Abstract

Creating Socially Competent Mobile Robots
Rethinking How to Measure Success

Nathan Tsoi

2025

Motivated by the value alignment problem, which is concerned with ensuring that an autonomous system's behavior matches user values, this dissertation proposes that *creating socially competent mobile robots requires rethinking how success is measured* in order to *align evaluation metrics with human values* and to this end, proposes the use of *context-aware simulation systems and subjective human feedback.*

Social robot navigation is concerned with an agent that must traverse the navigable space in an environment that is shared with people. Conducting such motion in dynamic and densely populated environments demands that a robot understands how humans perceive its behavior and then respond appropriately. This is a challenging task as small deviations in social behavior can significantly impact how people perceive and respond to the robot. Yet, research in social navigation often relies on objective metrics that fail to capture these subtle social factors, leading to policies that may optimize for obvious and easily measurable metrics like efficiency, but neglect the social aspect of how people perceive robot behaviors in terms such as competence.

The foundation for this dissertation's contributions is in the area of simulation. We introduce SEAN: the Social Environment for Autonomous Navigation and its follow-on project, SEAN 2.0, a high-visual-fidelity, extensible, and human-centric simulation tool. SEAN allows researchers to develop, test, and compare social navigation algorithms in safe, controlled environments. The contributions that my work makes in the area of simulation are useful for researchers throughout the development lifecycle of social navigation systems.

Building on work in simulation, this dissertation makes contributions to evaluating social navigation systems. Fair comparison of existing and future systems allows measurement of and future progress in the field. Critical to fair comparison is a characterization of different social contexts during navigation because social actions are context dependent. Inspired by social psychology, we propose a preliminary set of "Social Situations" that characterize some contexts during social navigation. We then conducted structured interviews with experts working to understand if there is an overarching objective metric which can be used for fair evaluation. We found that beyond safety, the ranking of different metrics varied by application domain. As part of the interviews, we also asked open-ended questions. Responses to these questions highlighted the need to incorporate subjective evaluation criteria, because objective measures alone are insufficient to capture the nuances of the social aspects of navigation.

Finally, with the understanding of how critical human perceptions are to the development of social navigation policies, we study the impact of methodological choices researchers can make when collecting human feedback. To enable this work, I led development of the SEAN Experiment Platform (SEAN-EP), which allows researchers to collect human-feedback using interactive, online surveys. Using SEAN-EP, we compare the gold-standard of an interactive, in-person study with scalable online, interactive surveys, and a typical video-based survey. We find that interactive methodologies are preferable to passive alternatives. Still, even with scalable, interactive data collection via SEAN-EP, querying humans for their perceptions of robot behavior requires a significant amount of time and effort. Therefore, we investigate whether it is possible to predict perceptions of robot performance using machine learning in data-limited regimes.

Collectively, the contributions of this dissertation provide a foundation for building and evaluating social navigation robots. By integrating context-aware simulation,

human-centered evaluation methodologies, and predictive models of subjective human feedback, this work enables more systematic alignment of robot behaviors with people's social expectations. These contributions open the avenue for future research identified in this dissertation, including the development of universally accepted summary metrics for social navigation success, the creation of simulation systems that capture a richer range of human behaviors, and the incorporation of human feedback into robot policies that learn and adapt to predicted human perceptions.

Creating Socially Competent Mobile Robots

Rethinking How to Measure Success

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Nathan Tsoi

Dissertation Director: Marynel Vázquez

December 2025

Contents

List of Figures

List of Tables

Acknowledgments

My Ph.D. work was possible only with the support of many family, friends, collaborators, and mentors. First, thank you to my Ph.D. advisor and academic mentor, Marynel Vázquez. Your research mentorship has been a formative and invaluable element in my academic career. I would also like to thank my committee members: Brian Scassellati (Scaz), Daniel Rakita, and Nicholas Roy. I especially want to thank Scaz for your advice and mentorship throughout my Ph.D. I thank Hamid Rezatofighi, Ian Reid, and Silvio Savarese for their guidance as an early-career researcher, as well as Bill Coughran and Chad Dyer for their guidance navigating the transition from industry to academia.

I am grateful for all the friends and colleagues at the Interactive Machines Group and the Social Robotics Lab. I have thoroughly enjoyed all our conversations over the years. Austin Narcomey, Chayan Sarkar, Debasmita Ghose, Ellie Mamantov, Emmanual Adéníran, Etiosa Omeike, Fern Limprayoon, Houston Claure, Jake Brawer, Kate Candon, Kayla Matheus, Meiying Qin, Nick Georgiou, Nicole Salomons, Qiping Zhang, Rebecca Ramnauth, Sarah Sebo, Sasha Lew, Sydney Thompson, and Timothy Adamson have made my time at Yale very special. It has been a pleasure to work alongside a fantastic group of undergraduate students including Alex Xiang, Amanda Hansen, Anjali W. Gupta, Booyeon Choi, Daniel Lee, Deyuan Li, Greg Schwartz, J.D. Zhao, Jeacy Espinoza, Jessica Romero, Joe Connolly, Joseph Valdez, Kaitlynn Taylor Pineda, Mofeed Nagib, Mohamed Hussein, Olivia Fugikawa, Peter Yu, Rachel

Sterneck, Subashri Ramesh, Xavier Ruiz, and Yofti Milkessa.

Thank you to all of the mentors and collaborators I've interacted with over the years in both industry and academia including Abhijat Biswas, Ada V. Taylor, Adam Fineberg, Alexander Toshev, Alexandre Alahi, Aniket Bera, Anthony Francis, Chengshu Li, Claudia Pérez-D'Arpino, Dylan Glas, Fei Xia, Hao-Tien Lewis Chiang, Haresh Karnan, Jie Tan, Jonathan P. How, Joydeep Biswas, Justin Hart, Luis J. Manso, Michael Everett, Mubbasir Kapadia, Naoki Yokoyama, Peng Xu, Peter Trautman, Phani Teja Singamaneni, Rachid Alami, Reuth Mirksy, Roberto Martín-Martín, Rohan Chandra, Ruta Desai, Sehoon Ha, Sören Pirk, Tsang-Wei Edward Lee, Xuan Zhao, and Xuesu Xiao.

I would also like to thank my friends from Trinity Baptist Church, who have been a constant support throughout my dissertation, along with my parents, Chris and Theresa, and my siblings, Lexi and Audrey. Thank you to my children, Lois, Isaac, Eliora, and Abigail, for your interest in my work and your endless encouragement. Finally, to my wife, Julie, thank you for your love and support.

Chapter 1

Introduction

Social robot navigation is an application area at the intersection of robotics, machine learning, and social robotics. Unlike traditional robot navigation, which focuses primarily on efficiency and collision avoidance in static or controlled environments, social robot navigation requires robots to move through dynamic human-populated spaces while adhering to implicit social norms and expectations. The complexity of this task arises from several factors: the unpredictability of human movement, the need to interpret and respond to social cues, the cultural and contextual variations in spatial norms, and the subjective nature of what constitutes socially acceptable robot behavior. While significant progress has been made in developing algorithms that perform well in controlled environments, adapting these algorithms to dynamic social settings requires understanding how humans perceive robot behavior and incorporating this understanding into both the learning and evaluation processes.

The challenge of creating socially competent robots is fundamentally driven by the value alignment problem. As described by Russell and Norvig [225], value alignment refers to ensuring that an autonomous system's behavior aligns with human values. This is particularly critical in social robotics, where misalignment between robot behavior and human expectations can lead to rejection of the technology, regardless of

the technical sophistication. The value alignment problem is especially challenging because human values are complex, context-dependent, and often difficult to formalize mathematically. In social navigation, what constitutes socially acceptable robot behavior varies based on environmental context, cultural norms, and individual preferences. A robot programmed solely to minimize path length or avoid collisions may technically accomplish its navigation goal but fail to navigate in a socially acceptable manner, causing discomfort or confusion among humans sharing the space.

The value alignment challenge is closely related to "the tyranny of metrics" which is a phenomenon where optimizing for the wrong metrics leads to behaviors that are misaligned with human values. The tyranny of metrics occurs when an evaluation focuses on easily quantifiable objectives while neglecting harder-to-measure qualities that ultimately matter more. In organizational contexts, this has been well-documented where optimization for specific metrics can produce behaviors contrary to an organization's broader mission. For example, the pharmaceutical company Mylan faced significant backlash over its pricing of the EpiPen, a life-saving device for severe allergic reactions. The company's focus on maximizing short-term profit metrics led to price increases of over 500% between 2007 and 2016, raising the cost from approximately $100 to over $600. This metric-driven decision, while temporarily boosting financial indicators, ultimately resulted in congressional investigations, public outrage, and lasting damage to the company's reputation and stock value. The tyranny of metrics in this case led Mylan to optimize for a narrow financial measure at the expense of the broader value of providing affordable access to life-saving medication [184].

Motivated by the value alignment problem and in an effort to avoid the tyranny of metrics, this dissertation contributes systems for training and evaluating social robot navigation in a way that is aligned with human values. Traditional navigation systems are often evaluated using metrics like path efficiency, time to goal, and collision

avoidance. While these metrics capture important aspects of navigation, they fail to account for critical social dimensions such as comfort, predictability, and perceived safety. The challenge is not just to optimize for the right metrics but to understand what the right metrics are in different social contexts.

This dissertation proposes that *creating socially competent mobile robots requires rethinking how success is measured* in order to *align evaluation metrics with human values* and, to this end, proposes the use of *context-aware simulation systems and subjective human feedback*.

This dissertation begins with an overview of Social Robot Navigation in Chapter 2. It takes inspiration from work that I co-authored with Anthony Francis and others following a symposium on Social Navigation in 2022 titled "Principles and Guidelines for Evaluating Social Robot Navigation Algorithms" [87], accepted to the ACM Transactions on Human-Robot Interaction (THRI) journal. The chapter introduces key definitions and highlights the relevance of social navigation in robotics. The chapter then examines the unique challenges posed by the social aspects of navigation. It also reviews existing simulation platforms, emphasizing the design decisions that support realistic and meaningful evaluation. Finally, it presents criteria for measuring success, incorporating both objective metrics and subjective assessments, along with methodologies for collecting human feedback.

Building on the high-level overview, Chapter 3 introduces the motivation for a human-centric simulation platform. It details the design considerations that guided the initial development of the simulation platform that I proposed, the Social Environment for Autonomous Navigation (SEAN), which places an emphasis on the potential for high-visual-fidelity, integration with the Robot Operating System (ROS) [212], and dynamic environments populated with pedestrians. This chapter is based on the work "SEAN: Social Environment for Autonomous Navigation," which won a best poster award - runner up prize at the 2020 Conference on Human-Agent Interaction

(HAI) [268].

Chapter 4 presents a detailed analysis of the latest version of the SEAN simulation system, a development effort that I led as part of this dissertation. This version, called SEAN 2.0, includes new system features that support the simulation of pedestrian motion through a novel crowd flow model. As part of this work, we provided a characterization of social context inspired by ideas from Social Psychology [13]. We then operationalized the characterization as a classifier based on this model which allows the identification of social context, referred to as "Social Situations." This work was published in the paper "SEAN 2.0: Formalizing and generating social situations for robot navigation," published in the IEEE Robotics and Automation Letters (RA-L) [273].

The ability to characterize and identify Social Situations is critical for the fair evaluation of different robot navigation policies in different environments and under different conditions. Once we can identify when robots are in similar Social Situations, a specific metric can then be applied to compare robot performance across different evaluation runs. The next logical question is what metrics should be used to evaluate policies under different social situations and thereby help one avoid the tyranny of metrics. Therefore, Chapter 5 presents a study in which structured interviews were conducted with robot experts. The experts were asked to rank common metrics for social robot navigation and answer open-ended questions about the most important aspects of social robot navigation. An insight from these interviews is that subjective human feedback is a critical component of evaluating social robot navigation. This chapter is based on the work "How Do Robot Experts Measure the Success of Social Robot Navigation?" published in the Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction [275].

The knowledge that subjective human feedback is a critical component of evaluating social robot navigation motivates Chapter 6. We present the SEAN Experimental

Platform (SEAN-EP), a tool for collecting human feedback for social robot navigation via interactive simulations that are incorporated into online surveys. This is based on the work "SEAN Experimental Platform: A Tool for Collecting Human Feedback for Social Robot Navigation," published in the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) [270]. SEAN-EP advances the goals of this dissertation by providing a scalable means of collecting subjective human feedback on the social aspects of robot navigation.

In Chapter 7, we use the SEAN-EP system to study the collection of human feedback via interactive methodologies. The two dimensions of Simulation vs. Real-World and Interactive vs. Video are compared. The gold-standard condition is the real-world, interactive condition, typical of human-subjects studies. The other typical condition, non-interactive video-based surveys is compared against online, interactive surveys via SEAN-EP. This study was presented in the work "Influence of Simulation and Interactivity on Human Perceptions of a Robot During Navigation Tasks," published in the ACM Transactions on Human-Robot Interaction (THRI) [276]. We find that while there are tradeoffs between these methodologies, interactive simulations are a useful tool for collecting human feedback.

Finally, Chapter 8 explores using machine learning to predict human perceptions of robot navigation performance from implicit feedback, reducing the need for manual supervision. We introduce the SEAN TOGETHER Dataset of VR-based human-robot interactions and show that spatial behavior features are key to inferring perceptions of robots. Our models, validated in both VR and real-world settings, perform well on a binary prediction task and generalize across users. This work suggests a path toward scalable, perception-driven robot learning. Our contributions were presented in the article "Predicting Human Perceptions of Robot Performance During Navigation Tasks" in the ACM Transactions on Human-Robot Interaction (THRI) [309].

Chapter 2

Mobile Robots that Navigate in Human-Centric Environments*

Social Robot Navigation (SRN), a subfield of Human-Robot Interaction (HRI), has a well-established and growing body of research. Several recent surveys provide overviews of the field, including Kruse et al. [144], Rios-Martinez et al. [218], Chik et al. [64], Charalampous et al. [53], Cheng et al. [60], Gao and Huang [89], Mavrogiannis et al. [174], Mirsky et al. [179], Möller et al. [183], Wang et al. [290]. Among these, Gao and Huang [89] present an extensive review including 177 papers focusing on social robot navigation algorithms This survey explores evaluation methods, scenarios, datasets, and metrics, drawing on these findings to highlight current limitations and suggest future research directions. Another recent survey by Mavrogiannis et al. [174] focuses on the core challenges of social navigation, emphasizing navigation algorithms, human behavior models, and evaluation. Wang et al. [290] contributes new evaluation metrics aligned with the principles proposed by Kruse et al. [144] including comfort, naturalness, and sociability. For readers seeking a broad overview

*Parts of this chapter were originally published as Anthony Francis, ..., Nathan Tsoi, ..., and others. (2023). Principles and guidelines for evaluating social robot navigation algorithms. In *ACM Transactions on Human-Robot Interaction (THRI)* [87].

of the field, I refer them to these surveys. However, given the diversity of existing success metrics, this dissertation proposes that creating socially competent mobile robots requires rethinking how success is measured. In particular, I propose that evaluation metrics must be aligned more closely with human values. To this end, this dissertation advocates for the use of context-aware simulation and the incorporation of subjective human feedback. In the future, my research aims to extend these insights to other domains of social robotics that can benefit from lessons learned in social robot navigation.

2.1 Defining Social Navigation

Research in Social Robot Navigation has the potential to transform how robots integrate into human environments. While simple, point-to-point navigation in controlled environments with well-defined parameters is generally considered a solved problem, enabling robots to navigate complex, dynamic, and human-centric environments in ways that are compelling to people is a significant challenge.

Humans are inherently social creatures. As we move through the physical world, we engage in social behaviors such as making eye contact, moving aside to yield space to others, and using subtle gestures to convey intent or acknowledge others. These behaviors reflect our social expectations. We naturally expect other agents, including robots, to exhibit some level of social behavior as well.

Understanding the nuances of human social behavior and implementing them in mobile robots is a challenging task. It requires the technical ability to detect and respond to subtle social cues, for example, by leveraging the ability to identify social interactions and respond in a way that reflects human expectations.

In their review of Human-Robot Interaction for social robots, Kanda and Ishiguro [127] argue that in addition to the basic functions of navigation (moving robots from

place to place) and manipulation (interacting with objects) robots must be capable of social interactions, which involve engaging with humans or other robots to complete tasks. They distinguish robots that encounter humans from those that interact meaningfully using features such as speech, via expressive faces, or gestures.

However, a robot that simply has socially interactive features does not mean that its interactions will be perceived as suitable to the people it interacts with. High-quality social interactions require subtle behaviors influenced by appropriate timing, adaptation, and the perception which leverages understanding during the two-way flow of communication between robot and human. Many systems studied by Kanda and Ishiguro required a human to teleoperate the robot in order to succeed, which highlights the challenges of developing fully autonomous social robots.

This raises the question of what distinguishes "social" robots from those that are merely interactive. In order to address this question, we considered the use of the term *social* and *antisocial* as it has applied to humans. One meaning of *social* is participating in a group, modifying one's behavior to meet the expectations and needs of others in the group while still achieving one's own goals. Another definition emphasizes empathy and skill in interpersonal engagement. That is, the understnading of others's feelings and adapting one's behavior to them. On the contrary, *antisocial* behaviors are those that fail to follow the customs of society or live without consideration for others' needs.

I extend these ideas to social robot navigation. A broad definition of social robot navigation encompasses three key components: mobility, social interaction, and the potential for operation in concert with humans. Firstly, the robot must be mobile, capable of moving in physical space, whether on the ground plane, in the air, or in water. Secondly, it must be capable of social interaction, engaging with humans on some level. Lastly, the robot should have the potential to operate in concert with people, navigating around them and communicating either verbally or non-verbally.

Importantly, the robot does not need to always operate in concert with people, but should be capable of doing so when the situation demands it. This multifaceted view emphasizes that social robot navigation is not just about locomotion as an application area, but about aligning robot behavior with people's social expectations.

2.2 Key Challenges in Social Robot Navigation

Designing mobile robots with sufficient social capabilities to complete navigation tasks while acting in a socially acceptable manner presents a number of challenges.

Due to the fact that social expectations change along with the current social situation, evaluation of progress in the field of social robot navigation must consider the social context in which a robot is currently operating. Defining context in a way that is both meaningful to humans and operationally useful for robots is a significant challenge. Even so, researchers should aim to describe the cultural, environmental, operational, task, and interpersonal context for social robots so that comparisons can be made across the findings of the many studies that span the field of social robot navigation.

At the most fundamental level, robots must ensure *safety*, avoiding physical harm not only to humans but also to property and other machines [149, 37]. Yet physical safety alone is not sufficient. Prior works have studied how robots should maintain *comfort*, avoiding behaviors that induce stress, violate personal space, or appear unnatural [144]. These considerations are often subtle; for example, maintaining appropriate proxemic distances [101, 218] and limiting erratic or jittery motion are key factors in human perceptions of comfort and social acceptability. Ensuring that a robot's actions are *legible*, i.e., that their goals and future behaviors are easily interpretable by observers, is another significant challenge [76]. This involves modifying trajectories, or supplementing motion with communicative cues, so that nearby hu-

mans can anticipate the robot's intentions.

Beyond making behavior clear and non-threatening, social navigation also demands a form of *politeness* which involves respecting social norms both in movement and communication [120, 213]. For example, robots should avoid cutting off pedestrians or forcing humans to yield in narrow passageways without appropriate signaling. These behaviors should be balanced with *social competency*, which entails recognizing and conforming to local norms and conventions for shared spaces [63, 59, 179]. For example, social robots could consider lane conventions, turn-taking, and implicit rules about how groups move and interact. Underlying all of this is the challenge of *understanding other agents*: predicting human behavior and adjusting robot behavior accordingly [204, 262, 30]. This may involve recognizing conversations, avoiding interrupting social interactions, or proactively preventing trajectory conflicts [52].

Agents are not always simply reactive, they often also demonstrate *proactivity*, taking initiative to resolve or prevent navigational deadlocks, or to facilitate human progress in shared environments [169, 244, 51]. In highly dynamic spaces, socially competent robots should not only act appropriately, but also adaptively. The notion of *contextual appropriateness* captures this requirement: robot behavior must be adjusted according to elements such as cultural norms, task urgency, physical environment, and interpersonal interactions [29, 189, 214]. For example, maintaining politeness may be deprioritized in time-critical applications such as emergency medical delivery, whereas the same behavior might be essential in a museum guide setting. These challenges are deeply interdependent and often in tension, requiring a principled yet practical approach to evaluating social robot behavior in complex contexts.

2.3 Simulation for Social Robot Navigation

Simulators cannot simulate all aspects of the real world equally well. Therefore, when designing a simulator for social robot navigation, developers must make deliberate choices about which features to prioritize. At the most basic level, a social navigation simulator must support the interaction between at least two agents. Beyond the minimum requirement, simulator design decisions should incorporate several key factors including the level of abstraction, fidelity of agent and scene representations, realism of physical interactions, and the sophistication of pedestrian modeling in terms of motion planning and degrees of freedom. Some simulators focus on computational efficiency by abstracting agents as simple geometric primitives (e.g., discs or cylinders) and operating in lightweight two-dimensional environments [262, 58]. Others strive for greater realism by incorporating articulated pedestrian motion, reactive behaviors, photorealistic three-dimensional scenes, and kinodynamic constraints, albeit often at the expense of computational tractability and ease of implementation [224].

The intended use of the simulator should also play a role in the design process. Considerations stem from uses such as algorithm development and benchmarking, reproducibility and calculation of particular metrics. These objectives typically favor structured, repeatable environments with well-controlled sources of variability. By contrast, when the objective is to study human perception or interaction dynamics, simulators must prioritize both visual and behavioral realism. In such cases, the modeling of pedestrian reactivity, variability in human internal states (e.g., curiosity, fear, or indifference), and rich environmental interactions become more critical [144]. Additionally, interoperability with popular development tools such as the Robot Operating System (ROS) [212] or reinforcement learning frameworks like OpenAI Gym [42], as well as support for common scene formats and multi-agent policy frameworks, can significantly influence the simulator's integration into existing research pipelines and its capacity to support a range of experimental workflows.

Another set of design decisions arise from the intended use of the simulator. For example, algorithm development and benchmarking require reproducibility and the ability to compute standardized metrics, often favoring structured environments and controlled variability. In contrast, simulators used for studying human perception or interaction dynamics must prioritize realism, both visually and behaviorally. Factors like pedestrian reactivity, variability in human behaviors as well as internal states (e.g., curiosity, fear, or indifference), and the richness of environmental interactions vary in criticality depending on the design goals.

The simulation platform proposed in this dissertation, the Social Environment for Autonomous Navigation (SEAN) [268, 273, 308], is focused on three key aspects of the design process. First SEAN is a platform that is useful for running algorithms implemented in the Robotic Operating System (ROS) [212], which is commonly used to implement social navigation algorithms for real robots. SEAN emphasizes compatibility with ROS, enabling seamless execution of algorithms developed for physical robots. Built on the Unity game engine, SEAN supports high-fidelity rendering and enables realistic outputs from simulated sensors. This includes simulated cameras, particularly useful for perception-focused research, as well as simulated LiDAR, which is commonly used for mapping and localization. Pedestrian motion is modeled at two levels: low-level collision avoidance behavior is governed by the Social Forces Model (SFM) [106], while high-level crowd flow is determined by a novel Behavior Graph [273]. In combination, this allows for expressive and controllable pedestrian behavior in complex and dynamic environments and also eliminates the need for the labor-intensive process of manually specifying the trajectory for each pedestrian. The evaluation of social navigation policies in SEAN is supported through deterministic initial conditions and a suite of objective metrics. However, because interactions in social navigation are inherently variable, SEAN additionally provides a formalization of "Social Situations," which is a framework for characterizing and identifying social

interactions across different trials. This unique set of features makes SEAN a powerful tool for advancing the study of social competence in robot navigation, while supporting the rigor, scalability, and reproducibility that modern research demands.

2.4 Measuring Success in Social Robot Navigation

Measuring success in social robot navigation remains a core challenge in the field. Unlike traditional robot navigation tasks, social navigation involves coordinating motion in spaces shared with humans, where success is not just a matter of reaching a goal but of doing so in a way that adheres to social norms, communicates intent clearly, and avoids discomfort or harm to people. This complexity makes it difficult to agree on a unified metric that captures a holistic notion of success in social contexts.

A central reason for this difficulty is the multi-faceted nature of human-robot encounters. Researchers must consider not only physical safety (e.g. avoiding collisions), but also psychological safety, social acceptability, and the interpretability of robot behaviors. Even seemingly straightforward terms like "safety" can take on different meanings depending on the context, ranging from physical proximity to perceived emotional comfort and moral appropriateness [149, 126, 37].

2.4.1 The Evolution of Metrics in Social Robot Navigation

Over time, the field has evolved from task-based success metrics (e.g., success rate, distance without incident [171]) to quality-based metrics (e.g., SPL [10]) and finally to social metrics that reflect the robot's impact on humans. Social metrics can include measures of personal space compliance [283], protocol-based surveys [206], or learned models based on human ratings [299, 68].

Metrics for social navigation, however, come with their own challenges. Real-world human-robot interactions are highly contextual, often dynamic, and affected

by factors like prolonged exposure, user learning, or deployment-specific challenges. Furthermore, different stakeholders may value different aspects of the interaction: a customer might prioritize approachability, while a logistics operator might prioritize efficiency. Experts in social robot navigation rank different metrics as more or less important depending on the application [275].

2.4.2 Subjective vs. Objective Evaluation

While objective metrics are well-defined and straightforward to compute, their connection to the quality of the *social* component of an interaction, can be tenuous. Therefore, to evaluate social interactions, researchers often rely on subjective metrics collected from human participants by querying them using scales such as the Robot Social Attribute Scale (RoSAS) [49] or the Perceived Social Intelligence (PSI) [24] scale, which capture a range of different aspects of a social interaction such as perceived comfort and social intelligence. Collecting human feedback during in-person interaction between a person and a robot is considered the gold standard, but in-person studies are expensive, time consuming, and difficult to scale. Repeatedly querying users mid-interaction can also disrupt the very social dynamics being measured. In response, there is growing interest in utilizing alternative methods of collecting human feedback or even predicting perceptions of robot performance from limited data.

2.4.3 Social Navigation Evaluation Protocols

Standardized evaluation protocols provide a critical component in the development of social navigation algorithms. Utilizing standardized evaluation protocols, recent benchmarks [35, 88, 104, 113, 133, 34] have incorporated both objective and subjective measures to facilitate meaningful comparison across methods. These benchmarks aim to drive progress in the field by assessing comparative performance against baseline and proposed methods, identifying the contribution of specific components through

ablation studies, or evaluating generalization to new environments. Objective aspects such as obstacle avoidance, trajectory smoothness, and task success can often be measured without human input, but subjective social principles like safety, comfort, and politeness require human input, making human-subjects studies essential. As a result, robust evaluation protocols must consider not only what can be measured computationally, but also how robot behaviors are perceived by people. Toolkits that allow for annotation of simulated trajectories and the development of predictive models from human-labeled data are therefore increasingly important as they may enable a research lifecycle where empirical findings inform both method development and evaluation protocols.

2.5 Data Collection for Social Robot Navigation

Data collection for social robot navigation can take different forms, ranging from in-person interactive studies, typically conducted in controlled environments, to online surveys that utilize videos or, recently even online interactive methods [275]. In-person studies provide an interaction-rich experience for participants during which they can observe and interact with a social robot in the real world. Following this interaction, the participant is asked to fill out a survey to rate their experience with the robot. A challenge of in-person data collection is that it can be expensive, time consuming, and logistically complex. Online studies can be a more scalable and reproducible approach, however they typically are not as interactive because the typical approach for online studies is to have the participant watch a video of a human-robot interaction, instead of actively participating in one. Still, online video-based studies can support rapid hypothesis testing and validation, and they are especially valuable for researchers who lack access to real-world deployment opportunities.

A recent alternative to video-based studies is the use of online, interactive meth-

ods, which allow for interactive and engaging experiences for participants, while still being a lower-cost, scalable alternative to in-person studies. Online interactive methods, such as browser-based simulations [270] add interactivity to the participant experience, enabling researchers to study human responses to robot behavior in real-time. Though differences still exist between real-world and online interactive studies, interactive studies have benefits over passive video watching. For example, users may stay more engaged and focused on the task at hand because of the two-way communication required to interact with the robot.

Chapter 3

Human-Centric Simulation as a Tool for Social Robot Navigation*

Simulation is useful along the whole development cycle of robotic systems including data collection, features development, testing, and deployment [221, 192]. Simulation is key for the verification of safety-critical systems and is particularly relevant for companies that make robots for mainstream audiences [198].

Driven by the gaming industry and demand for autonomous vehicles, the robotics community has recently experienced a rapid increase in the quality and features available in simulation tools. These advancements led to simulation environments for self-driving vehicles [75, 9] and aerial vehicles [178, 232, 98]. Crowd simulations have improved as well [246, 72, 15], although often independently of simulation environments for mobile robots, including environments that build on game engines [116, 142], or state-of-the-art rendering like Gibson [297] or ISAAC [193]. This disconnect has led to a gap in high-fidelity simulation environments for evaluating social robot navigation in pedestrian settings, e.g., service robots that need to operate nearby people

*Parts of this chapter were originally published as Nathan Tsoi, Mohamed Hussein, Jeacy Espinoza, Xavier Ruiz, and Marynel Vázquez. (2020). SEAN: Social Environment for Autonomous Navigation. In *Proceedings of the 8th international conference on human-agent interaction (HAI)*. [268].

Figure 3.1: SEAN's rendering of two virtual worlds: an outdoor city scene and a lab scene, both of which include dynamic pedestrians for studying social robot navigation.

and are subject to social conventions. Our work, depicted in Figure 3.1, is a step towards filling this gap.

We proposed SEAN, a Social Environment for Autonomous Navigation, as an extensible and open-source simulation platform. SEAN includes animated human characters useful for studying human-robot social interactions in the context of navigation. Similar to other recent simulators [178, 232, 116, 142], SEAN leverages modern graphics and physics modeling tools from the gaming industry, providing a flexible development environment in comparison to more traditional robotics simulators like Gazebo [141]. We provide two ready-to-use scenes with components that allow social agents to navigate according to standard pedestrian models. We provide integration with the Robot Operating System (ROS), which allows for compatibility with existing navigation software stacks. An important contribution of this work is a toolkit for repeated execution of navigation tasks and logging of navigation metrics.

3.1 Core Elements of Human-Centric Simulation

SEAN is composed of a collection of tools built around the Unity 3D game engine* and the Robot Operating System (ROS) [212] that allows for control of a mobile robot in a dynamic, simulated human environment. Unity implements the NVIDIA PhysX physics engine, which has been found to provide promising results for robot simulation [142]. Communication between ROS and Unity is implemented as a set

*https://unity.com/

of scripts executed as part of the Unity scripting run-time model and implemented via the ROS# library.[†] SEAN's architecture balances between a) ease of integration with navigation systems (or robot teleoperation) via ROS, b) high visual fidelity for creating immersive environments and enable vision-based navigation methods, and c) a cross-platform ecosystem that supports iterative development. SEAN works in Windows 10 and Ubuntu 18.04 with ROS Melodic.

The key tenets of our approach are usability and flexibility. While these often seem at odds, we seek these goals by providing a set of scenes, robots, and evaluation metrics within the platform to enable users to use the system with minimal preliminary work. Additionally, we maintain an open source repository and supporting documentation to allow the community to improve our social navigation environment.[‡]. Our contributions are an effort to begin to explore the challenging problem of fairly and reproducibly benchmarking algorithms for human-robot social interactions.

Scenes: A scene is a 3D environment in which a robot operates. With our initial release, we provide a high-fidelity model of a lab environment and a larger outdoor city scene (Figure 3.1). Because humans play a key part in the study of robot navigation in these environments, for each scene we have created reasonable start and goal positions for human agents to navigate. To this end, SEAN uses a combination of crowd flow prediction [234] and Unity's built-in path planning algorithm. The system is parameterized such that we can easily deploy an appropriate number of agents given the size and context of the scene. We can also vary the density of pedestrians across experiments in a repeatable manner. SEAN's online documentation explains how to create and modify scenes.

Robots and Sensors: SEAN provides 2 robot models ready to run: a medium size Clearpath Jackal, which is suitable for indoor and flat outdoor environments; and a Warthog with 254mm of ground clearance. The Warthog is more suitable

[†]https://github.com/siemens/ros-sharp
[‡]https://sean.interactive-machines.com/

for outdoor environments due to its bigger size (Figure 3.1). Because neither robot comes equipped with standard sensors, we outfitted them with a simulated Velodyne VLP-16,[§] a LIDAR scanner, and a simulated RGB camera.

Evaluation Toolkit: SEAN's toolkit for evaluating social navigation algorithms centers on the Trial Runner, which enables repeatable and automatic execution of navigation tasks. The Trial Runner performs a *trial* by executing a collection of point-to-point navigation *episodes*. Each episode begins with the Trial runner configuring the scene, actors, and robot positions. Pedestrians are assigned goal positions and a ROS navigation goal is used to indicate the desired final pose for the robot. As the robot navigates, the Trial Runner records relevant metrics. It starts a new episode once the robot has moved to a sufficiently close location to the destination or the episode times out. While the initial conditions for each episode are random by default, they are recorded at the beginning of an episode. This allows to replay the episode for fair comparisons of navigation methods.

SEAN currently tracks the following navigation metrics: whether or not the robot reached the goal position, how long it took to reach the final position, collisions with static objects, and the robot's final distance to the goal position. The latter metric is particularly useful for comparison in challenging tasks. In addition, SEAN can continuously track metrics related to social interactions. Currently, we track the closest distance between the robot and pedestrians, as well as the number of collisions with pedestrians, which are recorded separately from collisions with all other objects. These metrics are common in the social navigation literature [261, 194, 246, 173, 207] and serve as a starting point for comparisons among navigation approaches. We plan to expand this set of metrics in the future.

