How Do Robot Experts Measure the Success of Social Robot Navigation?

Nathan Tsoi
Yale University
New Haven, CT, USA
nathan.tsoi@yale.edu

Jessica Romero
Yale University
New Haven, CT, USA
jessica.romero@yale.edu

Marynel Vázquez
Yale University
New Haven, CT, USA
marynel.vaquez@yale.edu

ABSTRACT

We interviewed 8 individuals from industry and academia to better understand how they valued different aspects of social robot navigation. Interviewees were asked to rank the importance of 10 measures commonly used to evaluate social navigation policies. Interviewees were then asked open-ended questions about social navigation, and how they think about evaluating the challenges they face. Our interviews with industry and academic experts in social navigation revealed that avoiding collisions was the only universally important measure. Beyond the safety consideration of avoiding collisions, roboticists have varying priorities regarding social navigation. Given the high priority interviewees placed on safety, we recommend that social navigation approaches should first aim to ensure safety. Once safety is ensured, we recommend that each social navigation algorithm be evaluated using the measures most relevant to the intended application domain.

CCS CONCEPTS

• **Human-centered computing → Social navigation**.

KEYWORDS

Social Robot Navigation, Performance Measures, Interview

ACM Reference Format:
Nathan Tsoi, Jessica Romero, and Marynel Vázquez. 2024. How Do Robot Experts Measure the Success of Social Robot Navigation?. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3610978.3640636

1 INTRODUCTION

Research in social navigation studies how mobile robots can navigate in concert with people while adhering to social norms. Mobile robots need to operate in a wide range of social situations, which is defined by Tsoi et al. [15] as the physical environment, pedestrian behavior near the robot, and the robot's task. Prior works have studied social navigation in social situations that encompass airports [16], labs [14], and museums [7]. The number of pedestrians near the robot can range from a single person or a few people [11] to crowds of people [1]. The task is often A-to-B navigation, from one position to a goal position, but can also include delivery [8, 10], guiding [2, 6], following [5], serving as a receptionist in a building [4] and interacting with groups [13, 17]. Such a wide variety of social situations makes it challenging to compare different social navigation approaches.

Inspired by the wide range of social situations and corresponding approaches to social navigation, we asked if users of different approaches have different requirements and priorities. There are many different measures used to evaluate social navigation approaches [3, 9]. We hypothesized that users of social navigation robots in different application domains are concerned with different aspects of performance when evaluated by how they prioritize different evaluation metrics. For example, a robot delivering blood for a patient procedure in a hospital may be most concerned with taking the minimum time to deliver the blood. In contrast, a large and dangerous industrial robot in a warehouse may be more concerned with staying a safe distance from everyone in the warehouse.

To better understand how users value and prioritize the behavior of social navigation robots, we interviewed 8 roboticists working in the field of social navigation. The 8 individuals we interviewed were contacts at 8 robotics companies and research labs. They were experts in social navigation working in areas including autonomous delivery, hardware development, space robotics, data analytics, warehouse automation, and academic research.

2 RELATED WORK

Many different evaluation measures have been proposed to evaluate social navigation approaches. Measures can be quantitative or qualitative, the latter typically focused on human perception of robot behavior. We refer the reader to surveys that discuss these measures in detail [3, 9]. In the broader field of Human-Robot Interaction (HRI), common metrics have been reviewed by Steinfeld et al. [12]. In this work, we refer to both metrics and measures as "measures" due to the fact that many "metrics" used in social navigation and HRI do not adhere to the properties of a proper mathematical metric space. We chose to ask interviewees to rank some of the most common [3, 12] and readily available measures [15] covering navigation performance and social perception. We also asked open-ended questions to determine what other measures the interviewees prioritized.

