant's introduction to a given agent, and therefore is more likely to focus their trust assessment to the new feature of the agent (the voice cue), rather than it's task performance. A third theory is because the highway is a simpler map than the other two, both the predictability of both the drone's behaviour and a participant's steering intervention eased the "wizard's" role of triggering voice cues at times that made sense to the participant in context of the scenario.

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

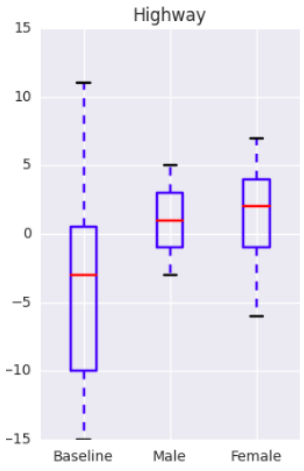


Fig. 6. Sum of Trust Critiques

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. We cannot determine the validity of this critique due to inability to discern definitive results from data filtered for only the first agent in each trial. However this does seem like a very intuitive observation.

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, we did indirectly presume that trust would increase, given that our objective in this study was to investigate factors for more effective collaboration. 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.

V. Conclusion

A. Limitations

The most complicated limitation of this study is the broad subjectivity of the data measurement due to individual biases of participants and the noise that creates in test environments. 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 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, just 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 fairly small sample size. This limited the 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 human-robot interaction, 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.

VI. Acknowledgements

The Principle Investigator would like to thank his supervisor, Prof. Dudek, for inspiring him to pursue his interests in robotics, and for taking him on as an undergraduate researcher. Also, much appreciation is deserved by members of the McGill University Centre for Intelligent Machine's Mobile Robotics Lab, for guidance and friendliness throughout. We would also like to thank all those who participated in the study. Finally, a special thanks to friends who have been supportive from start to finish, and even for helping with voice acting for the voice agents.

References

[1] 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.

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

[3] 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.