Table 3.1 provides example results by the Trial Runner for the ROS Navigation Stack [99], which was minimally tuned, and a teleoperated robot. Localization for the

[§]https://github.com/Field-Robotics-Japan/unit04_unity/

Table 3.1: Sample **J**ackal, **W**arthog results, via the ROS Nav. Stack, or teleoperated*. $\mu \pm \sigma$ over 10 episodes.

Scene	Robot	Elapsed (sec.)	Complete	Final Dist (m)	Ped. Dist (m)	Collisions
Lab	J	24.51 ± 19.36	60%	2.26 ± 2.92m	1.54 ± 1.76m	7.1 ± 9.4
Lab	J *	21.6 ± 28.08	88%	1.14 ± 1.99	0.92 ± 1.16	4.63 ± 5.83
City	J	37.09 ± 13.74	29%	9.54 ± 8.94	0.64 ± 0.42	20 ± 30.83
City	J *	38.54 ± 29.5	80%	4.59 ± 11.87	1.06 ± 0.67	3.1 ± 7.58
City	W *	31.7 ± 20.94	100%	0.48 ± 0.01	2.27 ± 1.08	0 ± 0

ROS Nav. Stack was performed via SLAM [96]. Low performance is attributed to not taking into account human actors during mapping and overly conservative navigation behavior in the dynamic environments [261].

Teleoperation was implemented through a ROS node that connected to a gamepad controller. The teleoperated Jackal did not reach 100% of the target goals because people blocked its way and the episodes timed out. Nonetheless, teleoperation was an interesting baseline for automated methods. It can also serve to gather demonstrations or human preferences for navigation trajectories in the future [145, 278].

3.2 Summary

Simulation plays a critical role in the development and evaluation of social robot navigation systems. The Social Environment for Autonomous Navigation (SEAN) platform supports systematic and repeatable performance evaluation while accelerating development by enabling early identification of navigation failures. While simulation is not a substitute for real-world testing, SEAN is designed to complement physical experiments by offering a controllable, reproducible, and human-centric environment for studying human-robot interactions. This chapter introduced the motivation for building a simulation platform focused on social navigation. It outlined the key design decisions behind SEAN, resulting in support for high visual fidelity, seamless integration with the Robot Operating System (ROS) [212], and the inclusion of dynamic environments that incorporate virtual pedestrians.

Chapter 4

Systems for Simulating and Evaluating Social Robot Navigation*

Simulating socially competent robot navigation requires more than modeling physical obstacles; it necessitates capturing the dynamics of interaction between agents, particularly between a robot and a human pedestrian, within a shared space. This foundational requirement distinguishes simulators for social navigation from traditional navigation, where other entities are treated as passive or purely reactive. Human-centric simulations aim to recreate scenarios in which the robot must perceive, interpret, and respond to human behaviors such as yielding the right-of-way, walking alongside a person, or avoiding interruption of human activities. To support such tasks, the designers of social navigation simulators must model not only the physical trajectories of agents, but also consider the social elements including verbal and non-verbal signals inherent in social human behavior.

*Parts of this chapter were originally published as 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. In *IEEE Robotics and Automation Letters (RA-L)* [273] **and** Qiping Zhang*, Nathan Tsoi*, and Marynel Vázquez. (2023). SEAN-VR: An Immersive Virtual Reality Experience for Evaluating Social Robot Navigation. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)* [308]. * indicates equal contribution.

SEAN 2.0 builds upon the preliminary research around the SEAN simulator [268] to create a comprehensive simulation system that supports the design and evaluation of social navigation algorithms.

4.1 Introduction

While a significant amount of work has been done to enable robots to effectively move in human environments [174], prior work has largely been fragmented by interaction scenarios [90]. For example, past work has focused on studying navigation in scenarios where robots cross human paths [122, 58], approach users [228, 82] or groups [266, 303], and move in crowded environments [256, 263]. This fragmentation raises the question: *how can we build robot navigation systems that handle different social contexts*?

Inspired by work in social psychology, we propose to reason about context for social navigation in terms of *social situations*, which consider the interplay between robot task and environmental factors. Social situations may occur in a given interaction *scenario*, consisting of three key elements: 1) the physical *environment* (such as a lab or warehouse), 2) *pedestrian behavior* in the environment, and 3) the robot's navigation *task* (involving motion from a start to a goal pose).

We contribute an open-source system, the Social Environment for Autonomous Navigation (SEAN) 2.0, for training and benchmarking social navigation algorithms. Unlike other robotics simulation environments capable of high visual fidelity, such as [75], SEAN 2.0 is designed so that robots can experience a range of different pedestrian behaviors, which result in varying social situations. To ground the concept of social situations in our system, we propose logic-based definitions for five social situations relevant to navigation. These definitions serve as situation classifiers in SEAN 2.0.

One approach to specifying pedestrian behaviors for simulated interactions in

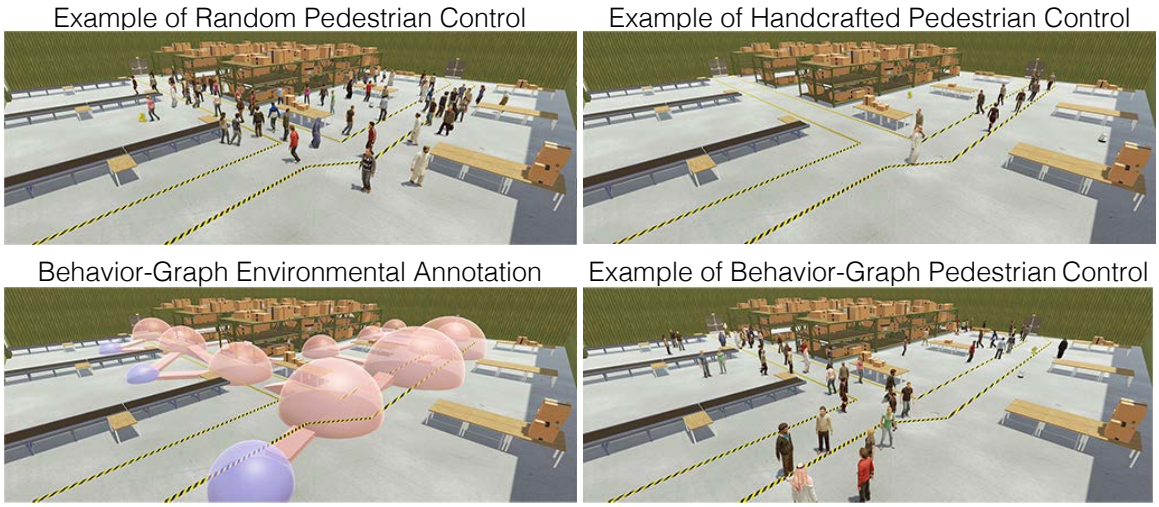


Figure 4.1: Different methods for specifying pedestrian behaviors in SEAN 2.0. The Behavior Graph method is a novel approach that uses an environmental graph-based annotation (bottom-left) to generate behavior (bottom-right).

social navigation is to handcraft a starting pose and a goal pose for each pedestrian in the scene. Handcrafting pedestrian behavior is time-consuming and specific to a single implementation, as evidenced by the limited number of social situations commonly employed when evaluating navigation policies [90]. Randomly choosing start and goal poses is an easy alternative to handcrafting; yet, as our experiments show, it is less likely to result in varied social situations in practice.

As part of SEAN 2.0, we propose a novel method for specifying pedestrian behavior based on a *Behavior Graph* annotation in the physical environment (Figure 4.1, bottom). We define the Behavior Graph such that nodes represent either static group formations or navigation waypoints, and compute flow between nodes based on graph parameters. Pedestrians traverse the scene by walking between different nodes in the graph. This creates opportunities for the robot to experience different social situations while avoiding time-intensive handcrafting of pedestrian motion (as in SEAN 1.0 [267]).

SEAN 2.0 also provides a range of components to enable the training and benchmarking of navigation policies, including vision and depth sensors, several physical

environments, different means of specifying robot tasks, and a range of evaluation metrics. To validate that SEAN 2.0 would be useful to the robotics community, we collected feedback from 7 roboticists who were early users of the system and incorporated their feedback in the final version of SEAN 2.0.

As part of our experimental evaluation, we studied the distribution of social situations that emerged in datasets gathered with different methods of pedestrian control according to logic-based social situation classifiers. We found that the Behavior Graph data resulted in more varied social situations than the data generated with handcrafted or random pedestrian motion. Also, policies trained on Behavior Graph data outperformed other learned policies that were trained using alternative methods for pedestrian behavior generation in SEAN 2.0. Finally, our experiments showed that analyzing navigation policies by social situation can reveal new insights about robot policy performance.

The five main contributions of this chapter are: 1) SEAN 2.0, a novel system for training and benchmarking social navigation systems; 2) a logical formalization of social situations; 3) multiple methods for pedestrian behavior generation including a novel Behavior Graph approach; 4) validation that our system is useful to users outside our team via feedback from other roboticists; and 5) experiments that show the usefulness of social situations in SEAN 2.0 towards the training and evaluation of robot navigation systems.

4.2 Related Work

4.2.1 Simulation Frameworks for Social Navigation

Our work builds on developments in robotics simulators. Recently, robotics simulators such as CARLA [75], iGibson [156], and Habitat [170] have focused on creating high-fidelity environments and have started to provide basic control of individual

pedestrians. Several works have extended the MORSE robotics simulator [78], adding humans that react to a robot [129] and humans in wide areas or narrow passages [81]. The MORSE simulator integrates with the Robot Operating System (ROS); however, visual fidelity is low in comparison to simulators based on game engines such as CARLA.

Crowd simulation frameworks such as Nomad [45], PED-SIM [92], and Menge [72] incorporate methods of individual and group behavior control into a system, but do not integrate robotics platforms to train and evaluate social robot navigation systems. Their strength lies in simulating pedestrian motion, not in integrating them in realistic physical environments or in the visualisation of pedestrians necessary for training state-of-the-art vision-based robotics algorithms.

Our prior work, SEAN 1.0 [267], was designed for training and evaluating social navigation algorithms. It supported integration with ROS and high-quality rendering of virtual pedestrians. However, it only provided simple waypoint navigation for the pedestrians, requiring time-consuming handcrafting of their behavior. More specifically, SEAN 1.0 allowed users to specify pedestrians' start and goal locations and implemented only one example set of such handcrafted pedestrian start and goal location annotations in three physical environments. SEAN 2.0 continues to provide the same ability to customize start and goal locations for pedestrians, but provides 13 sets of handcrafted start and goal positions per environment (39 total sets) across 5 social situations. This allows users of SEAN 2.0 to generate more varied pedestrian behavior with the handcrafted approach for pedestrian control out of the box. Further, we found that our proposed Behavior Graph, a novel method for specifying pedestrian behavior, was a superior method for training a navigation policy. See Section 4.5 for more details.

An alternative approach to evaluating social navigation policies is using prerecorded pedestrian trajectories as in SocNavBench [33]. Prerecorded pedestrian tra-

jectories offer realistic motion, but are not reactive to the robot during policy rollout. Our work complements [33] by allowing for dynamic human-robot interactions. To our knowledge, SEAN 2.0 is the only robotics simulation environment capable of high visual fidelity that provides easy-to-customize, dynamic pedestrian behaviors, including group formations.

4.2.2 Modeling Pedestrian Behaviors

Algorithms for the animation and control of virtual characters have been studied by different disciplines such as computer graphics [130], cognitive science [293], and computer vision [224]. The generation of collective behaviors has traditionally focused on modeling individual members of a crowd to elicit human-like behavior from the group. For example, flocks of animals inspired Reynolds et al.'s early work on modeling groups of pedestrians [216]. Pelechano et al. [200] focused their effort on imbuing human-like perception and decision making capabilities into individual agents to elicit more realistic group behavior. Collective behavior conditioned on a given environment can rely on annotations of the physical space, such as semantically relevant descriptors [203]. We take inspiration from these ideas and utilize environmental annotations in our proposed Behavior Graph.

To produce phenomena observed in human navigation, collision avoidance methods for individual agents are often used to complement collective behavior. Such methods, often referred to as microscopic models, include the Social Forces model [106] and the velocity obstacle method of the ORCA model [277]. While ORCA's primary benefit is collision-free movement between a large number of agents, the Social Forces model is easily extended by the addition of new forces. For this reason, our proposed Behavior Graph relies on the Social Forces model. In the future, SEAN 2.0 could be extended with other microscopic models.

4.3 Formalizing Social Navigation Context

There are many definitions of context in disciplines related to social navigation. For example, in social signal processing, context has been defined as the who, what, when, where, and why of interactions [282]. In Human-Robot Interaction (HRI), the term context has been used to refer to high-level environmental concepts such as an art gallery or a dining hall [189]. Likewise, context has been used to describe the relationship between agents (human and robot) in the scene, such as agents in a static group formation or standing in a line [8]. Task-based context has also been explored in HRI, often in the domain of engagement [50].

In this work, we propose to reason about context in social robot navigation based on the notion of social situations proposed by Argyle et al [13] in psychology. Those authors studied the interplay between internal and external determinants for human behavior. In their work, social situations encompass the intrinsic *goal* of a person and the extrinsic *environment* in which this person acts. Individuals' goals arise from an underlying drive that satisfies a specific need.

Consequently, we propose to reason about social situations in robot navigation as a construct that considers the interplay between a robot's *task* and *environmental* factors. Consider, for example, a situation when a robot must cross a pedestrian path to reach the other side [201]. This situation arises from the combination of the environmental factor of pedestrian traffic and the robot's start location relative to a navigation goal. Similarly, a robot approaching a group [303] could be considered another example of a social situation. In this case, the task is navigating to a specific goal position in a conversational group and depends on people's spatial arrangement.

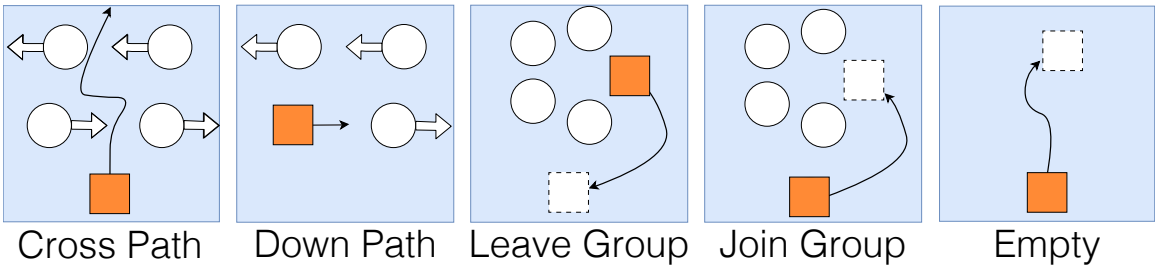


Figure 4.2: A brief visual description of *social situations*. Pedestrians are denoted as white circles and the robot as an orange square.

4.3.1 Logical Expressions for Social Situations

This section operationalizes the proposed notion of social situations in relation to five instances relevant to navigation, as shown in Figure 4.2. In particular, we consider situations that involve both pedestrians in motion and static group formations. Although our proposed set of social situations may not be complete for all robot navigation applications, it helps demonstrate the value of formalizing social situations for mobile robotics.

We use logic to formally define the proposed situations. The domain of predicates, defined below, consists of vectors in $\mathbb{R}^2$ that represent position, orientation (as unitary direction vectors), or velocity. These vectors can be provided in simulation or estimated in the real world.

- **Nearby Agent:** $Near(\boldsymbol{x_1}, \boldsymbol{x_2})$ is true when two agents at positions $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are separated by $||\boldsymbol{x_1} - \boldsymbol{x_2}|| < \mathsf{D}$.

- **Group Member:** $Member(\boldsymbol{x}, \boldsymbol{g})$ is true when agent $\boldsymbol{x}$ is a member of the group with center at position $\boldsymbol{g}$.

- **Walking:** $Walking(\boldsymbol{v})$ is true when an agent is moving at a velocity $\boldsymbol{v}$ where $||\boldsymbol{v}|| > \mathsf{V}$.

- **Perpendicular Trajectory:** $PerpTraj(\boldsymbol{d_1}, \boldsymbol{d_2})$ is true when the orientations

of two agents $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ are perpendicular within an error of $\pm\mathsf{A}$ rad:

$$cos(\frac{\pi}{2} + \mathsf{A}) \leq \frac{\boldsymbol{d_1} \cdot \boldsymbol{d_2}}{||\boldsymbol{d_1}||||\boldsymbol{d_2}||} \leq cos(\frac{\pi}{2} - \mathsf{A})$$

- **Parallel Trajectory:** $ParTraj(\boldsymbol{d_1}, \boldsymbol{d_2})$ is true when the orientations of two agents $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ are parallel within an error of $\pm\mathsf{A}$ rad:

$$cos(\mathsf{A}) \leq \frac{\boldsymbol{d_1} \cdot \boldsymbol{d_2}}{||\boldsymbol{d_1}||||\boldsymbol{d_2}||}$$

All predicates are defined by simple geometric relationships except for group membership. We assume group membership is provided by the simulator or a method such as [245, 255] which reasons about conversational formations [136]. Section 4.4.7 provides more details of our specific choice of other parameters for these predicates.

The five *Social Situations* (Figure 4.2) are expressed as:

Cross Path: a robot is at a position $\boldsymbol{x_r}$ with orientation $\boldsymbol{d_r}$. Also, it is nearby an agent at $\boldsymbol{x_a}$, moving at velocity $\boldsymbol{v_a}$, with orientation $\boldsymbol{d_a}$ perpendicular to $\boldsymbol{d_r}$.

$$\begin{aligned} &CrossPath(\boldsymbol{x_r}, \boldsymbol{d_r}, \boldsymbol{x_a}, \boldsymbol{v_a}, \boldsymbol{d_a}) \equiv \\ &Near(\boldsymbol{x_r}, \boldsymbol{x_a}) \wedge Walking(\boldsymbol{v_a}) \wedge PerpTraj(\boldsymbol{d_r}, \boldsymbol{d_a}) \end{aligned} \tag{4.1}$$

Down Path: a robot is at position $\boldsymbol{x_r}$ with orientation $\boldsymbol{d_r}$. Also, it is nearby an agent at $\boldsymbol{x_a}$, moving at velocity $\boldsymbol{v_a}$, with orientation $\boldsymbol{d_a}$ parallel to $\boldsymbol{d_r}$.

$$\begin{aligned} &DownPath(\boldsymbol{x_r}, \boldsymbol{d_r}, \boldsymbol{x_a}, \boldsymbol{v_a}, \boldsymbol{d_a}) \equiv \\ &Near(\boldsymbol{x_r}, \boldsymbol{x_a}) \wedge Walking(\boldsymbol{v_a}) \wedge ParTraj(\boldsymbol{d_r}, \boldsymbol{d_a}) \end{aligned} \tag{4.2}$$

Joining Group: a robot at position $\boldsymbol{x_r}$ has a navigation goal $\boldsymbol{x'_r}$, which corresponds to a location that would make the robot a member of a group with a center at $\boldsymbol{g}$. The robot is also near an agent at $\boldsymbol{x_a}$, which is a member of the same group. Note that

once the robot arrives at the goal, *JoinGroup* is no longer true.

$$\begin{aligned} &JoinGroup(\boldsymbol{x_r}, \boldsymbol{x_a}, \boldsymbol{x'_r}, \boldsymbol{g}) \equiv Near(\boldsymbol{x_r}, \boldsymbol{x_a}) \wedge \\ &Member(\boldsymbol{x_a}, \boldsymbol{g}) \wedge Member(\boldsymbol{x'_r}, \boldsymbol{g}) \wedge \neg(\boldsymbol{x_r} = \boldsymbol{x'_r}) \end{aligned} \tag{4.3}$$

Leave Group: a robot that is currently at a position $\boldsymbol{x_r}$, had a starting position $\boldsymbol{x''_r}$ which made it a member of a group with a center at $\boldsymbol{g}$. The robot is near an agent located at $\boldsymbol{x_a}$, which is a member of the same group.

$$\begin{aligned} &LeaveGroup(\boldsymbol{x_r}, \boldsymbol{x_a}, \boldsymbol{x''_r}, \boldsymbol{g}) \equiv \\ &Near(\boldsymbol{x_r}, \boldsymbol{x_a}) \wedge Member(\boldsymbol{x_a}, \boldsymbol{g}) \wedge Member(\boldsymbol{x''_r}, \boldsymbol{g}) \end{aligned} \tag{4.4}$$

Empty: a robot at position $\boldsymbol{x_r}$ has no other agents nearby. Let X be the set of positions for all other agents in the environment, then:

$$Empty(X, \boldsymbol{x_r}) \equiv \forall \boldsymbol{x} \in X, \neg Near(\boldsymbol{x}, \boldsymbol{x_r}) \tag{4.5}$$

The satisfiability of these logical expressions depends on the agents in the environment. Only when a sufficient number of agents are present, both moving and in static group formations, are all non-empty expressions satisfiable. For example, in environments without group formations, *JoinGroup* and *LeaveGroup* are not satisfiable. The size of the environment also impacts satisfiability. For example, consider an environment with a robot and a pedestrian. The *Empty* proposition is not satisfiable if the navigable space in the environment is smaller than the nearby distance D.

4.4 SEAN 2.0 System

SEAN 2.0 builds on our prior work, the Social Environment for Autonomous Navigation version 1.0 [267]. SEAN 1.0 and SEAN 2.0 both use the Unity game engine

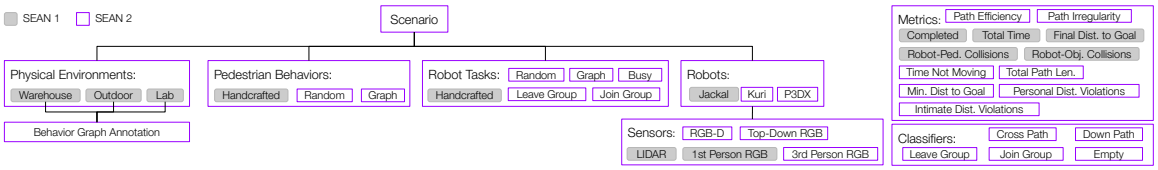


Figure 4.3: SEAN 2.0 system architecture including components that were re-used or adapted from SEAN 1.0 (rounded box, grey) and new to SEAN 2.0 (purple). Connections denote relationships between components in a Scenario. The Scenario, Metrics, and Classifiers are part of the SEAN 2.0 Unity API and exist for all scenes. The SEAN 1.0 Trial Runner [267] is superseded by Robot Tasks and the Metrics system in SEAN 2.0. Warthog is the only robot in SEAN 1.0 which is not in SEAN 2.0 due to it's unwieldy size relative to people. See the text for details.

and the Robot Operating System (ROS) [212] as underlying technologies, and can be integrated with online interactive surveys [269]. The core innovation in SEAN 2.0 is a variety of methods for specifying pedestrian behavior, including a novel Behavior Graph approach that induces the proposed social situations described in Section 4.3. In addition, SEAN 2.0 provides improved simulated sensors to facilitate vision-based, trained policies and two new features. One new feature is a set of logic-based social situation classifiers and the other is a revamped software architecture. Our goal was to make SEAN 2.0 easily configurable for users of its graphical user interface and easily extensible for users of its programming interface. Figure 4.3 shows the system components of SEAN 2.0, including those that are reused or adapted from SEAN 1.0.

4.4.1 Software Design

The usability of a software system depends directly on the underlying design decisions [26]. Therefore, we designed SEAN 2.0 following the singleton design pattern [237] and taking a convention over configuration approach [57].

The SEAN Unity API is implemented as a singleton GameObject, through which all key components can be accessed. Unlike other scripts that may be added or removed from the Unity scene at various times, this object exists throughout the duration of the simulation and provides the logic necessary to wrap other elements

that may be removed or added at various times. The singleton GameObject includes all elements discussed below such as pedestrian behaviors, robot tasks, social situation classifiers, metrics, and other utilities such as a simulated clock and a tool for creating ROS maps. The singleton GameObject can be added to any scene, thereby making it compatible with SEAN 2.0.

Classes in SEAN 2.0 use a convention over configuration approach by providing sensible defaults [57]. This makes our system easier to use than SEAN 1.0, which had an ad-hoc design. Feedback from early users of SEAN 2.0 indicates that our design choices provide a better user experience than SEAN 1.0 (see Section 4.5.1 for more details).

4.4.2 System Architecture

Our system architecture was designed to encapsulate the elements of a *Scenario*, which consists of the *Physical Environment*, *Pedestrian Behavior*, and *Robot Task*. These elements are shown in Figure 4.3 and correspond to objects in the SEAN 2.0 singleton GameObject. *Physical Environments* are locations in which a scenario occurs, such as a warehouse. *Pedestrian Behaviors* specify pedestrian motion. *Robot Tasks* specify a robot's start pose and goal pose.

4.4.3 Physical Environments

Environments correspond to the physical, static elements in a scenario and are composed of 3D meshes, textures, lights, and colliders that construct a Unity Scene. Static elements of the environment constrain agent motion and define a navigable area on the ground plane. SEAN 2.0 includes the *warehouse*, *lab*, and *outdoor* environments from SEAN 1.0 with annotations for our new pedestrian behaviors.

4.4.4 Pedestrian Behaviors

SEAN 2.0 supports three different approaches to high-level pedestrian control: *random*, *handcrafted*, and *graph*. High-level pedestrian motions are defined by their start and goal poses. They also depend on a low-level collision avoidance mechanism, which relies on the Social Forces model [106]. We extend the Social Forces model with consideration for pedestrians moving along one side of a hallway [152] and to stochastically vary the distances that they prefer to maintain from the robot [289]. The next sections describe the three methods for pedestrian behavior generation in SEAN 2.0.

Random Pedestrian Behavior: Start and goal locations are randomly chosen on the environment's navigable plane.

Implementation. There are no parameters other than the number of pedestrians in the scenario. This approach for pedestrian behavior generation is the easiest to implement, but no group formations are created by this method. We use this behavior as a baseline in our experiments to evaluate the effect of pedestrian density on policy performance relative to the other methods of pedestrian control in SEAN 2.0.

Handcrafted Pedestrian Behavior: Start, goal, and intermediary waypoint poses for the pedestrians are chosen manually in each environment and may be designed to resemble specific social situations. This is the most granular method of pedestrian control. The main challenge with this method is twofold: 1) it can be time-consuming, and 2) the many low-level decisions one has to make in regard to goal placement may not align with the intended high-level behavior. This challenge is further discussed in Section 4.5.

Implementation. In each of the environments, we implemented 13 unique sets of start and goal poses for pedestrians across five handcrafted scenarios to resemble each social situation described in Section 4.3.1. Pedestrians in the Join Group and Leave Group scenarios are configured in a static group formation typical of conversational

encounters [136]. In the Cross Path and Down Path scenarios, we specified navigation waypoints along a path so that the robot can cross parallel or travel perpendicular to the path of pedestrian waypoints. While choosing group locations and pedestrian waypoints, we also chose corresponding poses for the robot in specific tasks, described in Section 4.4.5.

Some handcrafted pedestrian behaviors are parameterized by the location of static group formations. Given a group center, we set the poses of individual group members by mimicking conversational formations in the Cocktail Party dataset [306] in a manner similar to [303].

Graph-based Pedestrian Behavior: We propose using a directed graph abstraction, which we call the Behavior Graph, to specify collective pedestrian behaviors. A graph annotation is overlaid in the environment (Figure 4.1, bottom-left). The graph parameterizes pedestrian motion via two types of nodes that determine pedestrian behavior. Nodes serve as either 1) navigational waypoints (through which an unrestricted number of pedestrians can continuously flow) or 2) as a location for a conversational group formation (where a number of pedestrians can enter a static group formation for a specific duration). The graph edges connect nodes that pedestrians can navigate through.

On initialization, individual pedestrians are stochastically assigned to starting locations, which correspond to specific nodes from the Behavior Graph annotation. During the simulation, pedestrians without an assigned goal position are first assigned to group nodes, until all group nodes are at capacity. When pedestrians assigned to group nodes reach their destinations, they remain at the group for a given duration. Once all the group nodes have reached capacity, the remaining pedestrians are stochastically assigned to waypoint nodes. This allows the simulation to maintain group formations and it allows pedestrians to automatically transition between navigation and being part of conversational groups, which is not easily achievable with

the random or the handcrafted pedestrian behaviors in SEAN 2.0.

The location of graph annotation nodes and the parameterization of accompanying edges can be used to modify pedestrian congestion in the environment. Depending on the graph's structure, certain areas may have more or less pedestrian congestion than what a user desires. Pedestrians can be directed away from the congested area by using edge weights associated with low or uni-directional flow.

Implementation. SEAN 2.0 provides one Behavior Graph annotation for each environment (warehouse, lab, and outdoor). Each Behavior Graph annotation consists of a graph where every pair of adjacent nodes is connected by two directed edges. Edges are weighted to control pedestrian congestion using one of three costs: c_{min}, 1, or c_{max}, where $1 < c_{max}$ and $0 < c_{min} < 1$. For example, consider the edges between nodes u and v. Users of SEAN 2.0 can constrain pedestrian flow using 4 sets of edge weights $\langle c_{uv}, c_{vu} \rangle$: $\langle 1, 1 \rangle$ for there to be medium flow between the nodes, $\langle c_{min}, c_{min} \rangle$ for high flow, $\langle c_{max}, c_{max} \rangle$ for low flow, and either $\langle 1, c_{max} \rangle$ or $\langle c_{max}, 1 \rangle$ for uni-directional flow.

The path to a pedestrian's goal node is computed over the edges between waypoint nodes using Dijkstra's algorithm [74], which considers the different edge costs and ensures that pedestrians do not disrupt groups. Pedestrians traverse the computed path using the Social Forces model [106], which allows them to perform local collision avoidance.

4.4.5 Robot Tasks

Tasks specify the robot's start (A) and goal (B) poses. Robot tasks can leverage ground truth information from the simulation, like group locations, to choose robot poses that may result in specific social situations.

SEAN 2.0 implements the following robot tasks:

- *RandomABNav*: uniformly samples a start and a goal pose for the robot from

the navigable plane in the environment.

- *BusyABNav*: samples a start and a goal pose for the robot near the largest cluster of pedestrians. Pedestrian poses from SEAN 2.0 are clustered via k-means.

- *Join Group*: a group center is sampled from graph nodes associated with group formations. Then, a point in the group is sampled as the goal location for the robot and a further away point is sampled as its start pose.

- *Leave Group*: a group center is sampled from graph nodes associated with group formations. Then, a point in the group is sampled for the start location for the robot and a further away point is sampled for the goal pose.

- *Handcrafted*: assigns a start and goal pose specifically chosen by a scenario designer. We implemented 5 handcrafted tasks corresponding to the Cross Path, Down Path, Join Group, Leave Group and Empty social situations.

Handcrafted tasks can only be used with handcrafted scenarios, as both robot and pedestrian poses are chosen by the scenario designer. Join and Leave Group tasks can only be used with the Behavior Graph method of pedestrian control as they depend on nodes in the graph. All other tasks are decoupled from the method of pedestrian control.

Tasks can be designed such that the robot is likely to experience a certain social situation. However, the social situations that we consider in this chapter only occur upon satisfaction of the propositions described in Section 4.3. For example, a robot aiming to complete a Join Group task is not guaranteed to enter a Join Group social situation, but given the task design it is likely to experience this situation.

4.4.6 Sensor Integration

SEAN 2.0 provides a simulated RGB camera and a simulated depth sensor. By convention, each robot implementation includes one simulated depth sensor and three RGB cameras. The depth sensor is positioned in a first-person perspective and the RGB cameras provide three angles: first-person, third-person, and top-down. This default configuration ensures that a standard API is available when accessing sensor data for any robot, allowing for comparison across robots.

4.4.7 Social Situation Classifiers

As part of SEAN 2.0, we provide five rule-based classifiers that implement the propositional predicates defining our five social situations from Section 4.3. The predicates that we define use parameters derived from Hall's work in proxemics [101] where applicable. For example, we consider two agents to be "nearby" when the distance between them is less than two times their personal space (1.2m). Our experiments in Section 4.5 indicate that social situation classifiers can help users better understand the distribution of data resulting from different pedestrian control methods. They can also help users identify how well a robot navigation policy can learn from and perform in different social situations.

4.4.8 Metrics

We implement a range of social navigation metrics which are aggregated over the duration of a Robot Task, including:

- *Path Efficiency*: ratio between the traveled and geodesic distance of the search-based path from the starting position.
- *Time Not Moving*: seconds that the robot was not moving.

- *Intimate Distance Violation*: number of times the robot approached a pedestrian within a distance of 0.45m.

Our documentation details all of the other metrics implemented in our system.* The underlying data needed to compute these metrics is available from the SEAN 2.0 API.

4.5 SEAN 2.0 System Evaluation

We first studied the usefulness of SEAN 2.0 for the robotics community by collecting and incorporating feedback from seven researchers who used our system. Then, we studied how robot navigation datasets generated via SEAN 2.0 affected the training and evaluation of social navigation algorithms in environments with varying pedestrian behavior.

4.5.1 User Feedback About SEAN 2.0

We initially gathered feedback about SEAN 2.0 from four robotics researchers at Yale University, University of Washington, University of Massachusetts Amherst, and Carnegie Mellon University who were previously unfamiliar with the present work.† They installed and used SEAN 2.0 and then provided written feedback. Although one researcher noted that occasionally pedestrians exhibited "unnatural behavior like running into each other, the robot, and obstacles," in general, the feedback was broadly positive. For example, one person said that the "lab scene and the warehouse scene both looked good." Another researcher familiar with the dynamics of the Kuri robot – one of the robots included in SEAN 2.0 – noted that the base dynamics were realistic and Kuri followed the navigation path "in a way that looks much like the real platform" when controlled by the ROS Navigation Stack. The one researcher who

*https://sean.interactive-machines.com/docs/metrics

†Our protocol was reviewed by our IRB and exempted from annual review.

had prior experience with SEAN 1.0 noted that SEAN 2.0 was "definitely easier to navigate and more user-friendly in terms of getting started with the simulator." This researcher later informed us via personal communication that they were able to successfully set up and use SEAN 2.0 to submit a paper for publication without our involvement. In contrast, when the user was using SEAN 1.0, we made many small changes to the SEAN 1.0 system to support their workflow.