Fairly evaluating different approaches to social navigation requires consideration of many factors, which are outlined by Francis et al. [3], including experimental design, evaluation measures, the social situations used for evaluation, benchmarking against other methods, datasets used, and simulators. Our interviews focused on the evaluation measures, but during the open-ended question

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640636

Category	Question
Demographic	What is your name and which organization do you represent?
Demographic	What is your role at this organization?
Market	What market does the company serve?
Success	Please rank these 10 metrics from most to least important. If there are additional metrics, you will be able to share them after this ranking.
Success	Are there other metrics used to measure success not in the list ranked?
Success	How would you rank their importance?
Success	How would you rank them relative to the metrics we provided?
Success	Do you consider the robot's navigation system as the main metric for success or are there other metrics outside of navigation that determine success?
Success	In what ways has your robot's navigation been changed when being around people to meet the demands of the application domain or market?
Success	Are subjective human opinions a success metric? If so, to what extent?
Success	Is there value in this [subjective] metric?
Success	What would you consider necessary changes still needed to improve the success of your robot?
Success	Are there changes still needed to be made to robots in your domain generally to improve their success in navigating around people?

Table 1: List of questions by category asked to participants during the video interviews. See the text for details.

portion of the interviews, some individuals mentioned other components they considered important, including their datasets, simulators, and how they designed experiments and incorporated end-user feedback.

3 METHOD

Social navigation robots work in a wide range of application domains and users in these different application domains may be concerned with different aspects of a robot's performance. We interviewed 8 individuals from industry and academia to better understand the priorities of users in different application domains. Our protocol was approved by our local Institutional Review Board and refined through pilots.

3.1 Hypothesis

Our hypothesis is that users in different application domains of social navigation robots are concerned with different aspects of performance when evaluated by how they prioritized different evaluation measures.

3.2 Recruitment

We recruited participants using personal communication methods including email and LinkedIn. We initially identified 25 organizations and established a point of contact at each. From the initial pool, 4 organizations were removed because their robots did not perform social navigation. From the remaining 19 organizations, 8 agreed to take part in the study and complete the interview. The representative of 1 organization did not complete the open-ended questions portion of the interview, but did rank the measures we provided. We include their ranking of the 10 measures we provided in our results. One respondent reverse-coded the rankings, which we corrected and included in our results.

3.3 Interviewees

We interviewed contacts at 8 organizations that addressed markets including space robotics, food delivery, general-purpose delivery robots, operations logistics, service robots, education, and computer vision for mobile robots. The individuals who participated in our interviews were from a range of roles within the organizations and held titles such as software lead, head of staff, head of AI and robotics, senior applied scientist, senior scientist, assistant professor, and Chief Executive Officer. Of these individuals, one was working at an academic institution and the rest worked at companies, startups, and industry research labs. Some individuals we interviewed who were working in industry previously worked as academic researchers and professors. The individuals we interviewed included people from two different countries, Spain and the United States of America. Within the USA, people were spread out across 7 different states.

3.4 Procedure

We collected data by conducting semi-structured interviews over 30-minute video calls using the Zoom teleconference platform. All of the information that we collected was anonymized to disassociate responses from any individual or company. Interviews for the study were conducted by the same research assistant and followed a predetermined script which had 5 main phases.

Interview Start (1): The interview began with the interviewer introducing herself and the following statement regarding our goal for the study: "We are conducting a study on robot navigation with the goal of collecting information about how different groups and companies are measuring success for mobile robots capable of navigating with or around people. We are specifically interested in learning more about how success is determined for different robots."

Voluntary Participation (2): Each participant was told that participation in the study is voluntary and they are free to decline to participate or end their participation at any time.

Recording Consent (3): Each participant was asked for consent to record the video call and transcribe the audio to text for the sole purpose of coding the interview question.

Interview Questions (4): Following verbal confirmation of their agreement to participate in the study, each participant was then asked 14 questions which included demographic information, the business market their organization serves, and questions about how they measure success in social navigation. This included a question that asked the participant to rank 10 measures commonly used in social navigation. The 10 measures were: completed navigation goals, path length, minimum distance to target, final distance to target, time not moving, path irregularities, path efficiency, distance violations, intimate distance violations, and collisions. We also