Based on the helpful feedback from our users, we made a number of improvements to SEAN 2.0. First, we tuned the parameters of our social forces model to decrease pedestrian-on-pedestrian collisions. Second, we improved the documentation for system setup and the integration of new robot policies. Third, we fixed several bugs, including adding localization information for two robot components and resolving an edge case where the robot's physics simulation was unrealistic when it collided with a pedestrian.

After incorporating this initial feedback, we shared the system with three more researchers at Yale University. One said that the system was "easy to use," another mentioned that they did not run into any issues while using the system, and another noted that "I can see how one would be able to implement their own controller through this [system]."

4.5.2 Emergence of Social Situations

We studied the distributions of social situations in three datasets collected from SEAN 2.0, where each dataset corresponded to a different method of pedestrian control.

Data Collection

Sensor data and robot control information were collected in SEAN 2.0 while a human expert navigated the robot in the warehouse environment using a joystick. The expert was first familiarized with the analog joystick controls and then directed to

navigate the robot to the goal in a polite manner, similar to the way they would navigate in real life. One hour of data was collected for each of the three proposed methods of pedestrian control: random, handcrafted, and graph-based. The random and graph methods were configured with an equal number of pedestrians ($n = 62$). The handcrafted scenarios had a variable number depending on the scenario's design, but on average, they had far fewer pedestrians ($n \approx [5, 10]$) due to the time and effort required to manually specify behaviors.

The social situation classifiers described in Section 4.4.7 were parameterized as follows. The maximum distance at which two agents were considered nearby was $\mathsf{D} = 2.4$m, within Hall's personal space [101]. The minimum speed of an agent considered to be walking was $\mathsf{V} = 1.4$m/s, near the average human walking speed. We set a small error value $\mathsf{A} = \frac{\pi}{12} = 15^\circ$ within which a pedestrian trajectory was considered parallel or perpendicular to the robot.

During data collection, robot tasks were chosen to induce the robot to experience a uniform distribution of social situations depending on the method of pedestrian control. For the random method, we used the RandomABNav task for the entire hour of data collection. For the handcrafted method, we collected 12 minutes of data from each of the 5 robot tasks designed to correspond to the 5 social situations formalized in Section 4.3. For the graph, we collected data evenly between the Join Group and Leave Group robot tasks.

The collected data was divided into examples composed of 5 depth images, a local navigation plan with 5 points of the expert's most recent trajectory sampled at 1hz,

Table 4.1: Percentage of examples belonging to each social situation.

	Social Situation				
Behavior	Cross Path	Down Path	Join Group	Leave Group	Empty
Random	14.51%	24.68%	0%	0%	66.44%
Handcrafted	0.73%	0.74%	3.43%	13.48%	81.62%
Graph	13.24%	22.08%	12.09%	19.51%	33.08%

a search-based global plan with the 10 nearest points between the robot position and the goal (where each point was at least 0.5m apart). Each example also contained five boolean flags corresponding to the social situation classifiers.

Results

Table 4.1 shows the distribution of social situations between the three datasets generated using SEAN 2.0 with different methods of pedestrian control. The number of examples in which the robot experienced the five social situations were not evenly distributed with the Handcrafted Behavior approach, even though we collected data for an equal amount of time in each Handcrafted scenario. In the Random pedestrian dataset, groups did not form, so no Join Group or Leave Group social situations were experienced. However, social situations were more evenly distributed among the Cross Path, Down Path, and Empty scenarios in comparison to data from the Handcrafted scenarios. The Behavior Graph dataset led to a more uniform distribution of social situations than the Handcrafted or Random datasets.

4.5.3 Evaluation of Navigation Algorithms

Experimental Setup

Whereas SEAN 1.0 [267] was evaluated using only the ROS navigation stack, we evaluated SEAN 2.0 as a benchmarking platform using two robot navigation methods: the ROS navigation stack with social cost layers [162], and a neural network controller following [207] (using depth images rather than LIDAR as input). For the latter method, we trained three controllers using each of the datasets created with the three different methods of pedestrian control outlined in Section 4.5.2. This was not possible with SEAN 1.0 because SEAN 1.0 only supported one method of pedestrian control.

We trained the neural network controllers end-to-end using PyTorch with equally

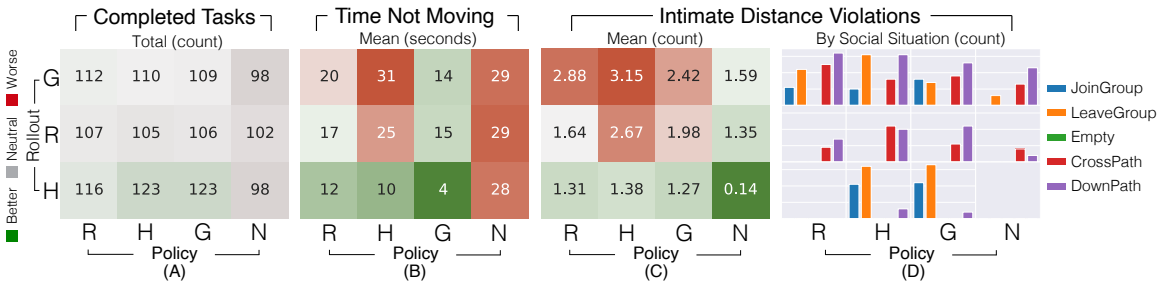


Figure 4.4: Select results for 4 policies in 3 scenarios aggregated over 123 episodes. The three learned policies are trained on data from the Behavior Graph (G), Random (R), and Handcrafted (H) methods of pedestrian control. We also evaluate the ROS Navigation (N) policy with social cost layers [162]. Three metrics are shown: Completed Tasks (A), Time Not Moving (B), and Intimate Distance Violations (C and D). Values for Time Not Moving are computed in seconds and averaged over all episodes. All other plots show the total or average count of the occurrences of the metric over all episodes. The percentage gain for a specific metric can be calculated between grid cells. For example, the Behavior Graph policy spends 55% less Time Not Moving than the Handcrafted policy when rolled out in the Behavior Graph environment.

weighted losses for the local planner and velocity controller modules. We used the AdamW optimizer with the default parameters, lr=0.001, wd=0.010, and a batch size of 1024 for all experiments. A search over a range of batch sizes ($\{128, 256, 512, 1024\}$) revealed similar performance so we chose a batch size which used the maximum amount of GPU memory, effectively decreasing the time required to train over a single epoch of the data. The training machine had 128GB of RAM, an Intel Xeon W-2155 CPU clocked at 3.30GHz, and an NVIDIA RTX 2080TI GPU. An early stopping window of 50 epochs was used after trying between 10 and 100 epochs. The loss did not always reach a local minima at 10 epochs, but the loss stabilized far before 100.

Results

Neural network navigation policies trained on the Handcrafted (H) dataset completed a similar number of episodes as policies trained on the Behavior Graph (G) data. Across all rollout environments, the average number of Completed Tasks was 112.7

for both H and G, as computed over the columns of Figure 4.4A. However, polices trained on Handcrafted (H) data spent more time on average not moving than the policies trained using the Behavior Graph (G) method. The policy trained using Random (R) pedestrians paused for less time on average than the policy trained on Handcrafted (H) data, but paused for more time on average than the policy trained using the Behavior Graph (G). All learned policies spent less time not moving than the ROS Navigation Stack (N). The average Time Not Moving per policy was 16.3s for R, 22s for H, 11s for G, and 28.7s for N, as computed over the columns of Figure 4.4B.

Intimate Distance Violations were more numerous for policies rolled out in the scenarios with the Behavior Graph (G) compared to the scenarios utilizing Random (R) and Handcrafted (H) pedestrians. The average number of Intimate Distance Violations per rollout scenario were 2.5 for G, 1.9 for R, 1.0 for H, as computed over the rows of Figure 4.4C. Dissecting the data by social situations in Figure 4.4D, we see the increase in violations in the Behavior Graph rollout scenario occurred mainly in the Cross Path and Down Path social situations. This type of analysis suggests that performing data augmentation for a learned policy and adjusting the training data distribution to include more samples from the under-performing social situation could increase performance. Additionally, dissecting metrics by social situation allows researchers to interrogate controller performance in specific contexts. For instance, delivery robots may need to be especially skilled at navigating down busy pathways in warehouse settings. Note that splitting the data by social situation was not possible in SEAN 1.0 as it did not contain a concept of social situations.

Figure 4.5: Virtual Reality (VR) capabilities incorporated into SEAN 2.0. A user controls an avatar through the VR interface in the Social Environment for Autonomous Navigation (SEAN).

4.6 Virtual Reality Capabilities in SEAN 2.0

We integrated Virtual Reality (VR) capabilities into SEAN 2.0 to enable user-centered studies of social robot navigation in immersive 3D environments. As shown in Figure 4.5, users can control a virtual avatar and look around using a head-mounted display (HMD) and VR controllers, creating a more engaging experience than traditional mouse-and-keyboard simulators.

While gaming remains the most common application of consumer-grade VR devices [305], VR is increasingly used in human-robot interaction (HRI) research to study how humans perceive and respond to robots in socially complex environments [164, 288]. Prior work has used VR to predict gaze behavior [112], improve communication [287], and enable telepresence control [134].

Our VR-enabled SEAN 2.0 allows researchers to simulate crowded environments with virtual pedestrians, offering a safe testbed for early-stage or exploratory algo-

rithms that lack formal safety guarantees, such as learned navigation policies [205, 207, 202]. This mitigates the risk of physical harm while preserving realism.

Evaluating social navigation performance often relies on subjective user feedback, such as perceptions of social appropriateness [89]. VR can enhance these evaluations by providing a more immersive and ecologically valid experience. Our system supports the broader adoption of VR in HRI research and can help address open questions about its methodological impact [294, 158].

4.7 Summary

SEAN 2.0 is a simulation platform designed to support the development and evaluation of socially compliant navigation algorithms in densely populated, dynamic environments. Motivated by the need for robots to adapt to diverse social contexts, SEAN 2.0 introduces a formalization of social situations and a range of features that enable rich, socially-aware interactions. These include customizable sensors, multiple navigation metrics, flexible task specification mechanisms, a suite of logic-based classifiers for recognizing social situations, and diverse pedestrian control strategies.

A key contribution is the introduction of the Behavior Graph, a new mechanism for specifying pedestrian behavior that supports both efficient scenario design and the emergence of varied, realistic social situations via complex pedestrian motion dynamics.

Critically, SEAN 2.0 includes a psychologically grounded model of social context, inspired by work in Social Psychology [13], which is operationalized through a rules-based classifier. This enables the system to recognize and label Social Situations during simulation, facilitating the study and evaluation of socially aware robot behavior.

To support user-centered evaluation, we also developed VR features for SEAN,

adding a virtual reality extension to the platform that allows participants an immersive VR experience in which they can respond to robot navigation behavior from a first-person perspective.

Chapter 5

How Do Robot Experts Measure Social Robot Navigation?*

The ability to characterize and identify Social Situations is critical for the fair evaluation of different robot navigation policies that are designed to operate across a wide range of environments and under varying conditions. Once similar Social Situations can be identified, specific metrics can be applied to compare robot performance across evaluation runs. However, it is essential to ensure that the most relevant metrics are selected for each context and thereby align the user's goals with the measured robot performance. To better understand what metrics matter most, in this chapter I present a study in which we conducted structured interviews with social navigation experts from both industry and academia. Experts were asked to rank the importance of ten commonly used social navigation measures and respond to open-ended questions about evaluating social navigation. A key insight from these interviews is that subjective human feedback plays a critical role in evaluation. Notably, avoiding collisions was the only universally important metric identified, highlighting the

*Parts of this chapter were originally published as Nathan Tsoi, Jessica Romero, 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)* [275].

foundational importance of safety. Beyond that, experts expressed varying priorities that depended on their application domains. Based on these findings, we recommend that social navigation algorithms first prioritize safety. Beyond this, social navigation algorithms should be evaluated using the most relevant metrics, which must be carefully selected by users given their application domain and specific goals.

5.1 Introduction

Mobile robots operate in a wide range of settings. Prior works have studied social navigation in settings that encompass airports [278], labs [271], and museums [107]. The number of pedestrians near the robot can range from a single person or a few people [202] to crowds of people [7]. The task is often A-to-B navigation, from one position to a goal position, but can also include delivery [151, 185], guiding [105, 28], following [94], serving as a receptionist in a building [93] and interacting with groups [266, 303]. 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 [176, 87]. We hypothesized that users of social navigation robots in different application domains are concerned with different aspects of performance when evaluated based on 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 naviga-

tion 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.

5.2 Related Work

Many different evaluation measures have been proposed to evaluate social navigation approaches. Measures can be quantitative or qualitative; the latter typically focuses on human perception of robot behavior. We refer the reader to surveys that discuss these measures in detail [176, 87]. In the broader field of Human-Robot Interaction (HRI), common metrics have been reviewed by Steinfeld et al. [235]. 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 [87, 235] and readily available measures [273] 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. [87], 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 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.

5.3 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.

5.3.1 Hypothesis

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

5.3.2 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.

5.3.3 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.

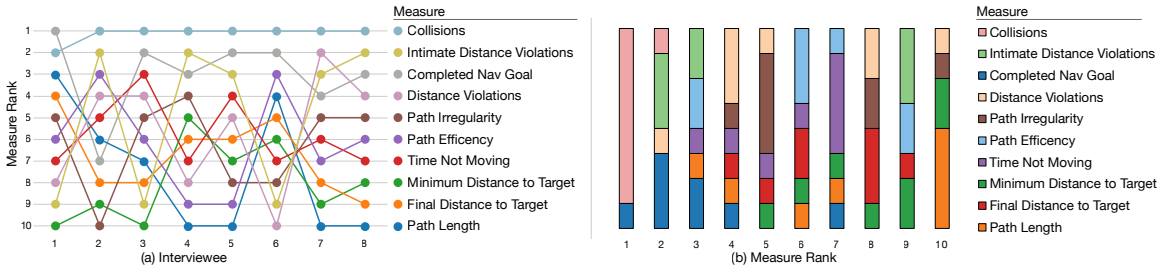


Figure 5.1: Two plots that show the measure ranking results visualized in different ways. (a) Ranking of social navigation evaluation measures by interviewee. Where 1 corresponds to the most important and 10 corresponds to the least important evaluation measure. (b) Interviewees who assigned the same 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.

5.3.4 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 her-

self 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 questions.

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 asked open-ended questions about the success of social navigation. The exact wording of these questions is detailed in Table 5.1.

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

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 5.1: List of questions by category asked to participants during the video interviews.

5.4 Results

We hypothesized that users of social navigation approaches in different application domains are concerned with different aspects of performance when evaluated based on how they prioritized different evaluation measures. We asked participants to rank 10 measures commonly used to evaluate social navigation approaches, shown in Figure 5.1, 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 phenomenon. 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 im-

portant 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 first 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.

5.5 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.

5.6 Summary

Our hypothesis was that users in different application domains of social navigation robots are concerned with different aspects of performance when evaluated based on how they prioritize 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 domains. 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.

5.7 Discussion

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 who have common goals. Finally, open-ended questions showed that interviewees were also interested in subjective metrics that may help evaluate the social aspects of a navigation robot's performance.

Chapter 6

Scalable Data Collection for Social Robot Navigation*

Recognizing that subjective human feedback is a critical component of evaluating social robot navigation, this chapter presents the SEAN Experimental Platform (SEAN-EP). SEAN-EP is a tool for collecting human responses to robot navigation behavior through interactive, online simulations embedded in surveys.

6.1 Introduction

Social robot navigation is a critical task for robotic applications such as service robotics [256, 80, 264], healthcare [44], and education [5, 128]. Thus, social robot navigation has long been studied from a technical and experimental perspective [138, 162, 261, 173, 144]. Yet, there is no agreed-upon protocol for evaluating these systems because: (a) robots often have different capabilities, making comparisons difficult; (b) implementing robust navigation systems is hard, thus baselines do not

*Parts of this chapter were originally published as Nathan Tsoi, Mohamed Hussein, Olivia Fugikawa, J.D. Zhao, 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)* [270].

necessarily represent the state of the art; and (c) there is a lack of standard human-driven evaluation metrics because the context of navigation tasks can significantly alter what matters to users [163, 260]. These issues have hindered advancements and make it difficult to understand the key challenges that the social robot navigation community faces today.

While simulations may not be perfect replicas of the real world, I believe that they provide a viable path towards standardizing the evaluation of social robot navigation systems in Human-Robot Interaction (HRI). Our rationale is twofold. First, simulations have long been leveraged to conduct early tests, stress tests, and verification of robotic systems [192, 198, 236]. They provide controlled environments to systematically study critical application scenarios, and can be used for benchmarking [166, 1, 2, 3]. Second, simulations can be integrated with real-time robotic software, facilitating sim-to-real transfer. For example, interfaces for the Robot Operating System (ROS) have enabled running robot stacks in a variety of simulators [248, 232, 116, 250].

Given prior work in robotics simulations, what is the key challenge that prevents us from fully leveraging simulations for evaluating systems in HRI? The problem is that the evaluation requires human input because the social aspects of robot navigation are subjective in nature. One option for gathering human input is to utilize web-based surveys and crowd workers. However, modern simulations are compute-intensive applications designed for local use in a desktop computing environment. That is, these simulators are *rich-client* applications that provide rich functionality independent of a remote server – in contrast to thin-client applications which are heavily dependent on remote processing, like browser-based web applications. This makes the modern simulators inaccessible to crowd participants who are limited to browser-based web applications.

Making a rich-client application, like a robotics simulation, available on the web

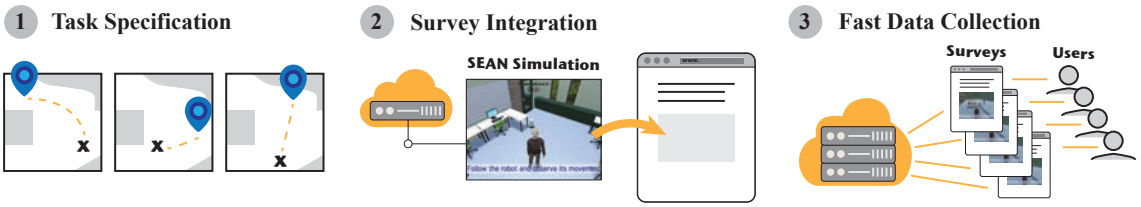


Figure 6.1: With SEAN-EP, researchers can scale HRI experiments in the context of navigation via 3 steps: (1) experimenters specify navigation tasks in the simulator, (2) they integrate interactive simulations based on the tasks with online surveys, and (3) they collect data in parallel from multiple users. See the text for more details.

is a non-trivial task. Perhaps one could think of re-implementing the software under the constraints of a web browser [247] and application-specific or web server-specific modules such as [55]. However, some re-implementations are too complex, time consuming, or even infeasible due to the lack of specific dependencies such as a programming language, physics engine, or rendering engine. Another option could be to use specific solutions that make applications such as word processor programs available in a web browser [55]. Unfortunately, these solutions do not generalize well to robotic simulators that require high performance graphics rendering via specialized hardware and libraries such as OpenGL. These challenges have often restricted human evaluation of social robot navigation via crowd-sourcing to video surveys (e.g., [207, 22]). While videos may lead to comparable results to in-person studies in some cases [295], they are passive media with low interactivity [301].

In this chapter, we propose a method of making rich-client, interactive robotic simulators accessible at scale on the web. As an example implementation, we introduce the SEAN Experimental Platform (SEAN-EP), an open-source system that allows roboticists to gather human feedback for social robot navigation via online simulations (Figure 6.1). Though our implementation uses the Social Environment for Autonomous Navigation (SEAN) [267] as the underlying simulator, any rich-client simulator that runs in Linux could be deployed using our method.

We validated our implementation and its usability through an online study about

social robot navigation. Further, we investigated whether human perceptions of robots differ based on whether they experience human-robot interactions via online simulations or watch them through videos. Interestingly, our results suggest that interactive surveys are less mentally demanding than non-interactive video surveys.

In summary, this chapter makes four main contributions:

1. A novel approach to deploy rich-client robot simulation environments at scale using standard web technologies. This method allows one to quickly gather human feedback in HRI.

2. SEAN-EP, a specific instantiation of the proposed approach based on the Social Environment for Autonomous Navigation. SEAN-EP is open-source and available online.*

3. Validation of our example implementation (SEAN-EP) through an online study about social robot navigation.

4. An experimental comparison of interactive simulations and videos for studying human perception of robot navigation.

6.2 Related Work

6.2.1 Robotics Simulation Environments

Progress has been made on developing photorealistic simulations which bridge the gap between virtual worlds and reality [297, 4]. With their high visual fidelity and responsiveness, game engines such as Unity and Unreal Engine have proved indispensable to robotic simulation of flying and mobile robots [232, 116, 142, 98]. Several of these simulation environments integrate with ROS to achieve realistic robot control and transfer results to the real world.

*https://github.com/yale-sean/social_sim_web

Within social robot navigation, crowd simulation and modeling of pedestrian behavior have improved as well [246, 72, 15]. However, there has been less work on combining crowd models with robotics simulation for robot navigation in human environments. One exception is the Social Environment for Autonomous Navigation (SEAN) [267], which we leverage in our work. SEAN builds on Unity and integrates with ROS, making it a good option in terms of photorealism and future sim-to-real transfer.

Modern robotics simulations are rich-client applications meant to run on a desktop computer or a powerful gaming laptop. While some simulators may utilize frameworks that provide the option to compile to WebGL for deployment on the web, like Unity, there are many challenges and limitations to this approach. This includes lack of direct access to IP sockets from WebGL due to security implications, limitations in rendering and illumination, lack of threading in JavaScript, and limited access to hardware [21]. While specific limitations can be addressed via engineering workarounds on a case-by-case basis, this approach lacks the flexibility of our method. Our method works with any kind of rich-client simulation that runs in Linux, even if it requires interacting with other software such as ROS components.

6.2.2 Leveraging the Web in HRI

While traditional HRI experiments are conducted in person, prior work has explored faster mechanisms that leverage the accessibility of the Web. For example, [61] explored using a two-player online game to build a data corpus for HRI research. Also, [220, 188, 300] developed web-compatible simulations based on Unity and WebGL, although these systems have the limitations discussed in Section 6.2.1.

Especially during early HRI system development, it has become common practice to gather human feedback via online video surveys [311, 249, 76, 207, 22]. Research has indicated that there is a moderate to high level of agreement for subjects' pref-

erences between live and video HRI trials [295]. Our work contributes to a better understanding of experimental methods by comparing human feedback from a video survey with feedback from an interactive survey.

Our approach is inspired by internet frameworks that provide methods to remotely interact with robotic systems and simulators. For example, the Robot Management System (RMS) [258] and Robot Web Tools [259] provide software for building web-based HRI interfaces and demonstrate their approach in a variety of tasks, including remote robot operation. These tools allow web clients to interface with ROS and the Gazebo simulator by transforming ROS-specific data streams into formats compatible with a web browser. While ROS applications can be made accessible on the web via RMS, our approach is relevant to all kinds of rich-client simulators that run on Linux, not just Gazebo.

Lastly, RoboTurk [168] allows for rapid crowdsourcing of high-quality demonstrations for robot learning in the context of manipulation. While our method could be used in the future to gather data for learning navigation policies through teleoperation, in this chapter we explore giving users control of a human avatar in the simulation. This methodology aims to bring their online, virtual experience closer to real-world human-robot interactions.

6.3 Method

We propose a general approach to deploy the graphical user interface (GUI) of rich-client robotic simulators on the web to facilitate and scale HRI experiments. Our approach builds on standard tools for graphical desktop sharing in Linux. It does not require the adaptation of rich-client simulators to other technologies, such as WebGL. With our approach, researchers can create interactive HRI surveys. These are surveys that, in some parts, include simulations in which the participant interacts with a

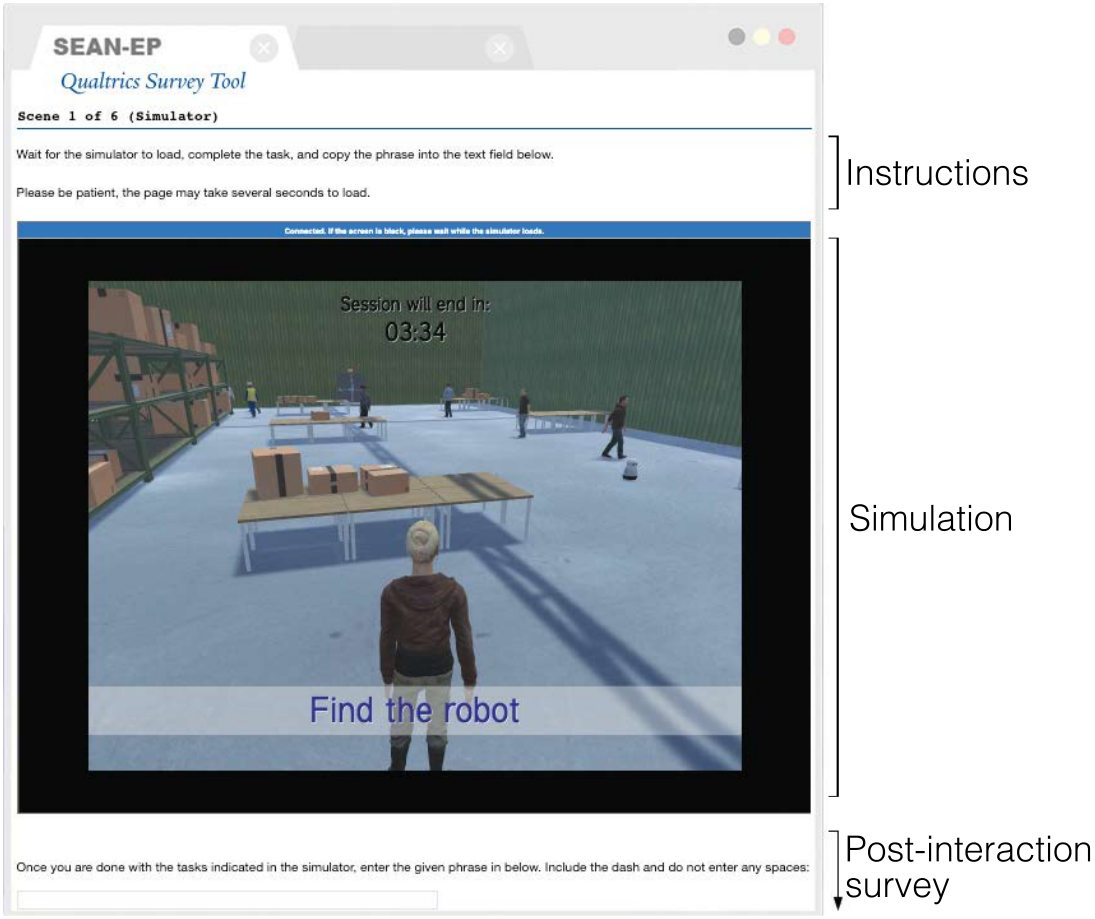


Figure 6.2: Screenshot of Qualtrics survey with embedded SEAN simulation. Best viewed in digital form.

robot. Figure 6.2 shows an example online survey used to study social robot navigation. This simulation was embedded in the survey using a particular instantiation of our approach, as later described in Section 6.4. After the simulations end, the online survey can query the participant for explicit feedback about his or her experience interacting in the virtual world.

Importantly, our method addresses parallelization challenges inherent to online studies typically run via crowdsourcing platforms. While simulation systems for in-person HRI studies are usually designed for one participant at a time, online surveys must cope with a potentially large number of people who participate in the study simultaneously. Our method provides a mechanism to scale simulations designed for a single user to many users in parallel. This is possible without changes to the underlying system.

The next sections describe in detail our method to make rich-client robotics simulations accessible on the web. We evaluate an implementation of this approach in Section 6.5.

6.3.1 Making Interactive Simulations Accessible on the Web

We propose to make rich-client simulations available in a standard web browser by running them on a remote server, and using a Virtual Network Computing (VNC) server to share the GUI of the simulator with a remote user.

While remote users typically connect to a VNC server via a desktop VNC client running on their machine, we use a browser-based VNC client running on the host server to allow browser-based access to the simulator GUI. The browser-based VNC client renders the GUI on a web page, through which our system can accept user input for the simulator, e.g., keyboard commands.

One important consideration when exposing the GUI of a simulator on the web, as described before, is that users are unauthenticated and untrusted. Thus, it is

important that the GUI of the simulator does not provide mechanisms to launch other processes on the remote server.

Because VNC connections are designed to be used by a single user, we propose the use of a web-based process orchestration tool to deploy and manage a large number of concurrent simulation environments. We call this tool the "Process Manager" because it controls the execution of processes associated with each simultaneous user. This tool is further described in the next section, where we explain in detail how to scale data collection on a single host.

6.3.2 Scaling on a Single Host

To scale human feedback collection, we can run multiple instances of the simulator on the remote server and provide individual remote users access to one of them. Figure 6.3a illustrates how we achieve this goal using a reverse proxy server and a Process Manager. The Process Manager is responsible for managing *user sessions*, which include an instance of the simulator (including its GUI), all other components necessary for the simulator to run, and a VNC server and browser-based client for the given user.

The reverse proxy routes web requests that are received via Hypertext Transfer Protocol Secure (HTTPS) to specific web servers on the host machine based on the URL path of the request. The target web server may be the Process Manager, which is in charge of initiating, maintaining, and terminating user sessions, or an existing web-based VNC client within the user's session.

Requests for simulations should have a specific URL path that includes a parameter for a unique user identifier, e.g., a Mechanical Turk ID. Additionally, they should include any other parameters needed to instantiate the simulation for the user. For instance, in the example implementation described in Section 6.4, the requests include start and goal poses for a user's avatar and a robot in the simulation.

When the Process Manager receives a request from the reverse proxy, it evaluates if it is a new request given the URL parameters. If that is the case, then the Process Manager launches the main components that make up an interactive simulation session and quickly redirects the request to the page of the web-based VNC client that corresponds to the user. Because the VNC web page is served on the same host, the reverse proxy gets the request that results from the URL redirection and appropriately routes it so that the user's web client can display the simulation's GUI. If an existing user requests a running simulation, the Process Manager simply redirects the request to the corresponding VNC URL.

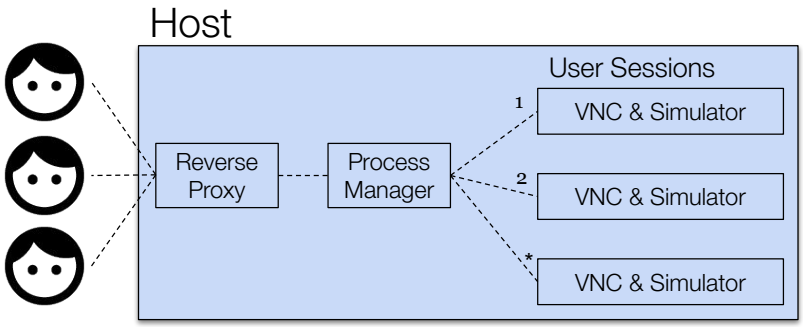
The Process Manager is also in charge of managing the maximum session duration. Sessions are allowed to run for a configurable amount of time before being automatically shut down. Once a session is closed, the resources can be re-allocated to new sessions for other users.

Handling simulation requests as described above is beneficial in 3 key ways: (1) it is easy to integrate simulations with online surveys because a single web address (with parameters) is used to handle all requests; (2) the entire connection between a remote user and the host machine is encrypted over an HTTPS connection; and (3) because the reverse proxy is routing requests rather than having users connect directly to each VNC instance, there is no need to expose many non-standard internet ports on the host.

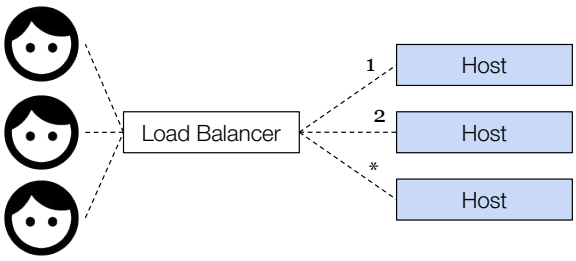
Crucially, this method allows a single host to handle the multiple concurrent requests required of an online study, where many users need access to interactive sessions to complete surveys.

6.3.3 Scaling Across Many Machines

Scaling of simulator sessions on a single host machine is limited by the hardware resources on the host. The ability to scale users' sessions across many machines, or



(a) Our method running on a single machine. The host machine accepts inbound connections via the NGINX reverse proxy.



(b) Our method running on multiple machines. Each host has all the components shown in Fig.6.3a.

Figure 6.3: Proposed methods to render the GUI of rich-client simulations on the web and scale HRI data collection.

scale "horizontally," removes this limitation.

Horizontal scaling can be achieved by adding a Load Balancer to our proposed system. The Load Balancer receives user requests and then acts as a "traffic cop" to evenly distribute the requests to the available host machines, as illustrated in Figure 6.3. This routing process must be "session aware" to associate users to the same host machine if they perform multiple requests.

With the guarantee that a single host will receive all requests for a unique user, the system state does not need to be shared across hosts, but can be managed in the same way as described in Section 6.3.2. Moreover, new hosts can be added dynamically to handle more concurrent sessions by simply notifying the Load Balancer.

6.4 SEAN-EP System

We created the SEAN Experimental Platform (SEAN-EP) in order to validate our method (Section 6.3) in the context of social robot navigation. Our goal was to test our method's feasibility in a realistic usage scenario and, through this effort, verify the key tenets of scalability and usability.

SEAN-EP uses SEAN [267] as the core simulator. SEAN provides photorealistic virtual worlds, crowd simulations for social robot navigation, and integration with ROS for robot control. We modified SEAN to use the Microsoft Rocketbox avatars library [95] for this work because these avatars are higher-quality than those used in [267].

6.4.1 System Implementation

We implemented our method as an open-source system and deployed it to virtual hosts with dedicated NVIDIA T4 GPUs using Amazon Web Services (AWS). We aimed to maximize the performance of our system and fully utilize the hardware resources to deliver a user-friendly and visually appealing simulated interaction that is free of glitches or lag. To this end, we chose TurboVNC as the VNC server, which accelerates data transfer by compressing images via libjpeg-turbo. We made TurboVNC available on the web using noVNC with websockify.[†] Notably, TurboVNC also supports VirtualGL for hardware-accelerated 3D graphics.

We used the open-source NGINX server as a reverse proxy and implemented the Process Manager using the popular Flask web framework. We developed a custom configuration for NGINX to properly route requests as specified by the Process Manager. Also, we configured an AWS Application Load Balancer with sticky-sessions to make it session-aware.

The Process Manager facilitates communication between the SEAN GUI and ROS.

[†]https://github.com/novnc/websockify

Because each session requires its own instance of ROS, we encapsulate ROS processes in a Docker container and expose a single network port for the SEAN GUI to communicate with its ROS instance.

We used ROS bag files as the main logging mechanism for human-robot interactions enabled by SEAN-EP.

6.4.2 Navigation Tasks

To evaluate our idea in practice and conduct an experiment about perceptions of social robot navigation (Section 6.5), we designed three tasks for users to complete in SEAN simulations. First, they had to find the robot. Second, they had to follow the robot and observe its movements, which required them to stay in proximity to the robot and observe its interactions with other people. Third, they had to navigate to a nearby location in the environment identified with a visual landmark. This last task incentivized them to navigate around the robot to reach their destination. Overall, these tasks motivated users to both interact with the robot in the virtual world and behave in naturalistic ways.

6.4.3 User Interface

We created a new user interface in SEAN to let a user control an avatar in the simulation and make the virtual experience similar to real human-robot interactions. We had two key requirements when designing the user interface: it had to be accessible to a wide range of users, and it needed to be simple enough to be explained in a short introductory tutorial. Given these requirements, we chose to implement an interface that is similar to a third-person video game, albeit with simplified controls. The main camera of the simulation follows the user's avatar as it moves. Users can press the up and down arrow keys to raise and lower the camera, changing the field of view of the environment as needed. In addition, they can use the keyboard commands W, A, S,

and D to move their character forward, left, backward, and right, respectively. These keyboard commands were captured in the users' browsers and seamlessly passed to the SEAN simulation using noVNC.

6.4.4 Data Collection via Online Survey

With our system, SEAN simulations can be integrated with standard online survey platforms via HTML iframe elements. The surveys can collect any additional data from users, such as demographic data or answers to questions about their experience in the simulations. An example is provided in the evaluation presented in Section 6.5, for which we integrated SEAN simulations with a Qualtrics survey (Figure 6.2).

6.4.5 Performance

SEAN-EP provides users web access to interactive SEAN simulations with a small amount of load time. When a user requests a new SEAN simulation session, the Process Manager starts a complete ROS environment, the Unity-based simulator GUI, and a VNC server and client. Despite all these many programs, the start-up time for a user session is 18.9s on average. Transferring the simulator's GUI to the user's browser through noVNC takes on the order of milliseconds with a low Internet connection speed in the U.S. (e.g., on the order of 10 Mbps). This means that the total wait time for users to access a SEAN simulation with SEAN-EP is significantly faster than compiling SEAN Unity worlds to WebGL. The reason is that the worlds are complex, resulting in simulations that are over 1.5GB in size after the compilation. With a global average fixed broadband download speed of 77 Mbps, transferring a single WebGL environment to the browser would take over 2.5 minutes.[‡]

[‡]Note that the complete SEAN simulation cannot be exported to WebGL due to ROS dependencies. Thus, we only report the expected time that it would take to load the Unity world after converting to WebGL.

6.5 SEAN-EP Evaluation

We used a Qualtrics online survey to validate the potential of our method to gather human feedback for social robot navigation. The survey included 6 interactive simulations, embedded through HTML tags, through which users could interact with a Kuri robot. The next sections detail our experimental protocol and results, with a special focus on user feedback obtained through the survey. Section 6.6 later compares using this type of interactive survey versus a video survey to gather human feedback about robot navigation. The protocols for these studies were approved by our local Institutional Review Board.

6.5.1 Method

The Qualtrics survey was designed to gather feedback about robot navigation in two simulated indoor environments. One environment was a warehouse that included 15 virtual humans, a Kuri robot, and the user's avatar (Figure 6.2). The other environment was a computer laboratory, which included one virtual human besides the robot and the user's avatar (Figure 6.1). The goal of the user in the simulations was to first find the robot, then follow it for 30 seconds, and finally navigate to a destination identified by a visual landmark.

Experimental Protocol. The survey began with a demographics section. Then, the participants were asked to behave politely in the simulator and were introduced to the task with a short simulation in the lab environment. This simulation served as a tutorial to explain the commands that the participants could use to move their avatar, identify Kuri, and practice navigation tasks. After the tutorial, the participants experienced 6 simulations in randomized order: 3 in the warehouse environment and 3 in the laboratory. For each simulation, there were specific start and goal locations for all agents. In particular,

the navigation goals of the robot and the human avatar were opposite to each other, so that they would easily encounter one another at some point in the virtual world. After each interactive simulation, the participants were asked a few questions about their experience, including whether they were able to identify the robot and whether it moved in the environment. At the end of the survey, the participants were asked about their overall experience.

Robot Control. The Kuri virtual robot was modeled after the real platform manufactured by Mayfield Robotics. It had a differential drive base and used a simulated 2D LIDAR and odometry information from Unity to localize against a known map. All path planning and execution were completed by the ROS Navigation Stack, which used a global and local costmap for object avoidance and social navigation around virtual humans [162]. We opted to use the Navigation Stack because it is widely used by many robots, including TurtleBot platforms, PR2, and the Clearpath Husky. Also, it is used as a classical baseline and ground truth by more modern learning-based approaches [207, 56].

Participants. We recruited 62 participants through Prolific, a crowdsourcing platform, for this evaluation. Participation was limited to individuals 18 years or older, fluent in English with normal-to-corrected vision. The participants had an average age of 32 years old and 29 were female. In general, the participants were familiar with video games (M=6.1, $\sigma = 1.1$), but not as familiar with robots (M=4.0, $\sigma = 1.4$) based on answers on a 7-point responding format (1 being the least familiar, 7 being the most familiar). They were paid $4.00 USD for completing the survey.

System Architecture. Because our interest was testing the proposed system, we ran participants under both strategies to scale simulations (Section 6.3). Half of the participants experienced simulations running on a single host machine and

thus were run in small batches to avoid overloading the host. The other half interacted with simulations distributed across four machines. Virtual machines were AWS g4 EC2 instances with 32 cores, 124GB of RAM, and 15.84GB of GPU memory. Given the requirements of our simulation environment, each machine was capable of running up to 30 interactive simulation sessions in parallel, based on its GPU memory and the size of our SEAN environments. In the case of scaling across many machines, we limited the number of sessions per host below the resource-constrained maximum to 10. We also added enough hosts behind the Load Balancer to accommodate the maximum number of total simultaneous participants in our study.

6.5.2 Results

We were able to successfully gather data by using a single host machine as well as multiple ones. With a single host, we ran on average 5 participants at a time, collecting all 31 responses in about 8 hours. With the multiple host approach, we ran all 31 participants at the same time and collected all 31 responses in 1.5 hours. As a reference, the average participant took about 31 minutes to complete the survey.

While the multiple host approach effectively reduced the time that it took to collect data through the surveys by 81%, it required more management overhead. This included managing instances in the pool and moving collected data from the machines to a shared store for analysis.

Task Completion. While not all participants followed the instructions by the book, a large majority tried and were able to complete the given tasks, validating that our system worked as intended. In only 23 of 372 interactive trials (6.18%), the participant's simulation session timed out before they reached their goal destination. Considering all 372 interactive sessions, there were only 8 interactive sessions (2.15%) in which participants did not move from the starting position.

Further inspection of the data revealed that the only two sessions in which the participants failed to find the robot in the simulation corresponded to timed out sessions in which their avatar moved. These simulations were in the warehouse environment, suggesting that they tried to find the robot, but the large space made it difficult for them to identify it. Because the study sessions were conducted in parallel, at scale, they help confirm the scalability of our method.

Navigation Behavior. Using ROS logs from SEAN, we checked how often the human's avatar and the robot were close to each other based on Hall's proxemic zones [101]. We set a threshold for intimate space of 0.45 meters, and found that in 195 sessions (52.4%) the robot came within this distance from the participant's avatar. In terms of personal space, in 316 sessions (84.9%) the robot came within 1.2 meters from the avatar. A benefit of simulations is that we can easily analyze proxemic behavior as shown by these results.

User Experience. At the end of the survey, the participants reported that the survey tasks required low mental demand (M=2.26, SE=0.19) and low physical demand (M=1.61, SE=0.15) on a 7-point responding format where 1 indicated the lowest demand. They did not have to work hard to accomplish the tasks (M=2.32, SE=0.18).

Some participants provided positive feedback for our system through open-ended questions. For example, one person said that *"the game was very well made and the controls are what I'm used to with my own gaming."* Another said they *"found the instructions were easy to follow."* When asked if the virtual world was confusing, participants strongly indicated it was not confusing (M=2.03, SE=0.18). However, a few participants reported confusing elements of the survey. Eight people believed the robot's motion was awkward. Additionally, 7 people thought that the control of their human avatar was unintuitive.

As a reference, the participants had an average internet speed of 105.56 Mbps (SE=11.09).

Overall, these results validate the feasibility of our method to enable online, interactive HRI studies.

6.6 Interactive Simulation vs. Video Feedback

We compared the interactive human feedback obtained using SEAN-EP with feedback obtained through a video survey, which is a typical approach to online HRI studies as discussed in Section 6.2. To this end, we recruited 62 more participants through Prolific. These participants provided feedback about the robot based on videos of the simulations that happened as part of our prior study (Section 6.5).

6.6.1 Method

Experimental Protocol. We expanded our data from Section 6.5 with data collected through a Qualtrics video survey. In general, the video survey followed the same format as the prior one. However, instead of having participants interact with the robot in a virtual world, each participant viewed the 6 video recordings of the simulations experienced by a participant from our validation study. After watching each video, they were asked about the observed robot.

Hypotheses. The data from Section 6.5 (Interactive condition) and the video survey (Video condition) were analyzed together to investigate two hypotheses:

H1. *The perception of the robot would differ between the conditions.* To test this hypothesis, we gathered ratings for the Competence and Discomfort factors of the Robotic Social Attributes Scale (ROSAS) [48] (Cronbach's α was 0.938 and 0.746, respectively). We also gathered participants' opin-

ions on whether the robot navigated according to social norms after each simulation session or corresponding video.

H2. *The perceived workload for the survey in the Interactive condition would be lower than in the Video condition.* We measured perceived mental and physical demand along with effort at the end of the surveys based on responses to the following questions from the NASA Task Load Index [102]: "How mentally demanding were the tasks?", "How physically demanding were the tasks?", and "How hard did you have to work to accomplish what you had to do?". Responses were collected on a 7-point responding format (1 being lowest, 7 being highest).

Participants. A total of 124 participants were considered for this experiment (62 from Section 6.5 plus 62 new participants). Their average age was 34 years old and 55 of them were female. We limited participation in the same way as Section 6.5. Participants were paid $4.00 USD for completing the survey.

6.6.2 Results

Human Perception of the Robot. There were 2 simulation sessions out of 372 in the Interactive condition in which the participants failed to identify the robot, and 53 sessions in which they said it was not moving. Meanwhile, there were 4 sessions out of 372 in the Video condition in which the participants failed to identify the robot, and 51 sessions in which they said that it was not moving. We excluded these sessions from further analysis about perceptions of the robot.

For the 288 pairs of sessions in which participants both saw the robot moving in the simulation and the video condition, we conducted a Wilcoxon signed-rank test for the paired data to check if the Competence and Discomfort factors from ROSAS (in 7-point responding format, 1 being lowest and 7 being highest) dif-

fered by condition. The test resulted in no significant differences for Discomfort. The median Discomfort was 2.33 for the Interactive condition and 2.17 for the Video condition. However, the Wilcoxon test revealed significant differences for robot Competence ($p < 0.0001$). The median Competence value was 3 for the Interactive condition and 3.83 for the Video condition. Lastly, an additional paired Wilcoxon signed-rank test indicated significant differences by Condition in terms of whether the robot navigated according to social norms. The mean rating was 3 over 7 points for the Interactive condition and 4 over 7 for the Video condition. These different results provided evidence in support of our first hypothesis (H1).

Workload. We conducted an additional paired Wilcoxon signed-rank test to evaluate potential differences in perceived workload across Conditions. The test indicated significant differences for perceived mental demand ($p < 0.01$). The median rating for mental demand in the Interactive condition was 2 points out of 7, while the median for the Video condition was 3. No significant differences were found for the ratings about physical demand. The median rating was 1 – the lowest possible – for both conditions. Lastly, we found significant differences for how hard the participants had to work to complete the surveys ($p = 0.023$). The median rating for the Interactive condition was 2 out of 7 points, and the median for the Video condition was 3. These results partially support H2.

We suspect that the different results across conditions were due to the interactive nature of the simulation, which provided better opportunities for the participants to evaluate the responsiveness of the robot to their actions and to other virtual humans than the video. However, further tests are needed to validate this assumption. For example, future tests could consider human perception of the robot and the perceived workload in both the real world and the simulated replica. Another aspect to consider is the user perspective. While we used a third-person perspective in SEAN-EP,

studying interactions perceived from a first-person perspective could better translate to the real world. An additional consideration for future work is the display type used by the participants.

6.7 Discussion

Our approach to make interactive simulations available on the web was effective in general. It allowed users to control their virtual avatars in rich-client simulations and quickly gather data to study social robot navigation. In the future, we would like to use SEAN-EP to also allow users to control the robot, so that we can collect example behaviors for social robot navigation. We also wish to explore other types of common navigation scenarios [215, 84, 173], e.g. walking alongside a robot or passing in narrow spaces.

We observed that most survey participants tried to complete the simulated tasks in a polite and naturalistic manner as directed. There were several people however, who explored undesired actions for their avatars. About 18% of the participants pushed the robot in the simulation, and about 12% collided with a human based on annotations from our video survey. In the future, it is important to explore incentives for participants to reduce these undesired behaviors.

Lastly, we evaluated our proposed approach using a single robotics simulator. Given the flexibility of our method, we would like to see it being used to make other rich-client simulators for Linux easily available on the web. This could facilitate human feedback collection in other HRI domains.

6.8 Summary

This chapter introduced a flexible method to enable crowd-sourcing of human feedback using interactive rich-client simulators deployed on the web. We then demon-

strated a particular instantiation of this approach, called SEAN-EP, in the context of social robot navigation.

We tested SEAN-EP with an online survey, which validated its ability to serve simulations to many users. Furthermore, we compared the results of evaluating robot navigation through interactive simulations using our method against evaluations based on video surveys. Our proposed interactive methodology resulted in different perception of the robot and lower mental demand for participants.

Chapter 7

Methodologies for Collecting Human Feedback in Human-Robot Interaction Research*

Building on the SEAN-EP platform introduced in Chapter 6, we investigate how different methodologies for collecting human feedback affect perceptions of a robot during navigation tasks. While SEAN-EP enables efficient collection of subjective feedback via interactive simulations, it raises important questions about how perceptions gathered via different methodologies compare to those from real-world encounters. To address this, we compare the gold standard of in-person studies with two online approaches, including non-interactive video-based surveys and interactive, online surveys that utilize SEAN-EP. We conducted a 2x2 between-subjects study (N=160) examining the effects of both the interaction environment (Real vs. Simulated) and the level of participant interactivity (Interactive vs. Video) on perceptions of a robot's competence, discomfort, social presentation, and social information pro-

*Parts of this chapter were originally published as Nathan Tsoi, Rachel Sterneck, Xuan Zhao, Marynel Vázquez. (2024). Influence of Simulation and Interactivity on Human Perceptions of a Robot During Navigation Tasks. In *ACM Transactions on Human-Robot Interaction (THRI)* [276].

cessing. Our results revealed a significant difference in the perceptions of the robot between the real environment and the simulated environment, as well as between passive observation and active engagement. Notably, simulated interactions and their corresponding videos elicited higher reported workload than real-world conditions. These findings suggest that results from video-based and simulation-based methodologies may not always translate to real-world human-robot interactions. In order to allow practitioners to leverage learnings from this study and future researchers to expand our knowledge in this area, we provide guidelines for weighing the tradeoffs between different methodologies.

7.1 Introduction

Different methodologies have been proposed to investigate human perceptions of robots in Human-Robot Interaction. Generally, the gold standard is to collect human perceptions through real-world, in-person studies [25]. However, in-person studies may carry with them administrative overhead, e.g., the recruiting of participants (perhaps through flyers, social media or word-of-mouth) and scheduling. Moreover, each participant must travel in order to interact with a researcher in a set physical space. In practice, the need for in-person interaction and the associated administrative overhead could negatively impact the number of participants in an in-person study. Inadvertently, this could limit the sample size and statistical power a study may achieve [110].

An alternative to in-person studies is to record interactions between a human and a robot in videos and then gather human perceptions of the robot using a web survey that includes the recordings. Because of the online nature of the survey, participants can be recruited via online crowdsourcing platforms [123], allowing researchers to scale data collection and accelerate the pace of research. However, video studies are not

a) Real-Interactive b) Sim-Interactive c) Real-Video d) Sim-Video

Figure 7.1: Experimental conditions of our 2×2 between-subject study. Our independent variables were the interaction environment (Real vs. Simulated environment) and the level of interactivity of the research methodology (Interactive participation vs. Video observation).

without limitations. First, video interactions between a human and a robot can lack diversity compared to in-person studies due to the limited number of scenarios used to create videos. Second, participants who observe interactions through the recordings are one step removed from the human-robot interaction. In this case, participants providing the survey responses are not interacting with the robot but, instead, they passively view the robot interacting with another person. Information flow between the robot in the video and the person providing the label is unidirectional, as opposed to bidirectional, which characterizes interactive encounters with technology [230, 39].

Recently, simulations of human-robot interactions have been used instead of in-person or video-based studies in HRI [292, 231, 270]. Modern web infrastructure allows researchers to deploy simulations within online surveys so that online study participants can virtually interact with a robot in a simulator within their web browser and then provide their perceptions of social robots [270]. Due to the virtual nature of this process, simulations have the potential to improve the efficiency and scalability of data collection in HRI while offering a higher level of interactivity than video-based studies. Prior studies have explored how human perceptions of social navigation robots may differ between some methodologies, such as between videos and simulations [270]. Other studies have explored the potential benefits of in-person vs. virtual interactions [19]. Yet, open questions remain on how human perceptions of a mobile robot for social navigation might differ between such methodologies.

We conducted a study that utilized two navigation tasks to investigate human

perceptions of a mobile robot along 4 dimensions (competence, discomfort, social presentation, and social information processing). As shown in Figure 7.1, the study considered two independent variables. One variable concerned the level of interactivity of the research methodology (Interactive participation vs. Video observation). The second variable was the interaction environment (Real vs. Simulated environment), because simulations used in HRI do not always fully mimic the visual appearance of the real world.

Our results suggest that there are subtle tradeoffs that must be considered when choosing the methodology with which one conducts a study. In particular, our results revealed that interaction environment and interactivity can influence human perceptions of robots in HRI studies. Moreover, the task can also influence perceptions of a robot's performance. While simulations and video studies conducted online are pragmatic for HRI research, our results suggest that user perceptions of robots gathered with these methodologies may not always translate to perceptions from real-world human-robot interactions. In order to allow practitioners to leverage learnings from this study and future researchers to expand our knowledge in this area, we provide guidelines for weighing the tradeoffs between different methodologies in Section 7.6.

7.2 Related Work

This section discusses related work in regards to the types of research methodologies considered in our study. First, we discuss video-based evaluations and simulation in Human-Robot Interaction. Then, we discuss related work on robot embodiment and physical presence, which are important aspects of in-person studies.

7.2.1 Video-Based Evaluation in HRI

Video studies have often been used in HRI to collect data on human perceptions of robots [249, 114], measure human understandability of robot behavior [217, 76], and gather preferences over robot behavior [312, 147]. Videos have also been used to portray recordings of human-robot interactions in a way that seems responsive to human actions [190] and for early robot prototyping [109].

Video recordings of human-robot interactions allow participants to provide feedback regarding their perception of a robot without directly interacting with it. Collecting feedback without in-person interaction is useful when it is infeasible to have a participant interact with the robot due to safety concerns [296] or when there are restrictions imposed by infectious disease outbreaks [83], which can limit access to research materials and robots.

While in-person studies require experimenters to find local participants (e.g., using flyers or word-of-mouth), online video studies can leverage crowdsourcing platforms (such as Prolific or Amazon Mechanical Turk) to reduce recruitment bottlenecks. Furthermore, crowdsourcing can enlarge the participant pool beyond a researcher's immediate geographic location, allowing for cross-cultural studies (e.g., [69, 124, 167]). Finally, once a study is posted online, crowdsourcing also allows the scaling of HRI research by enabling many participants to view videos of interactions and provide their feedback in parallel. However, because it is impossible to fully control the environment in which the video-based study is administered in these cases, there could be biases in the data collection. For example, bias could be introduced due to the screen size used by participants [281]. Nevertheless, because crowdsourcing has gained significant popularity in HRI (e.g., [249, 76, 258, 208, 132, 146, 23, 270]), we also used it in our study about human perceptions of a mobile robot.

7.2.2 Simulation in HRI

In HRI studies, simulations have been used to investigate interactions between participants and robots who engage in a two-way flow of information, which is not present in videos. Early HRI simulators focused on providing graphical user interfaces for robot development and testing. For example, USARSim supported human-robot interaction research in the context of robot teleoperation [155]. Chernova et al. created an online multiplayer game that simulated human-robot interactions for learning interactive robot behavior [61]. Other robotics simulators allowed users to teleoperate human avatars to enable virtual interactions with robots. For instance, the Modular OpenRobots Simulation Engine (MORSE) [79] was integrated with human avatars to allow for virtual experimentation [153]. Also, the Social Environment for Autonomous Navigation 2.0 (SEAN 2.0) [274] integrated the Unity game engine with the Robot Operating System (ROS) to make it possible to train and evaluate social robot navigation policies.

A common limitation of simulation is the lack of visual realism. Rich-client simulations such as MORSE and SEAN 2.0 have partly addressed this limitation, but they typically require a powerful computer with a dedicated Graphics Processing Unit (GPU) to render the virtual world. Web technologies, such as SEAN-EP [270], have been used to increase accessibility to rich-client simulations by allowing a participant to interact with a robot in a simulated environment using a standard web browser. We used SEAN-EP in our study so that participants did not need to install simulation software locally or have a dedicated GPU.

One might naturally assume that more visual realism, via higher-fidelity simulations, is always better than less visual realism. Surprisingly, Truong et al. [265] found that lower fidelity simulations resulted in better sim-to-real transfer of robot navigation behavior. This result inspired us to compare human perceptions of a robot where visual realism can differ based on the interaction environment in which humans

observe human-robot interactions. In our work, these observations were obtained in fully realistic environments (showing real-world interactions in a lab), or they were obtained in a simulation of the lab environment.

Close to our work, Tsoi et al. [270] examined differences in human perceptions of a Kuri robot in two setups: participants either interacted with the robot in SEAN [268], or they observed videos of human-robot interactions in the simulation. They found that, for navigation tasks, a robot viewed in a video was perceived as more competent than one experienced interactively in SEAN. Additionally, participants in the interactive simulation condition reported less mental demand than participants in the video condition. However, no comparison was made with respect to real-world interactions, as in our study.

7.2.3 Physical Robot Embodiment and Presence

One important difference between in-person studies and both video and simulation methods is robot embodiment and presence. These concepts are related but capture different aspects of the interaction [182]. Robot embodiment describes the morphology and visual characteristics of a robot, which can differ between the real world and virtual environments. Type of presence describes where a robot is located, and thereby can influence the medium over which the same robot is experienced (typically in-person, via teleconference, or in a one-way video). There has been much interest in how perceptions of robots are influenced by robot embodiment and presence, but results are inconsistent.

Robot embodiment can influence human perceptions of a robot and human-robot interactions [285, 85, 157, 67, 73, 291, 241]. Robot embodiment is not a binary concept, but exists on a spectrum [85] ranging from disembodied agents which communicate only over text or speech [67, 291], to agents simulated on a screen using a 2-dimensional interface or avatar [73], to agents modeled in a 3-dimensional simula-

tion [268, 155, 274], to agents that exist with a physical presence in the real world. For example, Strait et al. [241] studied the effects of direct versus indirect speech on humans for an advice-giving robot where relevant factors in the study included robot appearance and robot presence. In another study, Wainer et al. [285] compared human perceptions of a co-located physical robot, a remotely located (telepresent) robot, and a simulated robot that explained and supervised a Towers of Hanoi puzzle. The study results suggested that physically embodied co-located interactions are more enjoyable than interactions with remote-located and simulated robots.

Research suggests that human behavior and human perception of robots can be influenced by robots' presence, although results vary in the literature. For example, Jung and Lee [125] and Lee et al. [150] found that the physical presence of a robot can influence its perceived social presence; however, Thellman et al. [253] found that the perceived social presence of a robot was not influenced significantly by its physical or virtual presence [253]. Other examples are found in Bainbridge et al. [19] and Salomons et al. [227], who compared physically present robots with a live video stream of robots on a book-moving task and an exercise task, respectively. These studies found that people were more likely to fulfill an unusual request by the robot, afforded greater personal space to it, and made fewer exercise mistakes when it was physically present. But in social robot navigation, Woods et al. [296] found that perceptions of a robot approaching people were consistent between video and real-world settings. Our study further expands this line of work on the effects of presence on human perceptions of robots.

7.3 Method

Prior work on human perceptions of robots in video, simulation, and in-person studies has been largely fragmented by the research methodologies. To more comprehensively

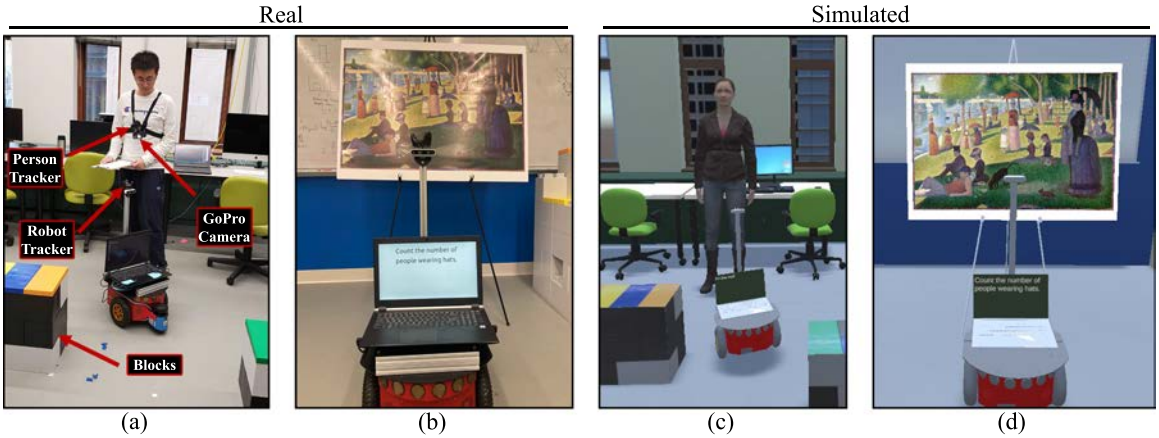


Figure 7.2: Photos of the Real (a and b) and Simulated (c and d) environments. The Interactivity level manipulated how the participant interacted with each of the environments. A participant in the Real-Interactive condition (a) wore a chest harness with trackers for localization and a GoPro camera while interacting with the robot in the real world. A participant in the Sim-Interactive condition (c) used keyboard controls to control an avatar through the virtual lab. Participants in the Video conditions watched video recordings of the interactive participants. During the art task, the robot guided a participant to a poster and communicated with the participant using text on the real (b) or simulated (d) laptop screen.

understand how human perceptions vary between these methodologies, we conducted a $2x2$ between-subjects study with a mobile robot in a laboratory setting. The two independent factors of our study were: *Interaction Environment* (Real vs. Simulated environment), and the level of *Interactivity* of the research methodology (Interactive participation vs. Video observation). Photos of all experimental conditions are shown in Figure 7.1. The difference between Real and Simulated interactions is shown in Figure 7.2. To the best of our knowledge, our study, which utilized two navigation tasks, is the first to compare human perceptions of robots obtained in real-world interactions with perceptions obtained from interactive simulations, where humans control a virtual avatar. We compared these human perceptions of a robot in real-world interactions and interactive simulations with perceptions of the robot after viewing a video recording. Our study protocol was approved by our Institutional Review Board.

7.3.1 Hypotheses

As shown in Figure 7.1, our two independent variables led to four conditions: Real-Interactive, Real-Video, Sim-Interactive, and Sim-Video. We studied whether these conditions had an effect on four aspects of human perceptions of the robot: Competence [86]; Discomfort [49]; Social Presentation, or "the robot's ability to appear to be a desirable social partner" [24]; and Social Information Processing, which captures social intelligence [24]. We also studied the effect of interactivity on perceived workload [91]. These measures are common in the Human-Robot Interaction literature [196, 148, 143, 90, 235].

Our first set of hypotheses focused on the idea that human perceptions of a mobile robot in the Real environment would differ from perceptions of the robot in the Simulated environment. These hypotheses were motivated by prior work that suggests that people's perception of a robot can vary between simulation and real-world interactions (e.g., [286, 159, 270]). In particular, Tsoi et al. [270] provided evidence that human perceptions of robots collected via video studies and compared to those collected using interactive, online simulations could differ, but did not compare them to observations obtained in real-world human-robot interactions. More specifically:

- **H1.** Human perceptions of the robot's competence (**H1a**), discomfort (**H1b**), social presentation (**H1c**), and social information processing (**H1d**) in the Real environment will differ from the Simulated environment.

Our second set of hypotheses tested the potential difference in human perception of a mobile robot between a participant interacting with a robot compared to a participant viewing an interaction with another person in a video. This hypothesis is motivated by the common use of videos in HRI studies, and the growing use of interactive simulations as a potential replacement [292, 231, 270]. Prior work suggests that people may perceive a robot more positively when physically present [157] and

that people may be influenced by co-present robots (e.g., [19, 111]).

- **H2.** Human perceptions of the robot's competence (**H2a**), discomfort (**H2b**), social presentation (**H2c**), and social information processing (**H2d**) will differ between interactive conditions (Sim-Interactive and Real-Interactive) and video-based conditions (Sim-Video and Real-Video).

Our third set of hypotheses considered data from the Real-Interactive condition as the gold standard for gathering human perceptions of robots. Then, because video observations lack interactivity in comparison to interactive simulations, we suspected that human perceptions collected with the Sim-Video and Real-Video conditions would be less similar to those obtained in the real world than the perceptions obtained with the Sim-Interactive condition.

- **H3.** Human perceptions of the robot's competence (**H3a**), discomfort (**H3b**), social presentation (**H3c**), and social information processing (**H3d**) in video-based conditions (Sim-Video and Real-Video) are more similar to the Sim-interactive condition than to the Real-Interactive condition.

Our fourth and final hypothesis is motivated by prior work that associates embodied and interactive experiences with low workload. For example, Wang et al. [291] found that robot agent embodiment resulted in lower perceived workloads during interaction with robotic agents compared to voice-only agents. Tsoi et al. [270] found partial support for lower perceived workload when completing an HRI survey that involved providing perceptions of a robot in interactive interactions compared to a survey that involved providing perceptions based on video observations

- **H4.** The Interactive conditions will lead to a lower perceived workload by participants than the Video conditions.

7.3.2 Participants

In total, we recruited 213 participants for our study. For the Real-Interactive condition, participants were recruited via flyers and word of mouth. Participants for all other conditions were recruited online using the Prolific crowdsourcing platform.

All the participants were at least 18 years old, had normal or corrected-to-normal vision, and were fluent in English. The participants in the Real-Interactive condition were required to be able to walk comfortably and stand for the duration of the study (20-30 minutes). Participants in the online portion of the study were limited to those on non-mobile devices, such as laptops and desktop computers to ensure a reasonable screen size on their device and the ability to control the virtual avatar in simulation using a physical keyboard.

We excluded 53 participants from analyses because 35 participants in an Interactive condition had incomplete video recordings due to technical issues or had incomplete surveys, 14 participants had other technical issues or did not follow directions, and 4 accidentally participated in the Sim-Video condition after participating in the Sim-Interactive condition.

Among the final 160 participants (40 per condition), 90 participants identified as male, 66 as female, 2 as non-binary, 1 as genderqueer, and 1 declined to state their gender. Additionally, 32 participants were between ages 25-34, 50 were between ages 35-44, 40 were between ages 45-54, 23 were between ages 55-64, 13 were between ages 65-74, and 2 were between ages 75-84. On average, the participants indicated neutral familiarity with robots on a 7-point scale ($M = 3.91, SE = 0.13$). The online participants had an average Internet speed of 163.46 Mbps ($SE = 15.86$), which was in line with prior use of SEAN-EP [270].