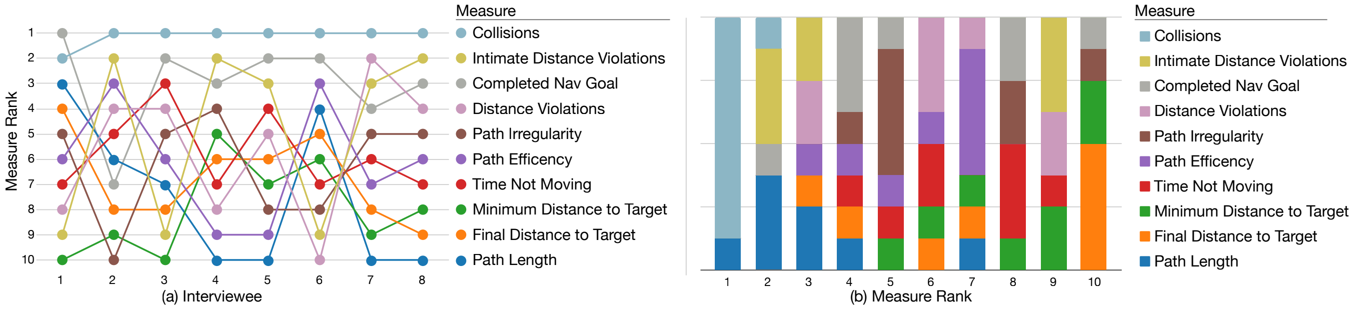

Figure 1: Two plots that show the measure ranking results visualized in different ways. (a) Ranking of social navigation evaluation measures by interviewee. Where 1 corresponds to the most important and 10 corresponds to the least important evaluation measure. (b) Interviewees that assigned the same rank to a metric where the bar length indicates the number of interviewees who assigned a given rank (x-axis) to each measure (color). Best viewed digitally.

asked open-ended questions about the success of social navigation. The exact wording of these questions is detailed in Table 1.

Interview End (5): The interview ended with an open-ended question regarding the participant's other thoughts surrounding the topics discussed during the call.

4 RESULTS

We hypothesized that users of social navigation approaches in different application domains are concerned with different aspects of performance when evaluated by how they prioritized different evaluation measures. We asked participants to rank 10 measures commonly used to evaluate social navigation approaches, shown in Figure 1, from most (1) to least important (10). While we did see variation in most rankings, the collisions measure was surprisingly ranked most important by all but one participant.

We performed a qualitative analysis of the open-ended interview questions by aggregating them and identifying themes in the responses. This process revealed the same phenomena. Across all interviewees, the primary concern was safety, but after this consideration, priorities varied widely. Interviewees' primary concerns, after safety, included their robot's ability to localize, user privacy, communication (via lights, speech, and motion), task throughput, engineering time required to recover from an error, the interpretability of motion, and enjoyability of interacting with the robot.

The variation in interviewee considerations indicates that a wide range of evaluation measures are appropriate for handling the wide range of social situations that robots encounter. Quantitative measures are necessary to evaluate social navigation approaches from the perspective of task performance. Qualitative measures can be used to measure how end-users perceive the performance of the robot, which is important for evaluating social considerations such as interpretability and enjoyability of interaction with the robot.

We observed the hypothesized differences in priorities across application domains, which were reflected in different evaluation measures. We also observed a difference in priorities given different roles within an organization. Individuals involved in the engineering and design processes were firstly concerned with the lower-level behavior of their robot. Individuals in leadership roles were more concerned with task-level and organizational-level goals. We saw this difference primarily in the open-ended questions where engineers and designers were concerned with the lower-level measures commonly used in social navigation, while institutional leaders were interested in measures that related to organizational-level, financial success, such as task throughput and minimizing engineering time.

5 LIMITATIONS

Our study had several limitations. First, while all interviews were conducted via Zoom, one interview ran over time and responses to some questions were emailed to the interviewer following the Zoom call. Another limitation is that we did not provide detailed descriptions of the evaluation measures. For example, we did not define the difference in distance between intimate distance violations and simple distance violations, but instead stated that intimate distance violations were when the robot came closer than a distance violation. We chose to omit details such as precise distances because we wanted to avoid biasing participants' responses given interviewees' different use cases. Finally, although we interviewed 8 individuals, from a wide range of organizations, further interviews could be conducted in the future.