7.3.3 Setup

For the Real-Interactive condition, the experiment was conducted in a laboratory room on a university campus in the United States. The room contained physical obstacles consisting of EverBlock construction blocks, as shown in Figures 7.1(a) and 7.2(a). There were also four distinct pieces of artwork on easel stands positioned in the corners of the room. A close-up photo of one of the pieces of artwork in the real laboratory environment is shown in Figure 7.2(b).

We designed our study such that a robot, controlled by the ROS Navigation Stack with Social Cost Layers [162], autonomously navigated near the participant to jointly complete two tasks: the *Follow Task* and the *Art Task*. The Follow Task was designed to place the participant's focus on the robot throughout the interaction. Follow tasks are typical for robots that serve as tour guides and have been investigated in the past in social navigation [43, 191, 222, 186]. Meanwhile, we designed the Art Task to allow participants to observe the robot's movement during a more dynamic and complex navigation task. These tasks are further described in the next section. Importantly, the robot that we used in the study was a Pioneer 3-DX on which we affixed a laptop, oriented with the screen pointing forward, to allow for robot communication with the participant. We also attached a depth sensor and localization beacon to the robot.

The participants in the Real-Interactive condition wore a GoPro camera on their chest (as in Figure 7.2(a)) to record videos from a first-person perspective while completing study activities. HTC Vive Trackers were used to localize the robot and the participants. Also, the participants used a custom web application on a mobile phone, which we provided, to do task-specific actions. This included pressing a button on the phone to begin each task and recording their answers to survey questions. The web application was also used to display text on the robot's laptop.

For the Sim-Interactive condition, we modeled the laboratory room used for the Real-Interactive condition as well as the Pioneer robot using the Unity game engine

and SEAN 2.0 [274]. Figures 7.1(b), 7.1(d), 7.2(c), and 7.2(d) illustrate the virtual world that we created for the study. In addition, we used SEAN-EP [270] to embed our simulation in a Qualtrics web survey, which gathered participants' demographics data and all other relevant measures regarding their experience of virtual human-robot interactions. The participants used their keyboards to control a virtual avatar in the SEAN simulations and to complete the same activities as in the Real-Interactive condition.

For the Real-Video and Sim-Video conditions, we used recordings of participants' interactions with the robot in the real-world lab and the virtual re-creation, respectively. A GoPro camera worn by participants in the Real-Interactive condition (as in Figure 7.2(a)) was used to record the interactions that were observed by participants in the Real-Video condition. For the Sim-Video condition, we used SEAN 2.0 to save video recordings of the human-robot interactions that happened under the Sim-Interactive condition. The recordings were made from the perspective of the virtual avatar that was controlled by a human in SEAN. In order to ensure participants in the Video condition were able to understand what the robot was communicating, we added captions to all videos which displayed the same text that was shown on the robot's laptop screen. We did not use audio in the simulation or the videos due to the difficulity of generating realistic audio. An example of the captions is provided in Figures 7.1(c) and 7.1(d). The videos were then embedded in a Qualtrics survey like the one used for the Real-Interactive condition.

7.3.4 Procedure

At the beginning of the study, the participant provided demographic information (as in Section 7.3.2). Then, the participant continued on to complete the study's four phases: 1) Introduction, 2) Follow Task, 3) Art Task, and 4) Closing. In each task, the participant was specifically asked to pay attention to how the robot moved.

Phase 1: Introduction. In the Real-Interactive condition, the participant was introduced to the robot by an experimenter who told them that they would interact with the robot through a series of tasks. Then, the experimenter assisted the person as they put on the GoPro chest harness to record their activities during the study. In the Sim-Interactive condition, the participant completed a walk-through tutorial that showed them the virtual Pioneer robot and their randomly assigned avatar. The walk-through then explained how to navigate the simulated lab. In the Real-Video and Sim-Video conditions, the participant was given text instructions indicating that they would watch videos of a person or avatar interacting with a robot. The participant was also shown an image of the robot to familiarize the person with the Pioneer 3-DX platform.

Phase 2: Follow Task. In the Real-Interactive condition, the participant was instructed to move to a specific marker on the floor and then press a button on the mobile device to begin the follow task. Then, the participant followed the robot along a pre-defined path, which was composed of four segments.

The path involved navigating around EverBlock construction blocks placed throughout the room, as shown in Figures 7.2(a) and 7.2(c).

After following the robot along each of the four path segments, the participant answered survey questions about their impression of the robot. In the Sim-Interactive condition, the participant completed the same task but in a SEAN simulation.

For the Real-Video and Sim-Video conditions, we paired each participant with a study session that involved Real-Interactive and Sim-Interactive participation, respectively. Then, the videos of the Follow Task from the Interactive sessions were shown to the participants in the Video conditions. In this manner, a participant in Real-Video and Sim-Video conditions was able to watch recordings

of the task and answer survey questions about their impression of the robot in the videos, as in the Interactive conditions.

Phase 3: Art Task. In the Real-Interactive condition, the participant was told that there had been an art heist in the lab and some of the art had been replaced with fakes. The participant and the robot were tasked with collecting information about the four art pieces in the laboratory to help the experimenters figure out which were real and which were fake. Figure 7.2(b) displays one of the art pieces in the real world, and Figure 7.2(d) shows it in simulation. For each of the four art pieces, a participant performed the following steps:

1. The participant was directed to find the robot.
2. Once the person found the robot, a text message was displayed on the robot's computer screen which instructed them to follow it.
3. The robot then led the participant to a piece of artwork.
4. The participant was instructed via text on the robot's computer screen to count the number of a given object shown in the art piece.
5. After instructing, the robot moved away to a different location and waited for the participant to complete the object counting.
6. The participant provided their answer to the counting request using the mobile device and was directed to find the robot again to repeat the process for the next art piece.

The Art Task was designed so that the person and the robot would engage in more dynamic interactions than in the Follow Task. In this case, while the person was counting objects in an art piece, the robot moved far from the participant and waited until they completed counting the objects in the picture. Only when the participant started moving away from the picture did the robot

start to move back towards the person. Then, both the robot and participant moved towards each other and soon thereafter engaged in face-to-face or side-by-side spatial formations (e.g., as in [117, 304]).

In the Real-Video and Sim conditions, the description of the Art Task was provided in text before the participant began the task.

Also, in the Sim-Interactive condition, the participant used an interface which we implemented in the simulation to record their responses to the counting request by the robot. Meanwhile, in the Video conditions, the participant recorded their answers using the Qualtrics web survey. This survey included videos from Interactive conditions using the same participant-session pairing explained for the Follow Task.

Phase 4: Closing. Finally, the participant provided their impressions of their perceived workload for the tasks in the study.

In-person participants in the Real-Interactive condition were paid $15.00 USD per hour, rounded to the nearest 10-minute increment.

Participants in all other conditions completed the study online using Prolific. They were paid $5.00 USD as we estimated the online study sessions to take 20 minutes.

7.3.5 Dependent Measures

We measured 2 aspects of participants' experience during our study using widely adopted survey measures in HRI:

Human Perceptions of the Robot. We measured four aspects of human perceptions of the robot: 1) Competence, 2) Discomfort, 3) Social Presentation, and 4) Social Information Processing. The first two aspects were measured using the Robot Social Attributes Scale (RoSAS) [49], which includes robot Competence and Discomfort factors. The items were answered in relation to how the

robot moved during the tasks. Ratings for the Competence and Discomfort scales were gathered using a 7-point responding format ranging from 1 (Definitely Not Associated) with the robot to 7 (Definitely Associated), which was the same as the original RoSAS responding format.

Robot Social Presentation and Social Information Processing were measured using the short-form of the Perceived Social Intelligence (PSI) questionnaire [24]. The Social Presentation scale had a total of 7 items, all of which began with "This robot..." and ended with statements such as "enjoys meeting people," and "cares about others." The Social Information Processing scale had a total of 13 items, which started with "This robot..." and ended with statements like "responds appropriately to human emotion" or "can figure out what people think.". Ratings for PSI statements were gathered on a 5-point responding format ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), which was the same as the original PSI responding format.

For each scale, we aggregated responses across items to calculate a composite measure after confirming high internal reliability. The Cronbach's α values were 0.90 for Competence, 0.76 for Discomfort, 0.76 for Social Presentation, and 0.94 for Social Information Processing. The Cronbach's α value for each aspect we measured was within the 0.7 to 0.95 acceptable value range [252].

Perceived Workload. We used items from the NASA Task Load Index (TLX) [91] to assess the perceived workload for the Follow and Art Tasks. Perceptions of Mental Demand, Physical Demand, Temporal Demand, Effort, and Frustration were gathered on a 7-point responding format from 1 (lowest) to 7 (highest). The 7-point responding format was used for consistency in the responding format with the other scales. The 7-point format was chosen over the 5-point format because responding formats with 6 or more categories have been shown

to correlate better[211]. Example survey items included "How mentally demanding were the tasks?" (Mental Demand) and "How insecure, discouraged, irritated, stressed, and annoyed were you?" (Frustration). The Cronbach's α for the NASA TLX survey items was 0.75, which is within the 0.7 to 0.95 range of acceptable values [252].

7.3.6 Analysis

We analyzed the results by task (Follow and Art) in two ways. First, we fitted linear mixed-effect models for all dependent measures with fixed effects for Interaction Environment (Real or Simulation) and Interactivity (Interactive participation or Video observation). We also assigned a unique identifier, Session ID, to each Interactive study session, which was added as a random effect in our linear model. A linear mixed-effect model was used due to the hierarchical nature of the data, i.e., Participant ID was nested within Session ID. This allowed us to associate the experience in the Interactive conditions, from which we made videos of human-robot interactions, with the corresponding data in the Video conditions. Unless otherwise noted, we used the Restricted Maximum Likelihood (REML) method for model estimation [199]. A linear mixed model was used for model estimation instead of ANOVA because of the nested nature of the data, i.e., Participant ID was nested within Session ID. Nesting was necessary because the video-condition stimuli were generated from a recording of the Interactive condition, which resulted in the interactive data and corresponding video recordings being paired. Note that within the paired data, the participant who interacted with the robot (either in the Real environment or simulation) was not the same as the participant who watched the video, so a unique Participant ID was used to identify all participants. Second, because H3 considered the Real-Interactive condition as the methodology that

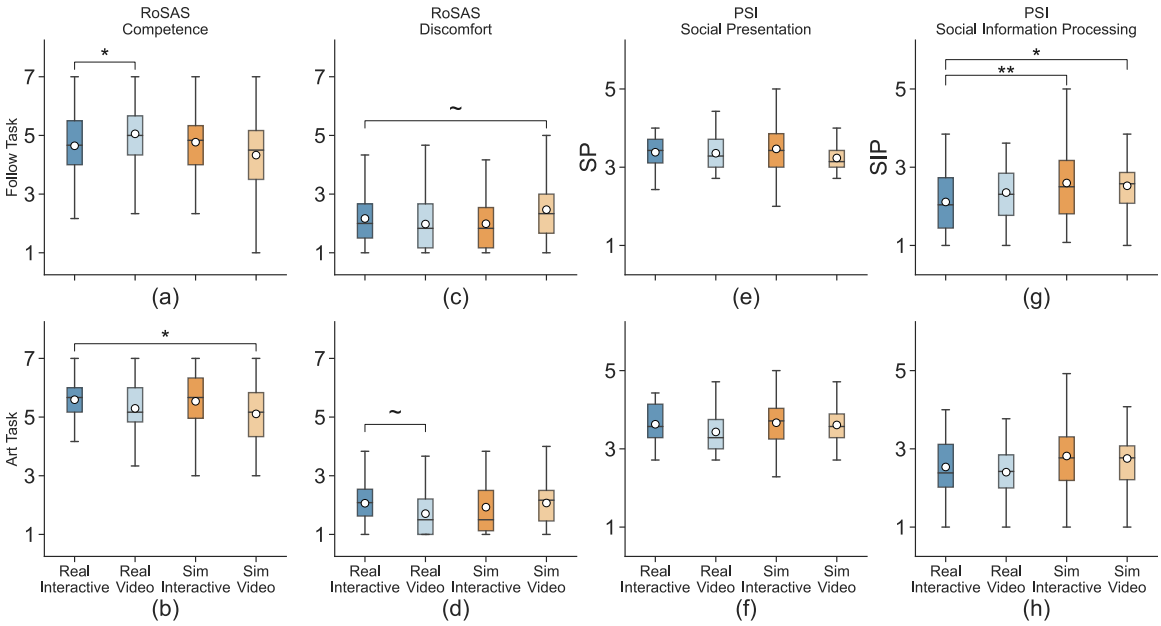


Figure 7.3: Contrast results for RoSAS Competence (a,b), RoSAS Discomfort (c,d), PSI Social Presentation (e,f), and PSI Social Information Processing (g,h) by task. Box plots span the first to third quartile, a dark grey horizontal line through the box indicates the median, and a white circle indicates the mean. Box plot whiskers extend to ±1.5 times the Interquartile Range. The $\sim$ indicates $p < 0.10$, * indicates $p < 0.05$, and ** indicates $p < 0.001$.

provides gold-standard results, we performed treatment contrasts between the Real-Interactive condition and all other conditions.

7.4 Results

7.4.1 Perceptions of the Robot

Competence

The linear mixed model analysis per task revealed significant effects. In particular, for the Follow Task, we found Interaction Environment to have a significant effect on Competence, $F(1, 156) = 4.30, p = 0.04$. The effect size, as measured by Cohen's d, was $d = 0.16$, indicating a very small effect. A post-hoc t-test showed that people perceived the robot to be significantly more competent in the Real condition ($M =$

$4.85, SE = 0.06$) than in the Simulated condition ($M = 4.55, SE = 0.07$). The linear mixed model analysis on the Art Task showed that only Interactivity had a significant effect on Competence, $F(1, 156) = 5.39, p = 0.022$. The effect size, as measured by Cohen's d, was $d = 0.18$, indicating a very small effect. A post-hoc t-test indicated that competence ratings were significantly higher for Interactive participation ($M = 5.56, SE = 0.11$) than for Video observation ($M = 5.20, SE = 0.11$).

Comparing the Real-Interactive condition as the baseline condition against three other conditions with treatment contrasts revealed that the Real-Video condition significantly differed from the Real-Interactive condition in the Follow Task, $F(1, 156) = 3.94, p = 0.05$. The effect size, as measured by Cohen's d, was $d = 0.22$, indicating a small effect. Specifically, compared to interacting with the robot in the real world ($M = 4.65, SE = 0.09$), participants watching videos of the robot interacting with someone else in the real world perceived the robot to be even more competent ($M = 5.05, SE = 0.08$). For the Art Task, only the Sim-Video condition was significantly different from the Real-Interactive condition, $F(1, 156) = 4.79, p = 0.03$. The effect size, as measured by Cohen's d, was $d = 0.24$, indicating a small effect. This suggests that compared to watching a video of a person interacting with the robot in simulation ($M = 5.11, SE = 0.16$), participants who interacted with the robot in the real world viewed it to be even more competent ($M = 5.59, SE = 0.14$). These results are shown in Figures 7.3(a) and 7.3(b).

Discomfort

The linear mixed model analyses on both tasks resulted in no significant main effects on discomfort.

The contrast analyses for the Discomfort responses in the Follow and Art Tasks led to no significant differences. However, the discomfort ratings in the Sim-Video condition were marginally different from the Real-Interactive ratings in the Follow

Task, $F(1, 156) = 3.57, p = 0.06$. The effect size, as measured by Cohen's d, was $d = 0.21$, indicating a small effect. This indicates that compared to watching a video of a simulation ($M = 2.47, SE = 0.08$), participants who interact with a robot in the real world may view the robot as less discomforting ($M = 2.17, SE = 0.07$). Additionally, discomfort in the Real-Video condition was marginally different from the Real-Interactive condition in the Art Task, $F(1, 156) = 3.48, p = 0.06$. The effect size, as measured by Cohen's d, was $d = 0.21$, indicating a small effect. This indicates that compared to interacting with a robot in the real world ($M = 2.06, SE = 0.13$), participants who watch a video of the real-world robot interacting with another participant may view the robot as less discomforting ($M = 1.71, SE = 0.13$). These results are shown in Figures 7.3(c) and 7.3(d).

Social Presentation

The linear mixed model analyses and the treatment contrasts per task showed no significant effects on Social Presentation ratings. In general, most ratings were neutral in the Follow Task and slightly positive in the Art Task, as shown in Figures 7.3(e) and 7.3(f). The slight increase in Social Presentation perceptions for the Art Task was expected because the task involved more complex interactions than the Follow Task, as indicated in Section 7.3.4.

Social Information Processing

The linear mixed model analysis on Social Information Processing for the Follow Task revealed a significant main effect of Interaction Environment on the ratings, $F(1, 157) = 6.71, p = 0.01$. The effect size, as measured by Cohen's d, was $d = 0.41$, indicating a small effect. A post-hoc t-test indicated that people perceived the robot as better able to process social information in the Simulated condition ($M = 2.56, SE = 0.09$) than in the Real condition ($M = 2.23, SE = 0.09$). The linear

mixed model analysis for the Art Task also indicated that Interaction Environment had a significant effect on Social Information Processing, $F(1, 157) = 5.02, p = 0.03$. The effect size, as measured by Cohen's d, was $d = 0.35$, indicating a small effect. The post-hoc test indicated that ratings were higher for the Simulated environment ($M = 2.79, SE = 0.10$) than for the Real environment ($M = 2.47, SE = 0.09$).

The contrast analyses on the Follow task indicated a significant difference in Social Information Processing ratings between the Sim-Interactive and Real-Interactive conditions, $F(1, 156) = 7.29, p = 0.008$, as well as between the Sim-Video and Real-Interactive conditions, $F(1, 156) = 5.31, p = 0.02$. The effect sizes, as measured by Cohen's d, were $d = 0.60$ and $d = 0.52$, respectively, indicating a medium effect for both contrasts. This suggests that compared to interacting with the robot in the real world ($M = 2.11, SE = 0.12$), participants viewed the robot as more capable of processing social information when interacting with it in simulation ($M = 2.60, SE = 0.15$) and when viewing it in a video in simulation ($M = 2.53, SE = 0.11$). These results are shown in Figure 7.3(g). For the Art Task, the contrast analyses showed no significant differences on Social Information Processing with respect to Real-Interactive. The results for the Art Task are shown in Figure 7.3(h).

7.4.2 Perceived Workload

We analyzed the perceived workload with linear mixed model analyses that included Interaction Environment (Real or Simulation), Interactivity (Interactive participation or Video observation) and their interaction as main effects. Also, we added Session ID as a random effect. In the case of workload, we did not perform contrast analyses as in Section 7.4.1 because H4 did not consider the Real-Interactive condition as a specific baseline for comparison.

The average ratings for Physical Demand and Temporal Demand were $1.48(SE = 0.07)$ and $1.76(SE = 0.08)$, respectively. We found no significant effects on these

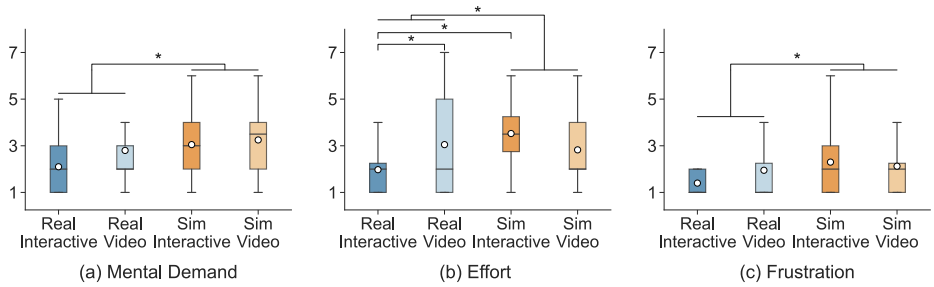


Figure 7.4: Perceptions of Mental Demand, Effort, and Frustration by condition: Real-Interactive, Real-Video, Sim-Interactive, and Sim-Video. Box plots span the first to third quartile, a dark grey horizontal line through the box indicates the median, and a white circle indicates the mean. Box plot whiskers extend to ±1.5 times the Interquartile Range. The * symbol indicates $p < 0.05$.

measures.

Interaction Environment had a significant effect on Mental Demand ($F(1, 156) = 8.60, p = 0.004$), Effort ($F(1, 156) = 6.94, p = 0.009$) and Frustration ($F(1, 156) = 5.77, p = 0.017$). The effect sizes, as measured by Cohen's d, were Mental Demand $d = 0.46$, Effort $d = 0.42$, and Frustration $d = 0.38$, indicating small effects. The post-hoc t-test on Mental Demand indicated that participants provided higher ratings in the Simulated environment ($M = 3.15, SE = 0.16$) than in the Real environment ($M = 2.45, SE = 0.18$). The distribution of Mental Demand ratings is shown in Figure 7.4(a). Likewise, in the case of Effort, the post-hoc test showed that the ratings in the Simulated environment ($M = 3.18, SE = 0.18$) were significantly higher than those in the Real environment ($M = 2.51, SE = 0.19$), as shown in Figure 7.4(b). Finally, the post-hoc test for Frustration revealed that participants felt more "insecure, discouraged, irritated, stressed and annoyed" with the Simulated environment ($M = 2.21, SE = 0.17$) than with the Real environment ($M = 1.68, SE = 0.15$). Figure 7.4(c) shows the distribution of results for Frustration.

Interactivity had no significant effect on Mental Demand or Frustration; however, we found an interaction effect between Interaction Environment and Interactivity on Effort, $F(1, 156) = 12.45, p < 0.001$, $R^2_{Adjusted} = 0.10$. A post-hoc Tukey HSD test indicated that the Effort for the Real-Interactive condition ($M = 1.98, SE =$

0.17) was significantly lower than for Real-Video ($M = 3.05, SE = 0.32$) and Sim-Interactive ($M = 3.53, SE = 0.26$).

7.5 Discussion

In our first set of hypotheses, our results indicated some support. Results showed a significant difference between perceptions of the robot in simulation compared to the real environment. In particular, we found higher Competence ratings (H1a) for the robot in the real laboratory environment than in simulation, although the effect was small. We suspect the difference was due to the greater level of visual realism exhibited by the real robot [286]. Also, we found that the real robot was perceived as less capable of processing social information than the simulated robot (H1d). Social information processing (SIP) refers to the robot's ability to perceive the social behaviors, emotional states (including desires), and cognitions (including beliefs) of nearby people [24]. The effect for SIP was larger than the effect for Competence, but still small. It could be that human perceptions about the robot's social information processing abilities were influenced by their virtual avatar in the simulations, which behaved in a much simpler way than people could in the real laboratory environment and looked less realistic as well.

We found evidence for some of our second set of hypotheses, which posited that human perceptions of the robot will differ between Interactive participation and Video observations. In particular, for the Art Task, participants viewed the robot as more competent with Interactive participation than when human-robot interactions were observed in Videos. Although the effect size was small, our results were surprising because they did not align with the results by Tsoi et al. [270], who compared human perceptions of the competence (H2a) of a Kuri robot in interactive SEAN simulations and in videos of the simulation. Beyond the fact that Tsoi et al. [270] did not consider

real-world interactions, we believe that the inconsistency in findings could be due to three reasons: 1) the laboratory environment used in our work had more obstacles and fewer people than the one used in [270]; 2) we used a Pioneer robot which could set different initial human expectations than the Kuri robot used in [270]; and 3) the Art Task was more complex than the Follow Task, and [270] only studied situations where participants followed the robot. Future work should investigate which factors specifically affect human perceptions of the competence of a robot between HRI studies involving Interactive participation and Video observation.

As to our third set of hypotheses, we obtained some evidence that human perceptions of the robot in the Video conditions are more dissimilar to the Real-Interactive condition than those in the Sim-Interactive condition. For example, contrast analyses indicated that robot competence (H3a) was significantly different between the Real-Interactive condition and the Real-Video conditions (for the Follow Task) and between the Real-Interactive and Sim-Video conditions (for the Art Task). No significant differences were found for competence between Real-Interactive and Sim-Interactive conditions. In terms of discomfort (H3b), we found trends that suggested similar differences but for the opposite task – compare Figures 7.3(a) with 7.3(c), and Figures 7.3(b) with 7.3(d). Again, no significant differences were found for discomfort between Real-Interactive and Sim-Interactive. However, for social information processing (H3d), Real-Interactive led to significantly different results than both Sim-Video and Sim-Interactive. This last result was unexpected and not in line with our hypothesis. Overall, the main takeaway from these results is that perceptions of robots gathered through video observation and interactive simulation studies may not always translate to real-world interactions.

Finally, we found only a small amount of evidence in support of our last hypothesis, which stated that cognitive load would be lower for Interactive participation than Video observations. More specifically, only perceived effort was significantly lower

for the Real-Interactive condition than for the Real-Video condition. Interestingly, most of our results in regard to workload were instead about differences between the Real and Simulated environments, including differences for mental demand, effort, and frustration. We thought that this result could be due to the fidelity of our SEAN 2.0 simulations. Although SEAN 2.0 generates the renderings through Unity and there is potential to make these simulations photo-realistic, our virtual laboratory environment looked much simpler than the real-world lab (as can be seen in Figure 7.1 and Figure 7.2). For example, while humans are adept at identifying coherent concepts from the visual clutter typically found in the real world [195], increased participant effort may be necessary to interpret and interact with the robot in the simulation environment, which contains a distribution of visual clutter different from the real world. In the future, exploring how environmental clutter affects human perceptions of robots in HRI could be an interesting avenue of research, for example, by comparing with experiments in simulation that incorporate real-world clutter [298]. Another factor to consider is the usability and computing experience of the different systems implemented for each condition, which may have also had an impact on participant workload. Overall, this is a first step towards a better understanding of how different methodologies can influence the perceptions of mobile robots for social navigation. We hope future HRI studies can explore this direction on a larger scale.

7.5.1 Limitations

First, we conducted our study with only one simulation environment (SEAN 2.0 [270]). It would be interesting to verify in the future if our results hold with other types of simulators, e.g., built using other game engines like Unreal [166] or with lower-fidelity like Gazebo [209]. Second, as with all simulations, our simulated environment and the videos thereof were not perfect replicas of the real world. In the future, it would be interesting to investigate the impact of factors such as the lack of audio in simulation,

which could have influenced perceptions of the robot in the Sim and Video conditions, the size and the resolution of the display or Head Mounted Display, and properties of the randomly assigned virtual avatars, such as gender, which may not match that of the participant. Third, we focused on investigating people's perceptions of robots using subjective responses to well-established questionnaires. However, future research could benefit from including behavioral outcomes, like proxemics measures [101], when comparing research methodologies for social robot navigation. When evaluating results for other tasks, perhaps other behavioral measures like teamwork efficiency [20], could be used instead. Lastly, it would be interesting to investigate to what extent the crowdsourcing setup that we used to gather data in three experimental conditions affected our results. In particular, one could imagine replicating our study in the future with 100% in-person participants, such that no participant is subject to the distractions and technical challenges that often arise with remote participation through crowdsourcing [281].

7.6 Guidelines for Methodology Selection

The choice of methodology is one of the many considerations that a researcher must evaluate when approaching new experimental questions in HRI. The primary considerations are time and cost. Ideally, minimal time is required to set up and complete the study while minimizing the cost. Although in-person user studies are the gold standard, often video studies are used. Video studies can allow crowd-sourcing of user feedback, which scales quickly, but the quality of responses can vary if participants are not engaged with or focused on the video. With recent technological advancements, interactive simulations can now scale with the use of crowd sourcing [270], they can encourage a participant to remain engaged with the task or detect if the person is not engaged. Other considerations include the availability of a real robot,

the safety of the task experienced via different methodologies, and the quality of the simulation along the dimensions of importance. Perhaps in the future, we may have widely available, photo-realistic, real-time, interactive simulations that will decrease the gap between methodologies. However, until this is the case, researchers should carefully consider the tradeoffs.

7.7 Summary

We investigated how people perceived the competence, discomfort, social presentation, and social information processing of a mobile robot during two navigation tasks. Our study compared methodologies with different Interaction Environments (Real vs. Simulated) and Interactivity (Interactive participation vs. Video observations). We found significant differences in human perceptions of a mobile robot when an interaction was experienced in the real world compared to simulation. In addition, we found significant differences in human perceptions when participants watched a video of a human-robot interaction compared to when they participated in the interaction, experiencing a two-way flow of information.

Overall, our study suggests that results from user studies that rely on video observations and interactive simulations may not always mirror human perceptions of robots in real-world HRI. Importantly, we found trade-offs between Real-Video, Sim-Video, and Sim-Interactive methodologies. First, our work provides initial evidence that suggests that human perceptions of a robot in video studies may be less similar to real-world in-person studies in comparison to interactive simulation studies. This suggests that an interactive simulation should be preferred over observing videos. Second, we found that participants perceived greater workload in simulated environments than in real-world environments. Lesser workload in the real-world may help explain why, in some prior work, humans preferred in-person human-robot interactions more

than simulated or video interactions [285, 19]. Also, our results with respect to workload suggest that Real-Video may be preferred over Sim-Video and Sim-Interactive. Ultimately, it is important to consider whether human perceptions are likely to translate to the real world, and human workload when choosing a methodology other than in-person studies to investigate human-robot interactions.

Chapter 8

Beyond Collecting Human Feedback: Predicting Human Perceptions of Social Robot Performance*

In the previous chapters, we explored scalable methods for collecting human feedback on a mobile robot's competence, social appropriateness, and overall performance, highlighting the benefits of interactive, scalable simulations. This chapter takes the next step: instead of asking people for feedback, we ask if machine learning methods can be used to infer how a robot is being perceived by a nearby human. To explore this question, we introduce the SEAN TOGETHER Dataset, which consists of human-robot interactions in VR. The dataset also includes behavioral cues such as pedestrian motion, eye gaze, and facial expressions. The dataset is labeled with human feedback supplied as 5-point ratings of perceived robot performance along 3 dimensions: competence, surprise, and intention. By learning to predict people's perceptions from these signals, this chapter aims to enable socially aware robots to detect

Parts of this chapter were originally published as Qiping Zhang, Nathan Tsoi*, Mofeed Nagib, Booyeon Choi, Jie Tan, Hao-Tien Lewis Chiang, Marynel Vázquez. (2024). Predicting Human Perceptions of Robot Performance During Navigation Tasks. In *ACM Transactions on Human-Robot Interaction (THRI)* [309]. * indicates equal contribution.

when interactions go wrong so that in the future, they can adapt their behavior to better align with human expectations and values. Our demonstration on real-world data shows that the models trained with VR data can generalize to real-world data, confirming the potential usefulness of our approach in real-world settings.

8.1 Introduction

As a scalable alternative to measuring subjective perceptions of robot performance through surveys, recent work in Human-Robot Interaction (HRI) has explored using *implicit* human feedback to predict these perceptions [14, 71, 239, 307]. These are communicative signals that are unintentionally exhibited by people [139]. They can be reflected in human actions that change the world's physical state [226] or can be nonverbal cues, such as facial expressions [71, 239] and gaze [181, 14], displayed during social interactions. Implicit feedback serves as a burden-free information channel that sometimes persists even when people don't intend to communicate [135].

We expand the existing line of research on predicting perceptions of robot performance from nonverbal human behavior to dynamic scenarios involving robot navigation. Prior work has often considered stationary tasks, like physical assembly at a desk [240] or robot photography [307], in laboratory environments. We instead explore the potential of using observations of the body motion, gaze, and facial expressions of a person to predict their perceptions of a robot's performance while a robot guides them to a destination in a crowded environment. These perceptions correspond to subjective opinions of how well a robot is performing the navigation task. Predicting them in crowded navigation scenarios is more challenging than in stationary settings because human nonverbal behavior can be a result of not only robot behavior, but also other interactants in the environment. Further, because of motion, nonverbal responses to the robot may change as a function of the environment. For example,

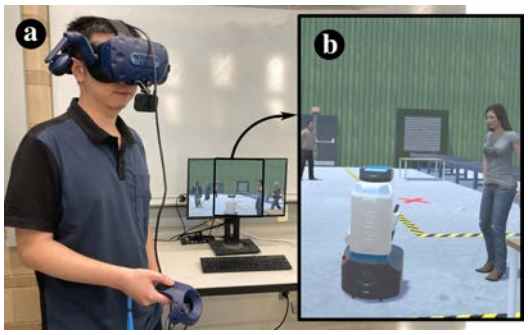


Figure 8.1: Data collection. Humans controlled an avatar in the simulation with VR (a) while they were guided by a Fetch robot (b). The screen on the desk shows what the user saw.

imagine that the person that follows the robot looks downwards. This could reflect paying attention to the robot, or be a result of the person inspecting their nearby physical space, which varies during navigation.

To study implicit feedback during navigation tasks, we performed a systematic data collection using the Social Environment for Autonomous Navigation (SEAN) 2.0 [273] with Virtual Reality (VR) [308].* Humans took part in the simulations through an avatar, which was controlled using a VR headset, as in Fig. 8.1. The headset enabled immersion and allowed us to capture implicit feedback features like gaze. Also, it facilitated querying the human about the robot's performance as navigation tasks took place. We considered robot performance as a multi-dimensional construct, similar to [307], because humans may care about many aspects of a robot's navigation behavior, as discussed in the social robot navigation literature [90, 176, 87].

Then, we studied fundamental questions about the value of implicit feedback signals in predicting subjective perceptions of robot performance using the VR data. First, we investigated to what extent humans can predict a person's perceptions of the robot's performance (along the dimensions of perceived navigation competence, surprising behavior, and clear intention during navigation). Predictions were made based on visualizations of observations of the human-robot interaction, as recorded

*Dataset and source code available at: `https://sean-together.interactive-machines.com/`.

in our VR navigation dataset. Second, we investigated how well various supervised learning models do this type of inference in comparison to humans. Third, we studied the generalization capabilities of supervised learning methods to users unseen at training time.

Our analyses bring understanding to the complexity of predicting humans' perceptions of robot performance in navigation tasks and enabled us to finally conduct a real-world demonstration in which a robot uses a machine learning model to predict how a human perceives it in a university campus. We conclude this paper by discussing the implications of our results for implementing autonomous systems that infer human perceptions of robot performance using implicit feedback in real-world navigation scenarios. We hope that our recommendations facilitate future efforts to make robots more aware of their failures during navigation [257], as well as facilitate aligning robot behavior to human preferences based on implicit feedback [177, 71, 62].

8.2 Related Work

This section discusses prior work in relation to our contributions. First, we discuss human perceptions of robot performance, especially in regard to robot motion. Then, we distinguish between explicit and implicit human feedback, the latter being the focus of our work. Finally, we briefly review data collection methodologies in HRI.

8.2.1 Perceptions of Robot Performance

Understanding human perceptions of robot performance is important. The perceptions can be used to evaluate robot policies [251, 161, 206] and to create better robot behavior [254, 181, 70, 31], increasing the likelihood of robot adoption. In this work, we focus on inferring three robot performance dimensions relevant to navigation [90]:

competence, *surprising behavior*, and *clear intent*. Robot competence is a popular performance metric in robotics [48], especially in robot navigation [175, 271, 12]. In our work, competent robot navigation behavior corresponds to effectively guiding a human to a destination. Surprising behavior violates expectations, which is often considered undesired [16, 87] and may require explanations by the robot [40]. Meanwhile, clear intentions means the robot enables an observer to infer the goal of its motion [76]. Prior work suggests that if humans fail to anticipate the motion of a robot because it acts surprisingly or its intent is unclear, they will likely have trouble coordinating their own behavior with it [229, 77]. There are other perceptions about a robot beyond robot competence, surprising behavior, and clear intent that one may want to model in Human-Robot Interaction, like human perceptions of discomfort with a robot [48, 137] or perceived safety [223, 6]; however, this is out of the scope of the present work.

8.2.2 Implicit Human Feedback

We distinguish between explicit and implicit human feedback about robot performance. Explicit feedback corresponds to purposeful or deliberate information conveyed by humans to robots, e.g., through preferences [36, 243] or survey instruments [17, 175]. Meanwhile, implicit feedback are cues and signals that people exhibit without intending to communicate some specific information about robot performance, yet they can be used to infer such perceptions. Inferring performance from implicit feedback can reduce the chances of excessively querying users for explicit feedback in robot learning scenarios [219, 97], thereby minimizing the risk of feedback fatigue [160]. Learning from implicit feedback is not without challenges, however, as it can be difficult to interpret [71, 239]. For example, this can happen due to inter-person variability in facial expressions [100], similar signals being produced for different reasons [46], or signals changing over time as interactions progress [47].

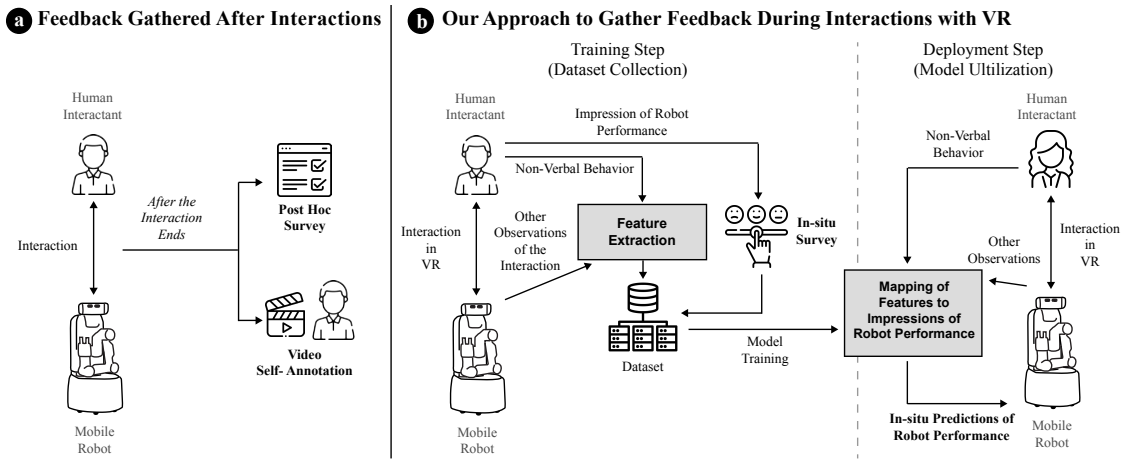


Figure 8.2: a) It is typical to gather explicit human feedback about robot performance using surveys after human-robot interactions conclude because interruptions by the experimenters can easily bias human-robot social encounters. Unfortunately, the feedback from surveys tends to be very limited, making it difficult to understand robot performance at a granular level. Alternatively, participants may complete video annotations of their experiences [308], but this can be time-consuming and taxing, especially in continuous navigation tasks. b) In this work, we first collect a dataset of human perceptions of a robot's performance by prompting participants *during* interactions using VR (Training Step in the diagram). Then, we use this explicit feedback to train models that infer human perceptions of robot performance based on observations of the interactions, especially including observations of human implicit feedback. The value of such a model is that once it is trained, it can be reused to estimate robot performance during new interactions (Deployment Step), without having to ask humans for explicit feedback as in the training step.

Our work considers a variety of nonverbal implicit signals, including gaze, body motion, and facial expressions, which have long been studied in social signal processing [282]. While in some cases these signals are treated as explicit feedback (e.g., to interrupt an agent [302]), we consider them implicit feedback because we do not prime humans to react in specific ways to a robot. As such, our work is closer to [71, 284, 177, 238, 46], which used these signals to identify critical states during robot operation, detect robot errors, and adjust robot behavior.

8.2.3 Data Collection in HRI: VR and Other Methodologies

Different kinds of HRI research methods have been used in the literature to gather interaction data, such as in-person user studies (e.g., [94, 261, 175]), observational

public data collections (e.g., [172, 131]), crowdsourcing studies (e.g., [41, 258, 119]), etc. See [25] for an introduction to these methods.

We considered different ways of conducting our data collection, but ultimately opted for gathering data with simulated human-robot interactions in VR for several reasons. First, in contrast to real-world data collection, simulation facilitated querying humans about their perceptions of robot performance during interactions and resulted in fewer negative consequences for interrupting the navigation task. This is illustrated in Figure 8.2. In lab studies, for instance, surveys that gather general perceptions of a robot are typically administered at the end of interactions to avoid interrupting the natural flow of events [307], which can cause unintended effects on collaborative tasks and interactants. In VR simulations, however, we can gather feedback in situ. We can freeze time during human-robot interactions, query a participant about their perceptions of robot performance through the VR display, and then resume the simulation as if the interruption had not occurred.

Second, we started our research by utilizing VR because simulations made interactions safer in contrast to those in the real-world. The reason is that we wanted to expose participants not only to good robot navigation behavior, but also to bad behavior. This was key for inducing a wide range of perceptions about robot performance during data collection and, thus, capturing varied implicit feedback. Prior work has used simulations in HRI for safety reasons as well [180, 115].

Third, in contrast to crowdsourcing data collection procedures, our in-person data collection reduced unrelated participant distractions [38] and minimized potential issues with participants' internet speed [118, 271]. Early in our research, we considered using interactive surveys [271] for our data collection while capturing implicit feedback signals through the webcams of remote participants (e.g., as in [46]). However, after testing both this setup and VR, we thought that the increased level of immersion afforded by VR was important to gather naturalistic feedback.

While we opted for using simulations in our work, they are not without limitations. In particular, simulations can result in a sim-to-real gap, as discussed before in HRI and other robotics areas (e.g., [32, 66, 65, 11, 158, 108]). This gap can emerge in HRI because of differences in physics between simulation and the real-world as well as the human-robot interactions in simulation not reflecting the real-world experience [108]. Indeed, prior work suggests that virtual robots may be perceived as more discomforting than real robots [158]. Thus, towards the end of this paper, we explored applying the insights from our work with VR data to a real-world demonstration, paving the path towards predicting perceptions of robot performance in real-world application scenarios. Being able to make such predictions opens up doors for adapting robot behavior to better align with human desires (e.g., by treating the predicted human perceptions as a reward signal in reinforcement learning [18]).

8.3 Problem Statement & Research Questions

We study if a person's perceptions of a robot's performance can be predicted using observations of their interaction in dynamic tasks involving navigation. Specifically, we aim to learn a mapping from a sequence of observations to an individual's reported perceptions at the end of the sequence (as in Fig. 8.2b). We consider multiple robot performance dimensions on a 5-point scale, as detailed later in Section 8.4.

Consider a dataset of observations and performance labels, $\mathcal{D} = \{(\mathbf{o}^i_{1:T}, y^i)\}$, where $\mathbf{o}_{1:T}$ is an observation sequence of length T, y is a performance rating given by a robot user at the end of the sequence, and i identifies a given data sample. We emphasize predicting a person's perceptions of a robot by considering observations of their implicit feedback. Thus, the observations $\mathbf{o}^i_t$ include features that describe the person's non-verbal behavior, such as their motion, gaze, and facial expressions. Also, the observations include features that describe the spatial behavior of all the agents

in the environment, the navigation task, and the space occupied by static objects. Given this data, we investigate three fundamental research questions:

1. ***How well can human observers predict a user's perceptions of robot performance?*** By answering this question, we obtain a human baseline for learning a function $f : \mathcal{O}_{1:T} \rightarrow \mathcal{Y}$, where $\mathcal{O}$ is the observation space and $\mathcal{Y}$ is performance. Also, through this question, we study the impact of two types of observations in the prediction task: observations that describe fine-grained facial expressions for a robot user and other observations about the user, the robot, and their environment. As mentioned earlier, observations of fine-grained expressions have gained popularity in recent work to infer human perceptions of an agent's behavior [71, 46, 307, 240]. Other observations (e.g. body motion and nearby static obstacles) can be more easily computed in real-world navigation tasks, but their usefulness on a robot's ability to infer users' perceptions of their performance is less understood.

2. ***Can machine learning methods predict perceptions of robot performance as well as humans?*** Ultimately, we are interested in bringing us forward to a future where machine learning models facilitate evaluating robot performance at scale, without having to necessarily ask users all the time for explicit feedback (as in the Deployment Step of Fig. 8.2b). Thus, we evaluate various machine learning models to approximate the function f, as defined for the prior question.

3. ***How well can machine learning models generalize to unseen users?*** In future robot deployments, a robot may interact with completely new users. Thus, we analyze the performance of various machine learning models in predicting perceptions of robot performance according to users for whom the model had no data at training time.

We study the above questions using data from SEAN-VR [308], as described in the next two sections. Later, in Section 8.5, we leverage our findings in VR to create a real-world demonstration through which we investigate predicting human perceptions of robot performance in two university environments.

8.4 Data Collection with SEAN and VR

For our VR data collection, we leveraged the SEAN 2.0 simulator, introduced in Chapter 4 [273]. SEAN 2.0 integrates with the Robot Operating System (ROS) [212] and supports Virtual Reality via the SEAN-VR Extension [308]. Participants used a Vive Pro Eye VR device to control an avatar in a warehouse (as in Fig. 8.1(a)). The VR headset captured implicit signals from the participants, like eye and lip movements.

During data collection, the participants had to follow a Fetch robot that guided them to a destination that was unknown to them a priori but was marked by a red cross on the ground. Fig. 8.1(b) shows a first-person view of the simulation during robot-guided navigation. The Fetch robot was controlled with ROS in SEAN. The environment contained other algorithmically controlled pedestrians and warehouse obstacles provided by SEAN 2.0.

The participants provided ratings of robot performance through the simulation's VR interface. The frame rate of the rendering of the virtual environment in the participants' first-person view in VR was over 30 frames per second. Our data collection protocol, described below, was approved by our local Institutional Review Board and refined via pilots.

8.4.1 Participants

We recruited 60 participants using flyers and by word of mouth. They were at least 18 years old, fluent in English, and had normal or corrected-to-normal vision. Overall, 19 participants identified as female, 40 as male, and 1 as non-binary or third gender. Most of them were university students, and their ages ranged from 18 to 43 years old. Participants were somewhat familiar with robots, as indicated by a mean rating of M = 4.20 (with standard error SE = 0.18) on a 7-point Likert responding format (1 being lowest). Yet, they were somewhat unfamiliar with VR (M = 3.72, SE = 0.20). No participant had prior experience with SEAN or social robot navigation in VR.

8.4.2 Data Collection Procedure

Protocol: A data collection session took place as follows. First, the participant provided demographic data. Second, the experimenter introduced the robot, explained the navigation task in which the participant was to follow the robot, and demonstrated how to use the VR device to control their avatar in SEAN and label robot performance. Third, the participant experienced four navigation tasks with the robot, each with a particular starting position and destination. For consistency, the pedestrians were controlled using the same behavior graph controller provided in SEAN 2.0 [273], and the robot used the same navigation logic across the tasks.

In each task, the robot guided the participant to the destination and repeatedly changed its behavior (as further detailed below). Importantly, the interaction was paused before and after each behavior change took place, at which point the participant was asked to evaluate the robot's most recent navigation performance. A typical data collection session was completed in 45 minutes to 1 hour. Participants were compensated US$15 for their time.

Robot Behaviors: During a navigation task, the robot switched between one of these three types of behavior:

1. **Nav-Stack.** The robot navigated efficiently to the destination based on the path planned by the ROS Navigation Stack with social costs [162]. The planned paths generally minimized navigation time while avoiding collisions and invading personal space. This behavior lasted 40 seconds.

2. **Spinning.** The robot rotated at its current position, which we expected to be perceived as if the robot was confused. This behavior lasted 20 seconds. It was implemented by sending angular velocity commands to the robot's motion controller.

3. **Wrong-Way.** The robot moved in the wrong direction, away from the task's destination, effectively making a mistake during navigation. This behavior lasted 20 seconds and was implemented using the Navigation Stack with social costs as well, but with an incorrect navigation goal.

Unbeknownst to the participants, the robot switched to *Nav-Stack* behavior after *Spinning* or *Wrong-Way* during navigation. It randomly switched to *Spinning* or *Wrong-Way* after finishing *Nav-Stack*. The design was intended to maintain a consistent rate of sub-optimal behavior and avoid user boredom or significant confusion, which can be caused by more stochastic behavior patterns that are hard for participants to reason about. We expected the behaviors to elicit both positive and negative views of the robot, leading to a large variety of non-verbal reactions and perceptions of robot performance.

Perceptions of Robot Performance: During a navigation task, we paused the interaction at 4 seconds *before*, and at 8 seconds *after* the robot switched between behaviors. The elapsed time for the latter pause was longer in order to give people enough time to experience the latest robot behavior.

As shown in the supplementary video, perceptions of robot performance were provided through an interface embedded in the simulation. The interface asked the

participants to indicate their perceptions about the robot's most recent performance in regard to: 1) *"how competent was the robot at navigating,"* 2) *"how surprising was the robot's navigation behavior,"* and 3) *"how clear were the robot's intentions during navigation."* Participants provided ratings for these three dimensions of robot performance on a 5-point Likert responding format, e.g., with 1 being "incompetent", 2 being "somewhat incompetent", 3 being "neither competent nor incompetent", 4 being "somewhat competent", and 5 being "competent".

8.4.3 Observations

We organized observations of human-robot interactions, as recorded in SEAN-VR [308], into the features described below. More details about these features are provided in the Appendix.

Participants' Facial Expression Features: We captured the participants' eye and lip movements, as well as their gaze through the VR headset using the VIVE Eye and Facial Tracking (SRanipal) SDK. The eye and lip movements corresponded to 73 features that described the geometry of the face through blend shapes. The gaze was a 3D vector providing the direction of gaze of the person relative to their face.

Spatial Behavior Features: During navigation, we captured the poses of the robot, the participant, and the other automatically-controlled avatars on the ground plane of the scene. Then, we computed the poses of the avatars relative to the robot, considering only those within a 7.2m radius, as this region is typically considered a robot's public space [101, 233, 121]. Each of the features were (x, y, θ) tuples with x, y being the position and θ the body orientation (yaw angle) relative to a coordinate frame attached to the robot.

Goal Features: A navigation task had an associated destination or goal that the

robot had to reach. We converted the goal pose in a global frame in the warehouse to a pose in a coordinate frame attached to the robot. This pose described the robot's proximity and relative orientation to its destination.

Occupancy Features: During navigation, the robot localized [96] against a 2-Dimensional (2D) map of the warehouse. We used a cropped section of the map around the robot (of 7.2m $\times$ 7.2m) to describe the occupancy of nearby space by static objects.

8.4.4 Perceived Robot Performance

Perceptions of robot performance were as expected: ratings for competence and clear intention were generally higher for *Nav-Stack* than for *Spinning* and *Wrong-Way*, while the latter two tended to be more surprising than the former. Pairs of performance dimensions were significantly correlated with absolute Pearson r-values greater than 0.6. An exploratory factor analysis suggested that the dimensions could be combined into one performance factor (which explained 77% of the variance).

Using the features described before and the perceptions of robot performance provided by the participants, we created a dataset of paired observation sequences and target performance values. We further refer to this data as the SEAN virTual rObot GuidE with impliciT Human fEedback and peRformance Dataset (SEAN TOGETHER Dataset). As described below, we used this dataset to investigate the research questions in Section 8.3.

8.4.5 How Well Can Human Observers Predict a User's Perceptions of Robot Performance?

To better understand the complexity of inferring perceptions of robot performance, we evaluated how well human annotators could solve the prediction problem. To this

end, we administered an online survey through Prolific,[†] a platform for human data collection and online research studies. In our survey, human annotators observed visualizations of observations in our SEAN TOGETHER Dataset. Then, they tried to predict performance ratings provided by the people who followed the robot.

Method: For the survey, we randomly selected 2 data samples from each of the 60 participants in our data collection, with one gathered before and the other gathered after the robot's behavior changed. The observations in each sample corresponded to an 8-second 5-hz window of features right before the corresponding performance label was provided.

As shown in Fig. 8.3, data samples were visualized in two ways:

1. **Facial Rendering.** We created a human face rendering in Unity by replaying the facial expression features on an SRanipal compatible avatar, as shown in Fig. 8.3 (right). This visualization was motivated by the use of facial expressions in prior work on implicit feedback (e.g., [71]).

2. **Navigation Rendering.** We created a plot of features that described the navigation behavior of the robot and the avatars in the simulation. The plot showed features that, using existing perception techniques, may be easier to estimate than facial features in real-world deployments. These features are the spatial behavior features, the robot's goal location, the occupied space near the robot, and the gaze direction of the participant – the last of which could be approximated using an estimate of the person's head orientation [197]. Because prior work suggests that it is easier to make sense of implicit human feedback in context [46], the plot was always centered on the robot, making its surroundings always visible as in Fig. 8.3 (left).

We used the visualizations to create three annotation conditions that helped understand the value of different features: 1) ***Nav.-Only***: annotators only saw the

[†]`www.prolific.co`

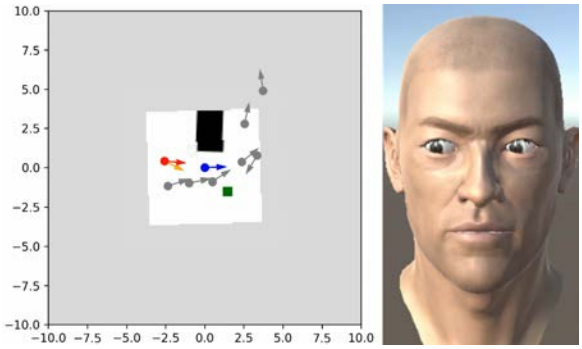


Figure 8.3: A data sample from the *Nav.+Facial* condition. The **left** plot shows gaze, spatial behavior, goal, and occupancy features: ●→ is the robot's pose; ●→ is the pose of the participant following the robot during the VR interaction; → indicates the gaze of the participant; ●→ are the poses of algorithmically controlled avatars; ■ is the destination position that the robot navigated towards; and occupancy in the environment is indicated by black pixels (occupied) and white pixels (unoccupied). The **right** visualization shows a rendering of the facial expression features of the participant.

navigation rendering (e.g., as in the left image of Fig. 8.4), and then completed the annotation; 2) ***Facial-Only***: annotators only saw the facial rendering (e.g., as in the right image of Fig. 8.4), and then completed the annotation; and 3) ***Nav.+Facial***: annotators saw the navigation rendering in the first page, then the facial rendering in the second page, and finally, saw a video with both visualizations next to each other (as in Fig. 8.3) in the last page and completed the annotation.

Each of the data samples was annotated by 10 unique people in each condition. The annotators were instructed to predict how the participant who controlled the avatar that followed the robot perceived the robot's performance. The samples they annotated were presented in random order. Each annotator was paid US$7.50 for approximately 30 min of annotation time. To encourage high-quality annotations, we also gave them a bonus of US$0.125 for each correct prediction that they made.

Annotators: We recruited a total of 100 annotators. Thirty-five of them identified as female, 60 as male, and 5 as non-binary or third gender. Ages ranged from 18 to 75 years old. Annotators indicated similar familiarity with robots (M = 4.12, SE = 0.14) as the data collection participants, though the annotators were slightly more familiar with VR (M = 4.50, SE = 0.16). See the Appendix for details on annotator

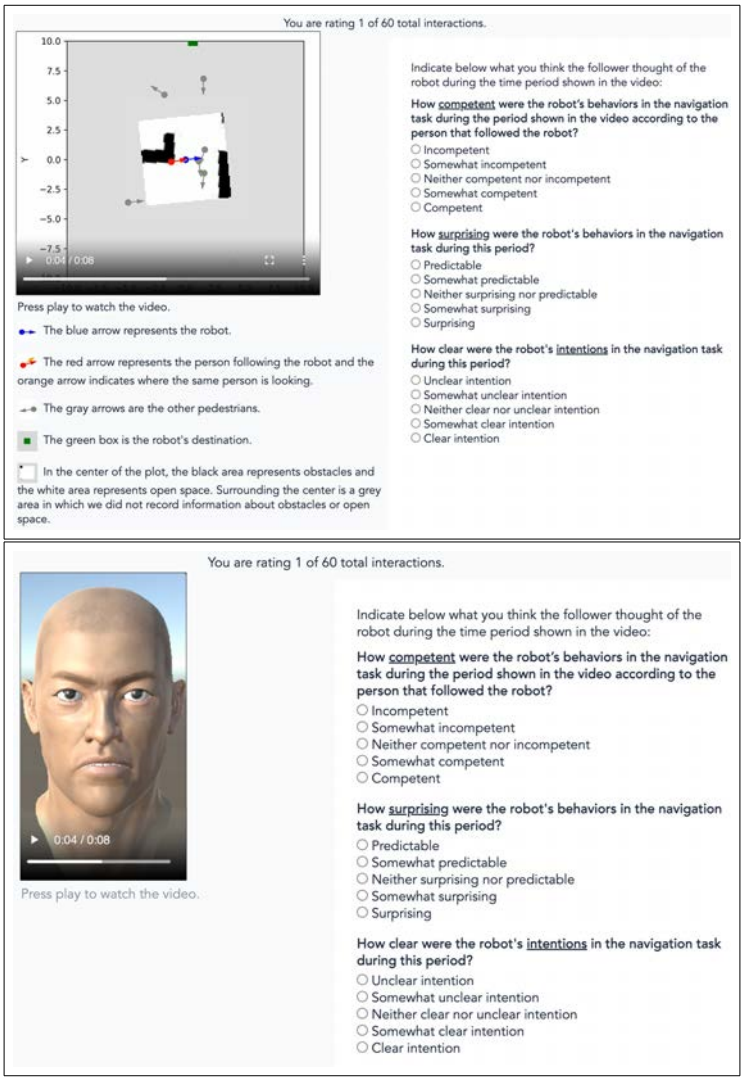


Figure 8.4: Layout of the interfaces used for video annotation for the human baseline. *Top:* Layout used for the *Nav.-Only* annotation condition, showing the navigation rendering on the left, and questions on the right. *Bottom:* Layout for the *Facial.-Only* condition.

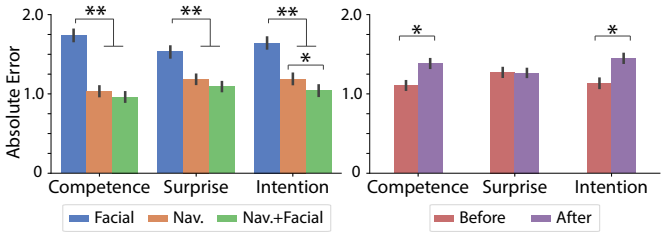


Figure 8.5: Errors for annotators' predictions by Annotation Conditions (*left*) and Before/After Robot Behavior Change (*right*). (**) and (*) denote $p < 0.0001$ and $p < 0.05$, respectively.

reliability.

Results: We used linear mixed models estimated with REstricted Maximum Likelihood (REML) [199, 242] to analyze errors in the predictions for each performance dimension. Our independent variables were Before/After Robot Behavior Change (*Before*, *After*) and Annotation Condition (*Facial-Only*, *Nav.-Only*, *Nav.+Facial*). Also, we considered Annotator ID as a random effect because annotators provided predictions for multiple data samples. Our dependent variables were the absolute error between an annotator's prediction and the performance rating in our SEAN TOGETHER Dataset.

We found that the Annotation Condition had a significant effect on the absolute error for Competence, Surprise, and Intention ($p < 0.0001$ in all cases). As in Figure 8.5 (left), Tukey HSD post-hoc tests showed that for Competence and Surprise, the errors for *Nav.+Facial* and *Nav.-Only* were significantly lower than *Facial-Only*, yet the difference between the former two conditions was not significant. For Intention, all conditions led to significantly different errors. *Nav.+Facial* resulted in the lowest error, followed by *Nav.-Only* and then *Facial-Only*. These results suggest that facial expressions provide information about perceptions of robot performance although, more generally, the features used to create the Navigation Renderings seemed to be the most critical for these predictions.

Before/After Robot Behavior Change had a significant effect on the prediction errors for Competence and Intention ($p < 0.0001$ in both cases). As in Figure 8.5

Table 8.1: Machine learning methods and human annotation (HA) performance predicting Competence, Surprise, and Intention. Methods: Random (R) sampling from the distribution of labels in the training set, Random Forest (RF), Multi-Layer Perceptron (MLP), Graph Neural Network (GNN), and Transformer (T). Arrows indicate that higher (↑) and lower (↓) results are better. Cells with (-) do not have results because a GNN trained on facial features only was effectively an MLP. The Best and Second results are highlighted.

		F_1-Score ($\mu \pm \sigma$) ↑			Accuracy ($\mu \pm \sigma$) ↑			Mean Absolute Error ($\mu \pm \sigma$) ↓		
		Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial
Competence	HA	0.16 ± 0.0	0.28 ± 0.1	0.29 ± 0.1	0.19 ± 0.1	0.40 ± 0.1	0.42 ± 0.1	1.74 ± 0.2	1.03 ± 0.3	0.99 ± 0.4
	R	0.18 ± 0.0	0.19 ± 0.0	0.17 ± 0.0	0.21 ± 0.0	0.21 ± 0.0	0.20 ± 0.0	1.73 ± 0.1	1.75 ± 0.1	1.81 ± 0.1
	RF	0.19 ± 0.0	0.37 ± 0.0	0.38 ± 0.0	0.33 ± 0.0	0.52 ± 0.0	0.52 ± 0.0	1.43 ± 0.0	0.88 ± 0.0	0.82 ± 0.0
	MLP	0.23 ± 0.0	0.29 ± 0.1	0.25 ± 0.1	0.28 ± 0.0	0.48 ± 0.0	0.44 ± 0.1	1.66 ± 0.1	1.07 ± 0.3	1.19 ± 0.4
	GNN	-	0.31 ± 0.1	0.33 ± 0.0	-	0.43 ± 0.1	0.39 ± 0.1	-	1.22 ± 0.3	1.04 ± 0.0
	T	0.21 ± 0.0	0.33 ± 0.0	0.33 ± 0.0	0.30 ± 0.0	0.43 ± 0.0	0.41 ± 0.1	1.58 ± 0.1	0.97 ± 0.0	0.95 ± 0.0
Surprise	HA	0.18 ± 0.0	0.24 ± 0.1	0.25 ± 0.1	0.20 ± 0.1	0.30 ± 0.1	0.32 ± 0.1	1.53 ± 0.3	1.19 ± 0.2	1.12 ± 0.2
	R	0.19 ± 0.0	0.21 ± 0.0	0.17 ± 0.0	0.20 ± 0.0	0.21 ± 0.0	0.18 ± 0.0	1.64 ± 0.1	1.60 ± 0.1	1.68 ± 0.1
	RF	0.29 ± 0.0	0.38 ± 0.0	0.34 ± 0.0	0.30 ± 0.0	0.40 ± 0.0	0.34 ± 0.0	1.30 ± 0.0	0.93 ± 0.0	0.98 ± 0.0
	MLP	0.24 ± 0.0	0.26 ± 0.1	0.24 ± 0.1	0.25 ± 0.0	0.30 ± 0.0	0.29 ± 0.1	1.23 ± 0.1	1.12 ± 0.2	1.08 ± 0.1
	GNN	-	0.29 ± 0.0	0.27 ± 0.0	-	0.30 ± 0.0	0.28 ± 0.0	-	1.13 ± 0.1	1.07 ± 0.1
	T	0.27 ± 0.0	0.29 ± 0.0	0.32 ± 0.1	0.28 ± 0.0	0.31 ± 0.0	0.33 ± 0.1	1.37 ± 0.1	1.07 ± 0.1	1.04 ± 0.1
Intention	HA	0.18 ± 0.0	0.25 ± 0.1	0.30 ± 0.1	0.21 ± 0.1	0.37 ± 0.2	0.41 ± 0.1	1.64 ± 0.2	1.19 ± 0.4	1.07 ± 0.2
	R	0.21 ± 0.1	0.19 ± 0.0	0.17 ± 0.0	0.23 ± 0.1	0.22 ± 0.0	0.19 ± 0.0	1.70 ± 0.1	1.73 ± 0.1	1.80 ± 0.1
	RF	0.28 ± 0.0	0.28 ± 0.0	0.24 ± 0.0	0.37 ± 0.0	0.43 ± 0.0	0.41 ± 0.0	1.45 ± 0.0	1.13 ± 0.0	1.14 ± 0.0
	MLP	0.27 ± 0.0	0.26 ± 0.1	0.22 ± 0.0	0.31 ± 0.0	0.41 ± 0.1	0.39 ± 0.1	1.86 ± 0.1	1.31 ± 0.3	1.51 ± 0.5
	GNN	-	0.28 ± 0.0	0.29 ± 0.0	-	0.37 ± 0.0	0.35 ± 0.0	-	1.32 ± 0.1	1.25 ± 0.1
	T	0.24 ± 0.0	0.29 ± 0.1	0.32 ± 0.0	0.33 ± 0.0	0.41 ± 0.0	0.40 ± 0.0	1.63 ± 0.1	1.21 ± 0.1	1.20 ± 0.1

(right), the error was significantly lower for samples *Before* a behavior change than for samples *After* a change for these performance dimensions. We suspect this was because the robot sometimes demonstrated two behaviors in the samples collected *After* a behavior change, but in the case of *Before* behavior change, the robot only showed one behavior, making these data samples more consistent and easier to reason about.

Table 8.1 shows the F_1-Scores for the annotator predictions (see HA rows). The low F_1 scores suggest that correctly predicting perceptions of robot performance on a 5-point responding format was difficult for humans. Despite this, we suspected that humans could do a more reasonable job at distinguishing perceptions of poor robot performance from other perceptions. If this was the case, then this could open up doors in the future to using this binary signal (instead of the more fine-grained feedback) as a reward signal to adapt robot behavior in navigation tasks, e.g., in line with [140, 165]. Thus, we transformed the ground truth ratings from

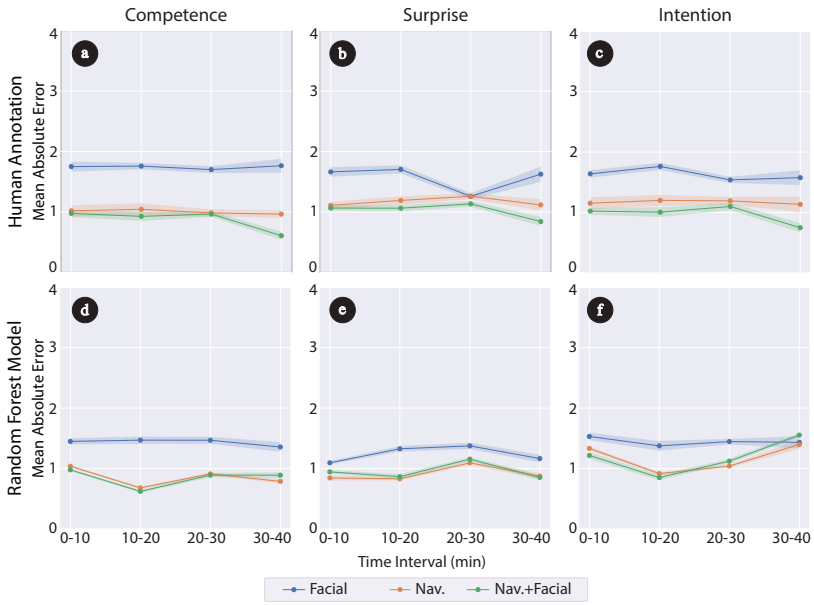


Figure 8.6: Mean Absolute Errors (MAE) of human annotation and Random Forest (RF) results over 10-minute intervals of the data collection sessions. MAE was computed for all data samples in each interval, and then the average and standard errors of MAE were calculated considering the performance of the 10 unique annotators (for human annotation results in (a)–(c)) or the 10 Random Forest models trained with different seeds in Table 8.1 (RF results in (d)–(f)).

our data collection to binary values, one corresponding to low performance (e.g., 1-2 ratings for competence) and another to medium-to-high performance (3-5 ratings for competence). Also, we transformed the annotators' predictions similarly. This led to F_1 scores of 0.69 for Competence, 0.64 for Surprise, and 0.69 for Intention. As expected, human annotators were better at telling the directionality of robot performance ratings than at predicting their exact magnitude.