6 CONCLUSION AND RECOMMENDATIONS

Our hypothesis was that users in different application domains of social navigation robots are concerned with different aspects of performance when evaluated by how they prioritized different evaluation measures. To evaluate this hypothesis, we interviewed 8 individuals from both academia and industry who are experts in social navigation. Data collected during these interviews showed that our hypothesis was partially supported. While minimizing collisions was almost universally the top priority, all other measures varied in priority across application domain. This was also supported by responses to open-ended questions which showed a variation in priorities across application domains. Moreover, interviews revealed that there was also a difference in priorities between people at different levels of an organization.

Given the difference in priorities regarding robot behavior across application domains and roles within an organization, we make three recommendations for the development and evaluation of social navigation algorithms. First, while most evaluation measures are prioritized differently, avoiding collisions is a near-universal goal. Therefore, all approaches to social navigation should first aim to ensure safety by utilizing an evaluation measure such as minimizing the risk of collision. Second, users in a given application domain should evaluate their robots using measures that matter most to their domain. If users in different domains were to share the prioritization of evaluation measures, this could serve as a starting point for collaboration between users that have common goals. Finally, given the potential for different priorities across roles within the same organization, we recommend that roboticists utilize low-level performance measures and roll up these low-level measures, such as time to completion for a trajectory, into measures that tie into organization-level goals, such as how low completion time might increase the task-efficiency of the robot which may equate to profit or research impact.

ACKNOWLEDGMENTS

The authors thank the NSF for partially supporting this work under Grant No. IIS-1924802. Any opinions and findings presented in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–971.

[2] Abir Bellarbi, Souhila Kahlouche, Nouara Achour, and Noureddine Ouadah. 2016. A social planning and navigation for tour-guide robot in human environment. In *2016 8th international conference on modelling, identification and control (ICMIC)*. IEEE, 622–627.

[3] Anthony Francis, Claudia Pérez-d'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, et al. 2023. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740* (2023).

[4] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C Schultz, et al. 2005. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1338–1343.

[5] Rachel Gockley, Jodi Forlizzi, and Reid Simmons. 2007. Natural person-following behavior for social robots. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 17–24.

[6] Päivi Heikkilä, Hanna Lammi, Marketta Niemelä, Kathleen Belhassein, Guillaume Sarthou, Antti Tammela, Aurélie Clodic, and Rachid Alami. 2019. Should a robot guide like a human? A qualitative four-phase study of a shopping mall robot. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*. Springer, 548–557.

[7] Mehdi Hellou, JongYoon Lim, Norina Gasteiger, Minsu Jang, and Ho Seok Ahn. 2022. Technical Methods for Social Robots in Museum Settings: An Overview of the Literature. *International Journal of Social Robotics* 14, 8 (2022), 1767–1786.

[8] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. 2012. Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 695–704.

[9] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. 2023. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction* 12, 3 (2023), 1–39.

[10] Bilge Mutlu and Jodi Forlizzi. 2008. Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 287–294.

[11] Claudia Pérez-D'Arpino, Can Liu, Patrick Goebel, Roberto Martín-Martín, and Silvio Savarese. 2021. Robot navigation in constrained pedestrian environments using reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1140–1146.

[12] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 33–40.

[13] Xuan-Tung Truong and Trung-Dung Ngo. 2017. To approach humans?: A unified framework for approaching pose prediction and socially aware robot navigation. *TCDS* (2017).

[14] Nathan Tsoi, Mohamed Hussein, Olivia Fugikawa, JD Zhao, and Marynel Vázquez. 2021. An approach to deploy interactive robotic simulators on the web for hri experiments: Results in social robot navigation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7528–7535.

[15] Nathan Tsoi, Alec Xiang, Peter Yu, Samuel S Sohn, Greg Schwartz, Subashri Ramesh, Mohamed Hussein, Anjali W Gupta, Mubbasir Kapadia, and Marynel Vázquez. 2022. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters* 7, 4 (2022), 11047–11054.

[16] Dizan Vasquez, Billy Okal, and Kai O Arras. 2014. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1341–1346.

[17] Fangkai Yang and Christopher Peters. 2019. AppGAN: Generative adversarial networks for generating robot approach behaviors into small groups of people. In *RO-MAN*.