Finally, we investigated the performance of human annotations over the span of data collection because prior work suggests that the expressiveness of people engaged in human-robot interactions can change over time [47], e.g., potentially due to changes in their expectations about the robot or due to fatigue. Figures 8.6(a)–(c) show the evolution of mean absolute errors for the human annotators' predictions over 10-minute intervals of interaction, considering each performance dimension. In general, human performance was very stable, suggesting no major bias over time

in participants' spatial behavior or facial expressions. Interestingly, the results also suggested that improvements in performance with an individual feature did not necessarily translate to improvements on the *Nav.+Facial* condition. Humans may have combined the information from the different implicit feedback modalities in subtle ways when making their predictions about how participants in VR perceived the robot.

8.4.6 Can Machine Learning Methods Predict Perceptions of Robot Performance as Well as Humans?

We compared human prediction performance with a variety of classifiers, including a random forest and neural networks.

Method: Machine learning (ML) models were evaluated on the same samples shown to the human annotators ($n = 120$). The rest of the data was used for training ($n = 2280$) and validation ($n = 569$). We trained one model for each combination of feature sets shown to the human annotators (*Facial-Only*, *Nav.-Only*, and *Nav.+Facial*). The *Nav.* feature set included occupied space near the robot, which we encoded using a ResNet-18 representation [103]. We repeated training for each model 10 times with varying random seeds. The Random Forest (RF) used 100 trees, and the depth was grown until leaves had less than 2 samples. The neural networks had a number of parameters on the same order of magnitude: 5.4×10^6 for a Multi-Layer Perceptron (MLP), 2.1×10^6 for a message-passing Graph Neural Network (GNN) [27], and 6.5×10^6 for a Transformer (T) [279]. Networks were trained using minibatch gradient descent with the Adam optimizer and cross-entropy loss. Learning rate, batch size, and dropout were chosen using grid search with validation-based early stopping [210]. We also compared all these models with a random sampling baseline.

Results: As is shown in Table 8.1, ML models outperformed both human-level performance and random baseline in all cases when measured via F_1-Score. When measured

using Accuracy and Mean Absolute Error, ML models performed the best, except for Intention when using *Nav.+Facial* features. These outcomes indicate that our implicit feedback data contained useful information that can be leveraged by ML models to predict users' perceptions of robot performance. Further, ML models trained with *Nav.-Only* and *Nav.+Facial* features outperformed those trained with *Facial-Only* features. This finding aligns with our observation in Section 8.4.5 on the criticality of the *Nav.* features in comparison to the *Facial* features on performance prediction.

Figures 8.6(d)–(f) show the evolution of mean absolute errors for the Random Forest model, which generally performed the best, over 10-minute intervals of interaction during the data collection. Similar to the results from human annotators (Figures 8.6(a)–(c), Sec. 8.4.5), the error for the RF model did not fluctuate drastically, although the performance for Intention prediction with *Nav.* and *Nav.+Facial* features decreased in the last two time intervals of data collection (having higher mean absolute error). The decrease in performance could be the result of a distribution shift, especially in the last interval, which had the fewest number of samples because not all interactions took the full 40 minutes. Also, a good proportion of the samples in the last time interval showed the end of navigation tasks, at which point the participants could have been more sensitive to robot navigation in the wrong direction. Indeed, there was a higher proportion of lower ratings for Intention in the last interval than in the other intervals, as shown in the Appendix.

To better understand differences in the prediction performance between ML and human annotators, we first identified the examples annotated by humans for which there was a difference greater than 1 in Mean Absolute Error between human annotators and the RF model that tended to perform best. Then, we inspected the 8-second navigation renderings of these data examples, as in Fig. 8.4 (left). Among examples where the RF model performed better than humans, 64% exhibited a major behavior pattern for the robot that persisted despite minor deviations. For example,

the robot navigated effectively to the goal most of the time, but was occasionally blocked and had to move around the obstacles. We hypothesize that ML did better in these cases because machine learning can leverage regularities in the data when making predictions without potentially getting distracted with the minor deviations. Among the examples where human annotators performed better, 68% showed the robot exhibiting more than one behavior (*Nav-Stack*, *Spinning*, or *Wrong-Way*) or the interaction involved unconventional reactions from humans, such as people interfering in the navigation task. We suspect that humans were better in these cases because they can leverage their prior knowledge about the world to better reason about uncommon variations in the data. For the RF, uncommon observations can be out-of-distribution samples that result in more prediction errors, especially considering the limited size of our dataset.

Taken together, these results motivated us to focus the analysis in the next section on the aggregate, overall results rather than the interval-based results.

8.4.7 Can Machine Learning Generalize to Unseen Users?

We investigated how well learning models could predict performance by a user whose data was held out from training.

Method: We used the models and training scheme from Section 8.4.6 with all fea-

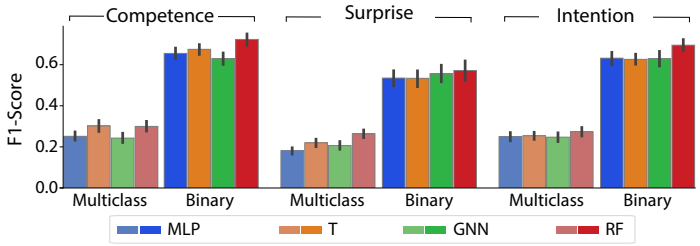


Figure 8.7: ML models trained on *Nav.+Facial* features using leave-one-out cross-validation and evaluated on the held-out participant's data. F_1-Scores are computed over 5 classes (Multiclass) and 2 classes (Binary). Error bars represent the standard errors calculated from the F_1-Scores per leave-one-out fold. See the text for details.

tures (*Nav.+Facial*), but split the data using leave-one-out cross-validation. For each fold, the data for one participant was used as the test set, and the remaining examples were split between training (80%) and validation (20%). We searched for new hyperparameters and computed results both on 5-classes and on binary classification. Binary targets and prediction labels were computed as in Section 8.4.5.

Results: Figure 8.7 reports F_1-Scores over all folds. The models generalized to unseen people with only a slight reduction in performance in comparison to Table 8.1. Also, the average F_1-Score across all performance dimensions improves from 0.25 in the multiclass case to 0.62 in the binary case. This makes the ML predictions more usable in practice. For example, in the future, we envision deploying the trained ML on new users (as in Fig. 8.2b) in order to detect low robot performance. This could be an indication that the robot made a mistake, triggering interaction recovery behaviors like apologies or explanations [257], which could increase trust in the system [54].

8.5 Real-World Demonstration

To investigate whether we could predict human perceptions of robot performance in other, more realistic scenarios than those observed in our VR data collection, we conducted a real-world demonstration with a modified Pioneer 3-DX mobile base. More specifically, we conducted a data collection with the mobile robot in two semi-public indoor environments of Yale University, and analyzed how well a random forest model could predict human perceptions of robot performance in the real-world setup. This real-world data collection, as further described below, was approved by our local Institutional Review Board.

The system that we built for real-world data collection was designed in consideration of: 1) we wanted to induce naturalistic interactions between the robot and pedestrians; and 2) we wanted to support the same data collection protocol used

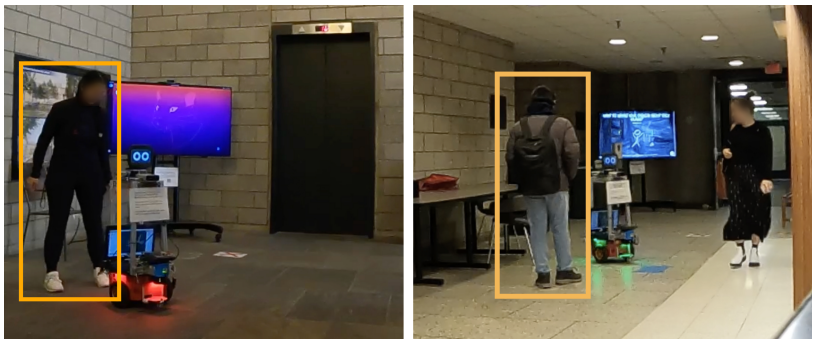


Figure 8.8: Real-world data collection in two indoor spaces of Yale University. The orange box highlights the follower, i.e., the person that followed the robot during navigation tasks. Other people could pass by the follower and the robot as in the *right* image during data collection. The robot had lights to indicate when it was navigating (green, *right* image) or had paused navigation (red, *left* image).

with SEAN, as in Section 8.4.2. Therefore, we did not recruit participants prior to the data collection. Instead, we operated the robot and, as pedestrians walked nearby, we asked them if they would be willing to follow the robot for a short period and answer brief surveys. In total, 45 pedestrians agreed to follow the robot for this demonstration.

Mobile Robot: The Pioneer 3-DX robot is a differential-drive mobile base and, thus, it moves in a similar way to the Fetch robot used in our VR data collection. We added to the Pioneer robot lights that illuminated green to indicate that it was navigating towards a location, and red to indicate that it had paused navigation. Over the Pioneer base, we built a frame that held a robotic screen face (similar to [280, 154]) on the very top of the robot, which allowed participants to easily distinguish the front of the platform. The frame also held two Kinect Azure RGB-D cameras right below the robot's head. Each camera had a 120-degree field of view. One was pointed forward and the other was pointed backwards, which allowed the robot to track people in front and behind it using the Kinect SDK. Additionally, the bottom section of the frame held a 2D LMS-100 Sick LiDAR and a gaming laptop with an Intel Core i7-8750H CPU, 32 GiB of RAM, and an Nvidia GeForce GTX 1070 GPU. The laptop ran the Robot Operating System to control the robot using

the ROS navigation stack [212] with social cost layers [162], which enabled the robot to avoid collisions with nearby people. Fig. 8.8 and our supplementary video show the robot in this demonstration effort.

Demonstration Protocol: We waited for pedestrians to walk by the robot in two locations on a university campus. One location was a subterranean pedestrian tunnel or concourse; the other one was an L-shaped entrance corridor to a building. When pedestrians passed by, we asked them if they would be interested in following the robot as it navigated to a nearby goal marked by a red cross on the ground. For those who agreed, we instructed them that the robot would navigate when it showed a green light. After short intervals of time, it would pause navigation, showing a red light, and they would be asked a few quick questions about their perceptions of the action that the robot just performed using a mobile device. The device showed the same questions about robot competence, surprising behavior, and clear intent (on a 5-point Likert responding format) as in our VR data collection. Also, the robot navigation behaviors and the timing of questions about robot performance matched those in Section 8.4.2.

Data: We focused on capturing *Nav.-Only* features (that described the navigation behavior of the robot and humans, as in Sec. 8.4.5) for two reasons. First, our prior results with VR data suggested facial expression features were not as critical to make predictions over human perceptions of robot performance as the other features. Second, facial expressions were often occluded, providing no information to the robot. In total, we collected 235 examples from this real-world demonstration, each consisting of *Nav.-Only* features and associated survey responses.[‡]

ML Models: Our primary aim was to understand the applicability of our approach to infer perceptions of robot performance in the real world. However, there were important differences in our VR and real-world data collection setups as a result of

[‡]The real-world data that we collected from this demonstration is available at: `https://sean-together.interactive-machines.com/`.

Table 8.2: F_1-Score ($\mu \pm \sigma$) for Random Forest models trained using *Nav.-Only* features from either the *Real*-world data, or *VR* data considering the nearest 5 people to the robot (as explained in Sec. 8.5) Results include multi-class classification based on the 5-point Likert responses (*Multi-cls*) and binary classification (*Binary*). Column 1 corresponds to training on VR data and evaluating on VR data (VR→VR), Column 2 corresponds to training on VR data and evaluating on real data (VR→Real), and Column 3 is training and evaluating on real data (Real→Real).

		(1) VR→VR	(2) VR→Real	(3) Real→Real
Multi-cls	Competence	0.30 ± 0.09	0.21 ± 0.18	0.27 ± 0.35
	Surprise	0.27 ± 0.08	0.26 ± 0.21	0.26 ± 0.27
	Intention	0.26 ± 0.08	0.20 ± 0.28	0.24 ± 0.34
Binary	Competence	0.69 ± 0.10	0.56 ± 0.41	0.61 ± 0.34
	Surprise	0.59 ± 0.18	0.58 ± 0.36	0.58 ± 0.33
	Intention	0.65 ± 0.08	0.55 ± 0.40	0.60 ± 0.40

real-world constraints. For example, the real robot had a more limited field of view compared to the simulation, where the ground truth motion for all people in the environment was available. Moreover, the real-world environments were less densely populated than simulation.

Therefore, to fairly compare our results across simulation and the real world, we trained two types of Random Forest classifiers, given that the RF model generally performed best in Table 8.1. One type of model was trained using VR data but we limited the field of view of the robot to 120-degrees forward and backward, as well as the maximum number of nearby people input to the model to five individuals. The other type of Random Forest model (with the same parameters) was trained using real-world data. Both types of models were trained considering 5-classes, with binary targets and prediction tables being computed as in Section 8.4.5.

Results: Table 8.2 shows the F_1-Score of models evaluated on the same type of data they were trained on (Sim or Real). For these results, we used leave-one-person-out cross-validation to train and evaluate generalization to new robot followers. That is, data from one person was held out for each fold. Also, Table 8.2 shows the performance of the model trained in simulation on real-world data. In this case, an RF

model was trained using all the VR data from the VR→VR case, and then evaluated on the test set for the leave-one-person-out folds for the real-world data. As one would naturally expect based on our prior results with VR data, binary classification resulted in higher performance than multi-class classification in all these cases.

In general, performance was higher for models trained and evaluated in simulation (Column 1), which could be the result of having more VR data than real-world data. The results for models trained and evaluated on real data (Column 3) were close to those that considered simulation data only (Column 1). This suggests that our methodology to collect real-world data and the RF model are promising for inferring perceptions of robot performance in the real world. Finally, reasonable performance was obtained for the model that was trained with VR data and tested on real-world data (Column 2). This highlights the potential of sim-to-real transfer of machine learning models trained on spatial features, as well as the potential of using our VR data to build computational models that predict human perceptions of robot performance in real-world interactions.

8.6 Discussion

We hope that future work leverages our findings to build effective models for mapping implicit human feedback to users' perceptions of robot performance in real-world social navigation tasks. To this end, we first recommend prioritizing robust people tracking and pose estimation over computing fine-grained facial expressions, especially when computational resources may be limited. Reasoning about spatial behavior features in the context of the task can facilitate achieving reasonable prediction performance with lower sensor requirements. Also, occlusions are likely more common for facial expressions than body tracking, as we observed in our real-world demonstration.

Second, it is important to consider the granularity of the predictions over perceptions of robot performance. We began our work gathering perceptions of robot performance on a 5-point Likert responding format, which we believed could reveal subtle aspects of human perceptions during navigation. However, we found that predicting perceptions of robot performance over 5 classes was challenging for both humans and ML models. While human prediction performance could have been affected by specific details of the visualizations that we used to gather our human baseline results, it is worth considering less granular feedback to favor prediction performance during robot deployments. In particular, for more practical usage of human feedback, we recommend building models that start by identifying poor robot performance (performing binary classification) and then, on top of that, try to predict more granular perceptions of robot performance.

Finally, if a robot is executing multiple behaviors, we recommend considering whether the robot switched behaviors recently when reasoning about performance predictions. As in our results, predicting performance recently after a behavior change can be more difficult than before, when the behavior was more consistent.

8.6.1 Limitations

Our work has several limitations that point to interesting future directions. In particular, we obtained human baselines for prediction performance, but used only a limited set of feature combinations that described interactions in a single VR environment and two real-world environments. In the future, it would be interesting to consider a broader set of feature categories in a more diverse range of environments. For instance, future work could investigate the value of more detailed human pose features (e.g., [310]) across a wider range of scenarios (public plazas or hospitals) where humans may behave differently due to their activity, stress, or other factors.

Facial expressions and the nuance of human motion are challenging to capture.

In our data collection with virtual reality, the use of VR could have biased observed nonverbal behavior as well as human perceptions of a robot, given the way humans provide input to the simulation via the VR device, the way the device captures their nonverbal behavior, and the sim-to-real gap.

We were limited by the features captured by the Vive Pro Eye VR headset, which describe the geometry of the face through blend shapes. We visualized this data by rendering the features on a virtual avatar head, and this could have affected the perception of subtle human facial expressions. In the future, it would be interesting to utilize more advanced devices such as the recently released Apple Vision Pro to create other datasets of implicit human feedback. The new Apple device can sense faces in a way that allows rendering higher quality avatars for users, and the data it captures could potentially improve the accuracy and robustness of ML models that predict robot performance.

In the future, inferred performance predictions could be used to adapt robot behavior. For example, a robot could use binary robot performance predictions as instantaneous rewards that guide changes in robot behavior to better align what the robot does with human preferences [140, 165, 71]. When the predictions indicate low robot performance or suggest drastic changes in perceptions of the robot's behavior, the robot could also opt for querying users explicitly about its performance to verify the predictions. Perhaps the responses can also be used to improve the prediction model.

8.7 Summary

This chapter contributed the SEAN TOGETHER Dataset, consisting of observations of human-robot interactions in VR, including implicit human feedback, and corresponding performance ratings in guided robot navigation tasks. Our analyses with

VR data revealed that facial expressions can help predict perceptions of the robot, but spatial behavior features in the context of the navigation task were more critical for these inferences. Our experiments also demonstrated the ability of humans and ML models to infer perceived robot performance from interaction observations. A general trend that we observed throughout this work was that predicting the directionality of perceptions of robot performance (as a binary classification task) was easier and, thus, seemed more practical than predicting exact performance ratings (on a 5-point scale).

As part of this work, we also conducted a real-world demonstration that showed the applicability of machine learning in predicting human perceptions of a mobile robot in indoor environments. We did not capture facial expression features for this demonstration, but rather focused on capturing features that described the navigation behavior of the robot and humans based on our prior findings. Both the models trained with VR data and real-world data showed promising generalization capabilities when evaluated on real-world data, confirming the potential of machine learning for predicting perceptions of robot performance from implicit feedback signals in social robot navigation. Our datasets, accompanying analyses, and demonstration facilitate future research on more scalable supervision of robot navigation behavior.

In the future, social robots could use implicit human feedback as supervision to interactively improve their behavior in the future. For example, human perceptions of robot performance predicted by machine learning models could be used as a reward function in a reinforcement-learning setup, where the robot improves its policy to learn how to best navigate with a user.

Chapter 9

Discussion

This dissertation explores the exciting and critical application area of social robot navigation, highlighting the limitations of traditional, purely objective metrics in capturing the nuances of socially competent robot behavior. A paradigm shift in how success is measured is necessary in order to better align the evaluation of social robot navigation systems with human values. To this end, I propose the use of context-aware simulation systems and subjective human feedback. Equipped with these tools, researchers should utilize a cyclical method of system development that relies on repeated measurement and improvement.

Motivated by the "tyranny of metrics," we proposed a suite of systems for training and evaluating social robot navigation in a way that is aligned with human values. SEAN 2.0 is a simulation system designed specifically for social robot navigation, incorporating novel components such as modeling pedestrian motion via a Behavior Graph, formalizing social contexts, and classifying social situations. Furthermore, we investigated how experts in social robot navigation prioritize different evaluation measures, revealing the criticality of subjective human feedback. We also addressed the challenge of scalable data collection for human-robot interaction through web-based interactive simulations and explored methodologies for collecting and even predicting

human perceptions of robot performance during navigation tasks. Our work lays the foundation for future research aimed at deploying socially competent robots that navigate effectively in human-centric environments, ensuring outcomes align with human values.

9.1 Common Themes

9.1.1 Choosing Metrics for Alignment Between Humans and Robots

This dissertation consistently pushes for the need to rethink how to measure success and makes the case that traditional metrics like path efficiency, time to goal, and collision avoidance, while important, fail to capture the nuances of socially competent behavior. We associate this outcome with the "tyranny of metrics," where optimizing for improperly chosen objectives can lead to behaviors misaligned with human values. We conducted a study using structured interviews of experts, which further reinforces this by revealing that while collision avoidance is universally important, other objective measures are prioritized differently depending on the application domain, highlighting the insufficiency of objective measures alone.

The dissertation proposes that to create truly socially competent robots, it is essential to understand and incorporate how humans perceive robot behavior. This theme is evident in the proposed three-pronged approach, which explicitly includes incorporating subjective human feedback in a scalable manner. The expert interviews revealed the critical role of subjective human feedback in evaluating social navigation. Furthermore, the development and evaluation of the SEAN-EP system focuses on enabling the collection of this crucial subjective human feedback data.

9.1.2 Human-Centric Simulation

Another common theme is the role of human-centric simulation and the proposed simulation system, SEAN, as a vital tool for safely developing, testing, and evaluating social navigation algorithms. This dissertation introduces the preliminary SEAN and the complete SEAN 2.0 system as a tool that can help address the challenges of value alignment. These platforms are designed to model human behavior, formalize social contexts, and provide a framework for evaluating robot policies in various social situations. The development of the Behavior Graph in SEAN 2.0 as a novel method for specifying pedestrian behavior and generating varied social situations underscores the importance of densely populated and dynamic simulation environments filled with virtual agents.

9.1.3 Human Feedback and Predicting Perceptions of Robot Behavior

The dissertation also highlights the challenges and advancements in achieving scalable data collection of human feedback for human-robot interaction in navigation contexts. The development of SEAN-EP is a direct response to this challenge, offering a method to deploy interactive robot simulations on the web as part of interactive surveys to gather human feedback at scale. The investigation into whether human perceptions differ between interactive simulations and video observations further explores methodologies for efficient and effective human feedback collection.

Finally, the dissertation looks towards the future with the theme of predicting human perceptions of robot performance as a crucial step towards creating robots that are inherently aligned with human values. The research on using nonverbal human behavior (body motion, gaze, facial expressions) to predict how humans perceive a robot's navigation performance suggests a path towards autonomous systems that can

understand and respond to human preferences without explicit feedback. This ability to predict human perceptions is seen as a way to improve robot decision-making and mitigate the risks of optimizing for metrics that are misaligned with human values.

9.2 Open Challenges

We identify several open challenges that are critical for the development of socially competent robots, including the need for agreement among stakeholders in social robot navigation on a summary metric for success, the need for simulation systems that incorporate more degrees of freedom in human behavior, and the incorporation of human feedback into learned policies.

9.2.1 Summary Metric for Success

Unlike traditional robot navigation, where success can be readily quantified by metrics like path efficiency and collision avoidance, social navigation involves navigating in spaces shared with humans, where success is not solely about reaching a goal but also about doing so in a socially acceptable manner. This involves adhering to social norms, communicating intent clearly, and avoiding discomfort or harm to people. The multi-faceted nature of human-robot encounters contributes to this difficulty, as researchers must consider physical safety, psychological safety, social acceptability, and the interpretability of robot behaviors. Even seemingly straightforward terms like "safety" can have different interpretations depending on the context.

9.2.2 Simulation Systems that Incorporate More Degrees of Freedom in Human Behavior

Another key open challenge lies in the development of simulation systems that incorporate more degrees of freedom in human behavior. While simulation plays a vital

role in the development and evaluation of social robot navigation systems, current simulators often have limitations in the realism and variability of modeled human behaviors. Many existing simulation frameworks for social navigation focus on basic collision avoidance or simple, scripted pedestrian movements. However, real-world human behavior is complex and influenced by a multitude of factors, including individual goals, social norms, emotional states, and interactions with other people.

To develop truly socially competent robots, simulation systems need to move beyond these simplified models and incorporate a richer spectrum of human behaviors. This includes modeling more nuanced social interactions like group formations, conversations, yielding behaviors, non-verbal communication (e.g., gaze, gestures), and reactions to unexpected robot actions. The introduction of the Behavior Graph in SEAN 2.0 represents a step towards addressing this challenge by providing a new method for specifying pedestrian behavior that can lead to the emergence of more complex and varied social encounters.

Future research should work towards creating simulation systems that can accurately and efficiently model the full complexity and unpredictability of human behavior in diverse social contexts. For example, simulators that incorporate more degrees of freedom into human communication would allow nonverbal communication, such as gestures and gaze as well as verbal communication through the integration of large language models and text-to-speech models.

9.2.3 Incorporating Human Feedback into Learned Policies

A crucial open challenge for achieving value alignment in social robot navigation is the effective incorporation of human feedback into learned robot policies. This dissertation emphasizes that subjective human perceptions are critical for evaluating and developing socially aligned navigation strategies. Relying solely on objective metrics can lead to robot behaviors that are technically efficient, but socially inappropri-

ate. Therefore, integrating human feedback into the learning process is essential for ensuring that robot policies align with human values and expectations.

We explored several approaches for collecting human feedback, including in-person studies, video-based surveys, and interactive simulations. The development of SEAN-EP demonstrates an approach to gathering human feedback in a scalable manner through online interactive simulations. Furthermore, our research on predicting human perceptions of robot performance from nonverbal cues like body motion, gaze, and facial expressions offers a potential path towards autonomous systems that can infer human feedback implicitly.

Effectively leveraging human feedback to guide the learning of robot policies remains a formidable challenge. This includes identifying how best to represent human preferences and incorporate them into learned policies. For example, we recently proposed a method for incorporating user preferences into a planning framework [187]. Future work could build on this by using inferred preferences to inform policy learning within the planning framework. Another promising direction involves using human feedback to design reward functions that incorporate subjective feedback. In cases where these functions have gradients that are not useful for gradient-based optimization, we proposed a technique that approximates the Heaviside step function and uses a soft-set version of the confusion matrix to enable gradient-based optimization [272]. Finally, a more general characterization of social contexts, and of user values related to subjective human behavior, could help ensure that learned policies generalize across diverse situations. The long-term goal of this work is to create truly social and competent robots.

9.3 Summary

This chapter covered the dissertation's contributions and overarching themes. It emphasized the need to shift from purely objective metrics towards incorporating subjective human perceptions for evaluating social robots. We reviewed the key contributions and common themes in this work. Finally, the chapter identifies key open challenges for the field, namely the need for a summary metric for success, simulation systems with more degrees of freedom in human behavior, and methods for incorporating predictive models that are capable of predicting human feedback into the learning of future social robot navigation policies.

Driven in a small part by this dissertation, social robots are becoming better equipped to understand human perceptions and thereby become more closely aligned with human values. Bridging technical innovation with the nuanced realities of social behavior brings the field closer to a future in which robots not only move through our shared spaces but do so with a sensitivity that earns human respect and acceptance. The future for social robotics is bright, with exciting opportunities for innovation in human-centric design and robot learning that place people not just as users, but as essential collaborators in shaping socially intelligent systems.

Chapter 10

Conclusion

This dissertation, which focused on the development of socially competent mobile robots, proposed that traditional objective metrics alone are insufficient to capture the nuances of social interactions that occur during robot navigation. The research emphasized a need to move beyond objective metrics such as time to goal and collision avoidance. Beyond these metrics, human perceptions and values should be incorporated into the design and assessment of social navigation algorithms. By focusing on a human-centric approach, this work provides a foundation for creating robots that can navigate and interact effectively in human-centric environments, with the ultimate aim of aligning robotic behaviors with human values.

The results presented across this dissertation underscore the importance of considering human factors in social robot navigation. Chapter 3 introduced the design decisions associated with building a human-centric simulator and presented the Social Environment for Autonomous Navigation (SEAN) as a tool for developing and testing social navigation algorithms. Building upon this, Chapter 4 detailed SEAN 2.0, which formalized social situations and incorporated the Behavior Graph method of pedestrian control, which supports dense and varied pedestrian behaviors. In combination, experiments utilizing Social Situations and the Behavior Graph demonstrated

SEAN 2.0's utility for training and benchmarking. Chapter 5 provided insights from expert interviews, revealing that while collision avoidance is a near-universal priority, the importance of other evaluation measures varies across application domains. Open-ended questions highlighted the importance of incorporating subjective human feedback.

Further exploring the collection of human feedback, Chapter 6 presented the SEAN Experimental Platform (SEAN-EP), a novel approach for deploying interactive simulations by embedding them in surveys and deploying them on the web in order to gather human perceptions of robots at scale. Chapter 7 compared different methodologies for collecting human feedback, based on whether the interaction was real or simulated, and interactive or video-based. This study highlighted that human feedback from simulation and video studies may not always directly translate to real-world human-robot interactions, yet interactive simulations can be a powerful tool for rapidly gathering subjective measures.

Finally, in Chapter 8 we introduced the SEAN TOGETHER Dataset and demonstrated that machine learning models can infer how a robot is perceived by humans during navigation. The ability to predict human perceptions opens the door for autonomous systems that can adapt their behavior in real-time to improve their social performance based on the behaviors of nearby people.

In sum, the contributions of this dissertation point towards a future for robotics in which socially capable robots can navigate the human-centric environments we inhabit and interact socially with nearby people in a way that is aligned with their values.

Bibliography

[1] AI Habitat Challenge. `aihabitat.org/challenge/2020`. 2021-02-28.

[2] AWS DeepRacer. `aws.amazon.com/deepracer`. 2021-02-28.

[3] Crowdbot challenge. `crowdbot.eu/crowdbot-challenge`. 2021-02-28.

[4] Nvidia isaac sim. `developer.nvidia.com/isaac-sim`. 2021-02-28.

[5] David J Ahlgren and Igor M Verner. Socially responsible engineering education through assistive robotics projects: The robowaiter competition. *IJSR*, 2013.

[6] Neziha Akalin, Annica Kristoffersson, and Amy Loutfi. Do you feel safe with your robot? factors influencing perceived safety in human-robot interaction based on subjective and objective measures. *International journal of human-computer studies*, 158:102744, 2022.

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

[8] Philipp Althaus, Hiroshi Ishiguro, Takayuki Kanda, Takahiro Miyashita, and Henrik I Christensen. Navigation for human-robot interaction tasks. In *ICRA*, 2004.

[9] Alexander Amini, Igor Gilitschenski, Jacob Phillips, Julia Moseyko, Rohan Banerjee, Sertac Karaman, and Daniela Rus. Learning robust control policies

for end-to-end autonomous driving from data-driven simulation. *IEEE Robotics and Automation Letters*, 5(2):1143–1150, 2020.

[10] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[11] Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, pages 671–681. PMLR, 2021.

[12] Georgios Angelopoulos, Alessandra Rossi, Claudia Di Napoli, and Silvia Rossi. You are in my way: Non-verbal social cues for legible robot navigation behaviors. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 657–662. IEEE, 2022.

[13] Michael Argyle, Adrian Furnham, and Jean Ann Graham. *Social situations*. Cambridge University Press, 1981.

[14] Reuben M Aronson and Henny Admoni. Gaze for error detection during human-robot shared manipulation. In *Fundamentals of Joint Action workshop, Robotics: Science and Systems*, page 5, 2018.

[15] Anoop Aroor, Susan L Epstein, and Raj Korpan. Mengeros: A crowd simulation tool for autonomous robot navigation. In *AAAI 2017 Fall Symposium on Artificial Intelligence for Human-Robot Interaction*, 2017.

[16] Chatchalita Asavanant and Hiroyuki Umemuro. Personal space violation by a robot: An application of expectation violation theory in human-robot interaction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 1181–1188. IEEE, 2021.

[17] Eleanor Avrunin and Reid Simmons. Socially-appropriate approach paths using

human data. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1037–1042. IEEE, 2014.

[18] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[19] Wilma A. Bainbridge, Justin W. Hart, Elizabeth S. Kim, and Brian Scassellati. The benefits of interactions with physically present robots over video-displayed agents. *Int. J. Soc. Robot.*, 3(1), 2010.

[20] David P Baker and Eduardo Salas. Principles for measuring teamwork skills. *Human factors*, 34(4):469–475, 1992.

[21] Hussein Bakri, Colin Allison, Alan Miller, and Iain Oliver. Virtual worlds and the 3d web–time for convergence? In *iLRN*, 2016.

[22] Santosh Balajee Banisetty and Tom Williams. Implicit communication through social distancing: Can social navigation communicate social norms? In *HRI LBRs*, 2021.

[23] Santosh Balajee Banisetty and Tom Williams. Implicit communication through social distancing: Can social navigation communicate social norms? In *HRI '21 Companion*, 2021.

[24] Kimberly A. Barchard, Leiszle Lapping-Carr, R. Shane Westfall, Andrea Fink-Armold, Santosh Balajee Banisetty, and David Feil-Seifer. Measuring the perceived social intelligence of robots. *THRI*, 9(4), September 2020.

[25] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. *Human-robot interaction: An introduction*. Cambridge University Press, 2020.

[26] Len Bass, Bonnie Elizabeth John, and Jesse Kates. Achieving usability through

software architecture. In *International Conference On Software Engineering*, volume 23, pages 684–686. Citeseer, 2001.

[27] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

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

[29] Aniket Bera, Sujeong Kim, Tanmay Randhavane, Srihari Pratapa, and Dinesh Manocha. Glmp-realtime pedestrian path prediction using global and local movement patterns. In *2016 IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 5528–5535. IEEE, 2016.

[30] Aniket Bera, Tanmay Randhavane, Emily Kubin, Austin Wang, Kurt Gray, and Dinesh Manocha. The socially invisible robot navigation in the social world using robot entitativity. In *2018 ieee/rsj Intl. Conf. on intelligent robots and systems (iros)*, pages 4468–4475. IEEE, 2018.

[31] Aniket Bera, Tanmay Randhavane, and Dinesh Manocha. Improving socially-aware multi-channel human emotion prediction for robot navigation. In *CVPR Workshops*, pages 21–27, 2019.

[32] Homanga Bharadhwaj, Zihan Wang, Yoshua Bengio, and Liam Paull. A data-efficient framework for training and sim-to-real transfer of navigation policies. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 782–788. IEEE, 2019.

[33] Abhijat Biswas, Allan Wang, Gustavo Silvera, Aaron Steinfeld, and Henny Admoni. Socnavbench: A grounded simulation testing framework for evaluating

social navigation. *THRI*, 2021.

[34] Abhijat Biswas, Nathan Tsoi, Allan Wang, Peter Yu, Xiao He, Liyao Fu, Gustavo Silvera, Marynel Vazquez, and Aaron Steinfeld. SEANavBench @ ICRA 2023 workshop website. `https://seannavbench2022.netlify.app/benchmark/overview`, 2022. [Online; accessed 1-Oct-2024].

[35] Abhijat Biswas, Allan Wang, Gustavo Silvera, Aaron Steinfeld, and Henny Admoni. Socnavbench: A grounded simulation testing framework for evaluating social navigation. *ACM Trans. on Human-Robot Interaction (THRI)*, 11(3):1–24, 2022.

[36] Erdem Bıyık, Aditi Talati, and Dorsa Sadigh. Aprel: A library for active preference-based reward learning algorithms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 613–617. IEEE, 2022.

[37] Paula Boddington. Epsrc principles of robotics: commentary on safety, robots as products, and responsibility. *Connection Science*, 29(2):170–176, 2017.

[38] Rita Borgo, Bongshin Lee, Benjamin Bach, Sara Fabrikant, Radu Jianu, Andreas Kerren, Stephen Kobourov, Fintan McGee, Luana Micallef, Tatiana von Landesberger, et al. Crowdsourcing for information visualization: Promises and pitfalls. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22–27, 2015, Revised Contributions*, pages 96–138. Springer, 2017.

[39] Terry K Borsook and Nancy Higginbotham-Wheat. Interactivity: What is it and what can it do for computer-based instruction? *Educational Technology*, 1991.

[40] Martim Brandao, Gerard Canal, Senka Krivić, Paul Luff, and Amanda Coles. How experts explain motion planner output: a preliminary user-study to inform the design of explainable planners. In *2021 30th IEEE International Conference*

on Robot & Human Interactive Communication (RO-MAN), pages 299–306. IEEE, 2021.

[41] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1):82–111, 2013.

[42] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[43] Wolfram Burgard, Armin B Cremers, Dieter Fox, Dirk Hähnel, Gerhard Lakemeyer, Dirk Schulz, Walter Steiner, and Sebastian Thrun. Experiences with an interactive museum tour-guide robot. *Artificial intelligence*, 114(1-2):3–55, 1999.

[44] Luis V Calderita, Araceli Vega, Sergio Barroso-Ramírez, Pablo Bustos, and Pedro Núñez. Designing a cyber-physical system for ambient assisted living: A use-case analysis for social robot navigation in caregiving centers. *Sensors*, 2020.

[45] Mario Campanella, Serge Hoogendoorn, and Winnie Daamen. The nomad model: theory, developments and applications. *Transportation Research Procedia*, 2014.

[46] Kate Candon, Jesse Chen, Yoony Kim, Zoe Hsu, Nathan Tsoi, , and Marynel Vázquez. Nonverbal human signals can help autonomous agents infer human preferences for their behavior. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, 2023.

[47] Kate Candon, Nicholas C. Georgiou, Helen Zhou, Sidney Richardson, Qiping Zhang, Brian Scassellati, and Marynel Vázquez. React: Two datasets for analyzing both human reactions and evaluative feedback to robots over time, 2024.

[48] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. The robotic social attributes scale (rosas) development and validation. In *HRI*, 2017.

[49] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. The robotic social attributes scale (rosas): Development and validation. In *HRI*, 2017.

[50] Ginevra Castellano, Iolanda Leite, Andre Pereira, Carlos Martinho, Ana Paiva, and Peter W McOwan. Detecting engagement in hri: An exploration of social and task-based context. In *SOCIALCOM-PASSAT*, 2012.

[51] Rohan Chandra, Mingyu Wang, Mac Schwager, and Dinesh Manocha. Game-theoretic planning for autonomous driving among risk-aware human drivers. In *2022 Intl. Conf. on Robotics and Automation (ICRA)*, pages 2876–2883. IEEE, 2022.

[52] Rohan Chandra, Rahul Maligi, Arya Anantula, and Joydeep Biswas. Socialmapf: Optimal and efficient multi-agent path finding with strategic agents for social navigation. *IEEE RAL*, 2023.

[53] Konstantinos Charalampous, Ioannis Kostavelis, and Antonios Gasteratos. Recent trends in social aware robot navigation: A survey. *Robotics and Autonomous Systems*, 93:85–104, 2017.

[54] Yuhang Che, Allison M Okamura, and Dorsa Sadigh. Efficient and trustworthy social navigation via explicit and implicit robot–human communication. *IEEE Transactions on Robotics*, 36(3):692–707, 2020.

[55] Borting Chen, Ho-Pang Hsu, and Yu-Lun Huang. Bringing desktop applications to the web. *IT Prof*, 2016.

[56] Kevin Chen, Juan Pablo de Vicente, Gabriel Sepulveda, Fei Xia, Alvaro Soto, Marynel Vàzquez, and Silvio Savarese. A behavioral approach to visual navigation with graph localization networks. In *RSS*, 2019.

[57] Nicholas Chen. Convention over configuration. http://softwareengineering.vazexqi.com/files/pattern.html, 2006.

[58] Yu Fan Chen, Michael Everett, Miao Liu, and Jonathan P How. Socially aware motion planning with deep reinforcement learning. In *IROS*, 2017.

[59] Yu Fan Chen, Michael Everett, Miao Liu, and Jonathan P How. Socially aware motion planning with deep reinforcement learning. In *2017 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1343–1350. IEEE, 2017.

[60] Jiyu Cheng, Hu Cheng, Max Q-H Meng, and Hong Zhang. Autonomous navigation by mobile robots in human environments: A survey. In *2018 IEEE Intl. Conf. on robotics and biomimetics (ROBIO)*, pages 1981–1986. IEEE, 2018.

[61] Sonia Chernova, Jeff Orkin, and Cynthia Breazeal. Crowdsourcing hri through online multiplayer games. In *AAAI Fall Symposium Series*, 2010.

[62] Mohamed Chetouani. Interactive robot learning: An overview. *ECCAI Advanced Course on Artificial Intelligence*, pages 140–172, 2021.

[63] Ernest Cheung, Aniket Bera, Emily Kubin, Kurt Gray, and Dinesh Manocha. Identifying driver behaviors using trajectory features for vehicle navigation. In *2018 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 3445–3452. IEEE, 2018.

[64] SF Chik, CF Yeong, ELM Su, TY Lim, Y Subramaniam, and PJH Chin. A review of social-aware navigation frameworks for service robot in dynamic human environments. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(11):41–50, 2016.

[65] HeeSun Choi, Cindy Crump, Christian Duriez, Asher Elmquist, Gregory Hager, David Han, Frank Hearl, Jessica Hodgins, Abhinandan Jain, Frederick Leve, et al. On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward. *Proceedings of the National Academy of Sciences*, 118(1):e1907856118, 2021.

[66] Jack Collins, Shelvin Chand, Anthony Vanderkop, and David Howard. A review of physics simulators for robotic applications. *IEEE Access*, 9:51416–51431, 2021.

[67] Filipa Correia, Samuel Gomes, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. The dark side of embodiment teaming up with robots vs disembodied agents. In *RSS*, 2020.

[68] Catie Cuan, Edward Lee, Emre Fisher, Anthony Francis, Leila Takayama, Tingnan Zhang, Alexander Toshev, and Sören Pirk. Gesture2path: Imitation learning for gesture-aware navigation. *arXiv preprint arXiv:2209.09375*, 2022.

[69] Kelly Cuccolo, Megan S Irgens, Martha S Zlokovich, Jon Grahe, and John E Edlund. What crowdsourcing can offer to cross-cultural psychological science. *Cross-Cultural Research*, 2021.

[70] Yuchen Cui, Pallavi Koppol, Henny Admoni, Scott Niekum, Reid Simmons, Aaron Steinfeld, and Tesca Fitzgerald. Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning. In *International Joint Conference on Artificial Intelligence*, 2021.

[71] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*, pages 604–626. PMLR, 2021.

[72] Sean Curtis, Andrew Best, and Dinesh Manocha. Menge: A modular framework for simulating crowd movement. *Collective Dynamics*, 2016.

[73] Andrea Deublein and Birgit Lugrin. (expressive) social robot or tablet? – on the benefits of embodiment and non-verbal expressivity of the interface for a smart environment. In *International Conference on Persuasive Technology*, 2020.

[74] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1959.

[75] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017.

[76] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[77] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 51–58, 2015.

[78] Gilberto Echeverria, Nicolas Lassabe, Arnaud Degroote, and Séverin Lemaignan. Modular open robots simulation engine: Morse. In *ICRA*, 2011.

[79] Gilberto Echeverria, Séverin Lemaignan, Arnaud Degroote, Simon Lacroix, Michael Karg, Pierrick Koch, Charles Lesire, and Serge Stinckwich. Simulating complex robotic scenarios with morse. In *SIMPAR*. Springer, 2012.

[80] Vanessa Evers, Nuno Menezes, Luis Merino, Dariu Gavrila, Fernando Nabais, Maja Pantic, Paulo Alvito, and Daphne Karreman. The development and real-world deployment of frog, the fun robotic outdoor guide. In *HRI*, 2014.

[81] Anthony Favier, Phani-Teja Singamaneni, and Rachid Alami. An intelligent human simulation (inhus) for developing and experimenting human-aware and interactive robot abilities. 2021.

[82] Anthony Favier, Phani-Teja Singamaneni, and Rachid Alami. Simulating intelligent human agents for intricate social robot navigation. In *RSS Workshop on Social Robot Navigation*, 2021.

[83] David Feil-Seifer, Kerstin S Haring, Silvia Rossi, Alan R Wagner, and Tom Williams. Where to next? the impact of covid-19 on human-robot interaction research. *THRI*, 2020.

[84] Gonzalo Ferrer, Anaís Garrell Zulueta, Fernando Herrero Cotarelo, and Al-

berto Sanfeliu. Robot social-aware navigation framework to accompany people walking side-by-side. *Autonomous robots*, 2017.

[85] Kerstin Fischer, Katrin Lohan, and Kilian Foth. Levels of embodiment: Linguistic analyses of factors influencing hri. In *HRI*, 2012.

[86] Susan T Fiske, Amy JC Cuddy, and Peter Glick. Universal dimensions of social cognition: Warmth and competence. *TiCS*, 2007.

[87] Anthony Francis, Claudia Pérez-d'Arpino, Chengshu Li, Fei Xia, Nathan Tsoi, et al. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740*, 2023.

[88] Anthony Francis, Claudia Pérez-D'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Hao-Tien Lewis Chiang, Michael Everett, Sehoon Ha, Justin Hart, Haresh Karnan, Tsang-Wei Edward Lee, Luis J. Manso, Reuth Mirsky, Sören Pirk, Phani Teja Singamaneni, Peter Stone, Ada Taylor, Peter Trautman, Nathan Tsoi, Marynel Vázquez, Xuesu Xiao, Peng Xu, Naoki Yokoyama, Roberto Martín-Martín, and Alexander Toshev. Benchmarking social robot navigation across academia and industry. In *Proc. of the AAAI 2023 Spring Symposium on HRI in Academia and Industry: Bridging the Gap.* AAAI, 2023.

[89] Yuxiang Gao and Chien-Ming Huang. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI*, page 420, 2021.

[90] Yuxiang Gao and Chien-Ming Huang. Evaluation of socially-aware robot navigation. *Front. Robot. AI*, 2022.

[91] Sandra G.Hart and Lowell E.Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 1988.

[92] Christian Gloor. Pedsim: Pedestrian crowd simulation. *URL http://pedsim. silmaril. org*, 5(1), 2016.

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

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

[95] Mar Gonzalez-Franco, Eyal Ofek, Ye Pan, Angus Antley, Anthony Steed, Bernhard Spanlang, Antonella Maselli, Domna Banakou, Nuria Pelechano, Sergio Orts-Escolano, et al. The rocketbox library and the utility of freely available rigged avatars for procedural animation of virtual humans and embodiment. *Frontiers in Virtual Reality*, 2020.

[96] Giorgio Grisetti, Cyrill Stachniss, and Wolfram Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE transactions on Robotics*, 23(1):34–46, 2007.

[97] Balint Gucsi, Danesh S Tarapore, William Yeoh, Christopher Amato, and Long Tran-Thanh. To ask or not to ask: A user annoyance aware preference elicitation framework for social robots. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7935–7940. IEEE, 2020.

[98] Winter Guerra, Ezra Tal, Varun Murali, Gilhyun Ryou, and Sertac Karaman. Flightgoggles: Photorealistic sensor simulation for perception-driven robotics using photogrammetry and virtual reality. In *IEEE International Workshop on Intelligent Robots and Systems (IROS)*, 2019. doi: 10.1109/iros40897.2019.8968116. URL `https://doi.org/10.1109/iros40897.2019.8968116`.

[99] Rodrigo Longhi Guimarães, André Schneider de Oliveira, João Alberto Fabro, Thiago Becker, and Vinícius Amilgar Brenner. Ros navigation: Concepts and

tutorial. In *Robot Operating System (ROS)*, pages 121–160. Springer, 2016.

[100] Hatice Gunes, Massimo Piccardi, and Maja Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. In *Affective Computing*. IntechOpen, 2008.

[101] Edmund T Hall and Edward T Hall. *The hidden dimension*, volume 609. Anchor, 1966.

[102] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*. Elsevier, 1988.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[104] Eric Heiden, Luigi Palmieri, Leonard Bruns, Kai O. Arras, Gaurav S. Sukhatme, and Sven Koenig. Bench-mr: A motion planning benchmark for wheeled mobile robots. *IEEE RAL*, 6(3):4536–4543, 2021. doi: 10.1109/LRA.2021.3068913.

[105] Päivi Heikkilä, Hanna Lammi, Marketta Niemelä, Kathleen Belhassein, Guillaume Sarthou, Antti Tammela, Aurélie Clodic, and Rachid Alami. 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*, pages 548–557. Springer, 2019.

[106] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 1995.

[107] Mehdi Hellou, JongYoon Lim, Norina Gasteiger, Minsu Jang, and Ho Seok Ahn. Technical methods for social robots in museum settings: An overview of the literature. *International Journal of Social Robotics*, 14(8):1767–1786, 2022.

[108] Padraig Higgins, Ryan Barron, Stephanie Lukin, Don Engel, and Cynthia Ma-

tuszek. A collaborative building task in vr vs. reality. In *Proc. of the International Symposium on Experimental Robotics (ISER)*, November 2023.

[109] Guy Hoffman and Wendy Ju. Designing robots with movement in mind. *JHRI*, 2014.

[110] Guy Hoffman and Xuan Zhao. A primer for conducting experiments in human–robot interaction. *JHRI*, 2020.

[111] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. Robot presence and human honesty: Experimental evidence. In *HRI*, 2015.

[112] Blake Holman, Abrar Anwar, Akash Singh, Mauricio Tec, Justin Hart, and Peter Stone. Watch where you're going! gaze and head orientation as predictors for social robot navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3553–3559, 2021. doi: 10.1109/ICRA48506.2021.9561286.

[113] Jarrett Holtz and Joydeep Biswas. SOCIALGYM: A Framework for Benchmarking Social Robot Navigation . In *Intelligent Robots and Systems (IROS), IEEE/RSJ Intl. Conf. on*, pages 11246–11252. IEEE, 2022. doi: 10.1109/IROS47612.2022.9982021.

[114] Yuhan Hu and Guy Hoffman. Using skin texture change to design emotion expression in social robots. In *HRI*. IEEE, 2019.

[115] Tom P Huck, Christoph Ledermann, and Torsten Kröger. Testing robot system safety by creating hazardous human worker behavior in simulation. *IEEE Robotics and Automation Letters*, 7(2):770–777, 2021.

[116] Ahmed Hussein, Fernando García, and Cristina Olaverri-Monreal. Ros and unity based framework for intelligent vehicles control and simulation. In *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pages 1–6. IEEE, 2018.

[117] Helge Hüttenrauch, Kerstin Severinson Eklundh, Anders Green, and Elin A Topp. Investigating spatial relationships in human-robot interaction. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006.

[118] Angel Hsing-Chi Hwang and Andrea Stevenson Won. Ideabot: investigating social facilitation in human-machine team creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[119] Tetsunari Inamura, Yoshiaki Mizuchi, and Hiroki Yamada. Vr platform enabling crowdsourcing of embodied hri experiments–case study of online robot competition. *Advanced Robotics*, 35(11):697–703, 2021.

[120] Ohad Inbar and Joachim Meyer. Politeness counts: Perceptions of peacekeeping robots. *IEEE Trans. on Human-Machine Systems*, 49(3):232–240, 2019.

[121] Walther Jensen, Simon Hansen, and Hendrik Knoche. Knowing you, seeing me: Investigating user preferences in drone-human acknowledgement. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[122] Collin Johnson and Benjamin Kuipers. Socially-aware navigation using topological maps and social norm learning. In *AIES*, 2018.

[123] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. Can we trust online crowdworkers? In *IVA*, 2020.

[124] Michiel Joosse, Manja Lohse, and Vanessa Evers. Crowdsourcing culture in hri: lessons learned from quantitative and qualitative data collections. In *3rd international workshop on culture aware robotics at ICSR*, 2015.

[125] Younbo Jung and Kwan Min Lee. Effects of physical embodiment on social presence of social robots. In *Proceedings of Presence*, 2004.

[126] Hiroko Kamide, Yasushi Mae, Koji Kawabe, Satoshi Shigemi, Masato Hirose, and Tatsuo Arai. New measurement of psychological safety for humanoid. In

2012 7th ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI), pages 49–56. IEEE, 2012.

[127] Takayuki Kanda and Hiroshi Ishiguro. *Human-robot interaction in social robotics*. CRC Press, 2017.

[128] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. A two-month field trial in an elementary school for long-term human–robot interaction. *T-RO*, 2007.

[129] Yuya Kaneshige, Satoru Satake, Takayuki Kanda, and Michita Imai. How to overcome the difficulties in programming and debugging mobile social robots? In *HRI*, 2021.

[130] Mubbasir Kapadia, Nuria Pelechano, Jan Allbeck, and Norm Badler. Virtual crowds: Steps toward behavioral realism. *Synth. lect. comput. vis.*, 2015.

[131] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Soeren Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *arXiv preprint arXiv:2203.15041*, 2022.

[132] Daphne E Karreman, Geke DS Ludden, and Vanessa Evers. Beyond r2d2: Designing multimodal interaction behavior for robot-specific morphology. *THRI*, 2019.

[133] Linh Kästner, Teham Bhuiyan, Tuan Anh Le, Elias Treis, Johannes Cox, Boris Meinardus, Jacek Kmiecik, Reyk Carstens, Duc Pichel, Bassel Fatloun, et al. Arena-bench: A benchmarking suite for obstacle avoidance approaches in highly dynamic environments. *IEEE RAL*, 7(4):9477–9484, 2022.

[134] Yuka Kato. A remote navigation system for a simple tele-presence robot with virtual reality. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4524–4529, 2015. doi: 10.1109/IROS.2015.7354020.

[135] Adam Kendon. Goffman's approach to face-to-face interaction. *Erving Goffman: Exploring the interaction order*, 1988.

[136] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[137] Hiroyuki Kidokoro, Takayuki Kanda, Dražen Brščic, and Masahiro Shiomi. Will i bother here?-a robot anticipating its influence on pedestrian walking comfort. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 259–266. IEEE, 2013.

[138] Rachel Kirby. *Social robot navigation*. PhD thesis, Carnegie Mellon University, The Robotics Institute, 2010.

[139] Ross A Knepper, Christoforos I Mavrogiannis, Julia Proft, and Claire Liang. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, pages 283–292, 2017.

[140] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

[141] Nathan Koenig and John Hsu. The many faces of simulation: Use cases for a general purpose simulator. In *Proc. of the 2013 IEEE International Conference on Robotics and Automation*, volume 13 of *ICRA*, pages 10–11, 2013.

[142] Anna Konrad. Simulation of mobile robots with unity and ros: A case-study and a comparison with gazebo. Master's thesis, Department of Engineering Science, University West, 2019.

[143] Christian U Krägeloh, Jaishankar Bharatharaj, Senthil Kumar Sasthan Kutty, Praveen Regunathan Nirmala, and Loulin Huang. Questionnaires to measure acceptability of social robots: a critical review. *Robotics*, 2019.

[144] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch.

Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

[145] Markus Kuderer, Henrik Kretzschmar, and Wolfram Burgard. Teaching mobile robots to cooperatively navigate in populated environments. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3138–3143. IEEE, 2013.

[146] Minae Kwon, Sandy H Huang, and Anca D Dragan. Expressing robot incapability. In *HRI*, 2018.

[147] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P Losey, and Dorsa Sadigh. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *HRI*. IEEE, 2020.

[148] Alexis Lambert, Nahal Norouzi, Gerd Bruder, and Gregory Welch. A systematic review of ten years of research on human interaction with social robots. *JHCI*, 2020.

[149] Przemyslaw A Lasota, Terrence Fong, Julie A Shah, et al. A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics*, 5(4): 261–349, 2017.

[150] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *Int. J. Hum. Comput. Stud.*, 2006.

[151] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. 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*, pages 695–704, 2012.

[152] Sophie Legros and Beniamino Cislaghi. Mapping the social-norms literature: An overview of reviews. *Perspect. Psychol. Sci.*, 2020.

[153] Séverin Lemaignan, Marc Hanheide, Michael Karg, Harmish Khambhaita, Lars Kunze, Florian Lier, Ingo Lütkebohle, and Grégoire Milliez. Simulation and hri recent perspectives with the morse simulator. In *SIMPAR*. Springer, 2014.

[154] Alexander Lew, Sydney Thompson, Nathan Tsoi, and Marynel Vázquez. Shutter, the robot photographer: Leveraging behavior trees for public, in-the-wild human-robot interactions. *arXiv preprint arXiv:2302.00191*, 2023.

[155] Michael Lewis, Jijun Wang, and Stephen Hughes. Usarsim: Simulation for the study of human-robot interaction. *JCEDM*, 2007.

[156] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *CoRL*, 2021.

[157] Jamy Li. The benefit of being physically present. *Int. J. Hum. Comput. Stud.*, 2015.

[158] Rui Li, Marc van Almkerk, Sanne van Waveren, Elizabeth Carter, and Iolanda Leite. Comparing human-robot proxemics between virtual reality and the real world. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 431–439. IEEE, 2019.

[159] Rui Li, Marc van Almkerk, Sanne van Waveren, Elizabeth Carter, and Iolanda Leite. Comparing human-robot proxemics between virtual reality and the real world. In *HRI*, 2019.

[160] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. A review on interactive reinforcement learning from human social feedback. *IEEE Access*, 8:120757–120765, 2020.

[161] Shih-Yun Lo, Katsu Yamane, and Ken-ichiro Sugiyama. Perception of pedestrian avoidance strategies of a self-balancing mobile robot. In *2019 IEEE/RSJ*

International Conference on Intelligent Robots and Systems (IROS), pages 1243–1250. IEEE, 2019.

[162] David V Lu, Dave Hershberger, and William D Smart. Layered costmaps for context-sensitive navigation. In *IROS*, 2014.

[163] Matthias Luber, Luciano Spinello, Jens Silva, and Kai O Arras. Socially-aware robot navigation: A learning approach. In *Intelligent robots and systems (IROS), 2012 IEEE/RSJ Intl. Conf. on*, pages 902–907. IEEE, 2012.

[164] Sean D. Lynch, Julien Pettré, Julien Bruneau, Richard Kulpa, Armel Crétual, and Anne-Helene Olivier. Effect of virtual human gaze behaviour during an orthogonal collision avoidance walking task. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 136–142, 2018. doi: 10.1109/VR.2018.8446180.

[165] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*, pages 2285–2294. PMLR, 2017.

[166] Ratnesh Madaan, Nicholas Gyde, Sai Vemprala, Matthew Brown, Keiko Nagami, Tim Taubner, Eric Cristofalo, Davide Scaramuzza, Mac Schwager, and Ashish Kapoor. Airsim drone racing lab. In *Neurips 2019 competition and demonstration track*. PMLR, 2020.

[167] Maxim Makatchev, Reid Simmons, Majd Sakr, and Micheline Ziadee. Expressing ethnicity through behaviors of a robot character. In *HRI*, 2013.

[168] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, Silvio Savarese, and Li Fei-Fei. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *CORL*, 2018.

[169] Swathi Mannem, William Macke, Peter Stone, and Reuth Mirsky. Exploring

the cost of interruptions in human-robot teaming. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2023.

[170] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019.

[171] Eitan Marder-Eppstein, Eric Berger, Tully Foote, Brian Gerkey, and Kurt Konolige. The office marathon: Robust navigation in an indoor office environment. In *2010 IEEE Intl. Conf. on robotics and automation*, pages 300–307. IEEE, 2010.

[172] Roberto Martín-Martín, Hamid Rezatofighi, Abhijeet Shenoi, Mihir Patel, J Gwak, Nathan Dass, Alan Federman, Patrick Goebel, and Silvio Savarese. Jrdb: A dataset and benchmark for visual perception for navigation in human environments. *arXiv preprint arXiv:1910.11792*, 2019.

[173] Christoforos Mavrogiannis, Alena M Hutchinson, John Macdonald, Patrícia Alves-Oliveira, and Ross A Knepper. Effects of distinct robot navigation strategies on human behavior in a crowded environment. In *HRI*, 2019.

[174] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Aaron Steinfeld, Pete Trautman, and Jean Oh. Core challenges of social robot navigation: A survey. *arXiv:2103.05668*, 2021.

[175] Christoforos Mavrogiannis, Patrícia Alves-Oliveira, Wil Thomason, and Ross A Knepper. Social momentum: Design and evaluation of a framework for socially competent robot navigation. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(2):1–37, 2022.

[176] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navi-

gation: A survey. *ACM Transactions on Human-Robot Interaction*, 12(3):1–39, 2023.

[177] Emily McQuillin, Nikhil Churamani, and Hatice Gunes. Learning socially appropriate robo-waiter behaviours through real-time user feedback. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 541–550. IEEE, 2022.

[178] Wei Meng, Yuchao Hu, Jiaxin Lin, Feng Lin, and Rodney Teo. Ros+ unity: An efficient high-fidelity 3d multi-uav navigation and control simulator in gps-denied environments. In *IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society*, pages 002562–002567. IEEE, 2015.

[179] Reuth Mirsky, Xuesu Xiao, Justin Hart, and Peter Stone. Conflict avoidance in social navigation—a survey. *ACM Transactions on Human-Robot Interaction*, 13(1):1–36, 2024.

[180] Daxton Mitchell, HeeSun Choi, and Justin M Haney. Safety perception and behaviors during human-robot interaction in virtual environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, 2020.

[181] Noriaki Mitsunaga, Christian Smith, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Adapting robot behavior for human–robot interaction. *IEEE Transactions on Robotics*, 24(4):911–916, 2008.

[182] Ali Mollahosseini, Hojjat Abdollahi, Timothy D. Sweeny, Ron Cole, and Mohammad H. Mahoor. Role of embodiment and presence in human perception of robots' facial cues. *Int. J. Hum. Comput. Stud.*, 2018.

[183] Ronja Möller, Antonino Furnari, Sebastiano Battiato, Aki Härmä, and Giovanni Maria Farinella. A survey on human-aware robot navigation. *Robotics and Autonomous Systems*, 145:103837, 2021.

[184] Jerry Muller. *The tyranny of metrics*. Princeton University Press, 2018.

[185] Bilge Mutlu and Jodi Forlizzi. 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*, pages 287–294, 2008.

[186] Amal Nanavati, Xiang Zhi Tan, Joe Connolly, and Aaron Steinfeld. Follow the robot: Modeling coupled human-robot dyads during navigation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3836–3843. IEEE, 2019.

[187] Austin Narcomey, Nathan Tsoi, Ruta Desai, and Marynel Vázquez. Learning human preferences over robot behavior as soft planning constraints. *arXiv preprint arXiv:2403.19795*, 2024.

[188] Mollik Nayyar, Zachary Zoloty, Ciera McFarland, and Alan R Wagner. Exploring the effect of explanations during robot-guided emergency evacuation. In *ICSR*, 2020.

[189] Aastha Nigam and Laurel D Riek. Social context perception for mobile robots. In *IROS*, 2015.

[190] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *HRI*, 2016.

[191] Ali Noormohammadi-Asl, Kevin Fan, Stephen L Smith, and Kerstin Dautenhahn. Human leading or following preferences: Effects on human perception of the robot and the human-robot collaboration. *arXiv preprint arXiv:2401.01466*, 2024.

[192] Illah R Nourbakhsh, Katia Sycara, Mary Koes, Mark Yong, Michael Lewis, and Steve Burion. Human-robot teaming for search and rescue. *IEEE Pervasive Computing*, 4(1):72–79, 2005.

[193] NVIDIA. Isaac simulator. `https://developer.nvidia.com/isaac-sim`, 2022.

[194] Billy Okal and Kai O Arras. Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2889–2895. IEEE, 2016.

[195] Aude Olivia, Michael L Mack, Mochan Shrestha, and Angela Peeper. Identifying the perceptual dimensions of visual complexity of scenes. In *Proceedings of the annual meeting of the cognitive science society*, 2004.

[196] Valerio Ortenzi, Akansel Cosgun, Tommaso Pardi, Wesley P Chan, Elizabeth Croft, and Dana Kulić. Object handovers: a review for robotics. *T-RO*, 2021.

[197] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054. IEEE, 2016.

[198] Hong Seong Park, Jeong Seok Kang, Hyeong Seob Choi, Sang Woo Meng, and Si Wan Kim. Simulation-based interface testing automation system and method for robot software components, December 3 2013. US Patent 8,601,436.

[199] HD Patterson. Maximum likelihood estimation of components of variance. In *IBC*. Biometric Soc., 1975.

[200] Nuria Pelechano, Jan M Allbeck, and Norman I Badler. Controlling individual agents in high-density crowd simulation. In *SCA*, 2007.

[201] Claudia Pérez-D'Arpino, Can Liu, Patrick Goebel, Roberto Martín-Martín, and Silvio Savarese. Robot navigation in constrained pedestrian environments using reinforcement learning. *arXiv:2010.08600*, 2020.

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

[203] Christopher Peters and Cathy Ennis. Modeling groups of plausible virtual pedestrians. *CG&A*, 2009.

[204] Björn Petrak, Gundula Sopper, Katharina Weitz, and Elisabeth André. Do you mind if I pass through? Studying the appropriate robot behavior when traversing two conversing people in a hallway setting. In *2021 30th IEEE Intl. Conf. on Robot & Human Interactive Communication (RO-MAN)*, pages 369–375. IEEE, 2021.

[205] Mark Pfeiffer, Michael Schaeuble, Juan Nieto, Roland Siegwart, and Cesar Cadena. From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots. In *2017 ieee international conference on robotics and automation (icra)*, pages 1527–1533. IEEE, 2017.

[206] Sören Pirk, Edward Lee, Xuesu Xiao, Leila Takayama, Anthony Francis, and Alexander Toshev. A protocol for validating social navigation policies. *arXiv preprint arXiv:2204.05443*, 2022.

[207] Ashwini Pokle, Roberto Martín-Martín, Patrick Goebel, Vincent Chow, Hans M. Ewald, Junwei Yang, Zhenkai Wang, Amir Sadeghian, Dorsa Sadigh, Silvio Savarese, and Marynel Vázquez. Deep local trajectory replanning and control for robot navigation. In *ICRA*, 2019.

[208] Ashwini Pokle, Roberto Martín-Martín, Patrick Goebel, Vincent Chow, Hans M. Ewald, Junwei Yang, Zhenkai Wang, Amir Sadeghian, Dorsa Sadigh, Silvio Savarese, and Marynel Vázquez. Deep local trajectory replanning and control for robot navigation. In *ICRA*, 2019.

[209] Louise Poubel. Service robot simulator, 2020. URL `https://github.com/osrf/servicesim`.

[210] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[211] Carolyn C Preston and Andrew M Colman. Optimal number of response cate-

gories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1):1–15, 2000.

[212] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: An open-source robot operating system. In *ICRA*, 2009.

[213] Tanmay Randhavane, Aniket Bera, Emily Kubin, Austin Wang, Kurt Gray, and Dinesh Manocha. Pedestrian dominance modeling for socially-aware robot navigation. In *2019 Intl. Conf. on Robotics and Automation (ICRA)*, pages 5621–5628. IEEE, 2019.

[214] Carmine Recchiuto and Antonio Sgorbissa. Diversity-aware social robots meet people: beyond context-aware embodied ai. *arXiv preprint arXiv:2207.05372*, 2022.

[215] Ely Repiso, Anaís Garrell, and Alberto Sanfeliu. Adaptive side-by-side social robot navigation to approach and interact with people. *IJSR*, 2019.

[216] Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH*, 1987.

[217] Laurel D. Riek, Tal-Chen Rabinowitch, Paul Bremner, Anthony G. Pipe, Mike Fraser, and Peter Robinson. Cooperative gestures: Effective signaling for humanoid robots. In *HRI*, 2010.

[218] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. From proxemics theory to socially-aware navigation: A survey. *Intl. Journal of Social Robotics*, 7(2):137–153, 2015.

[219] Claire Rivoire and Angelica Lim. The delicate balance of boring and annoying: Learning proactive timing in long-term human robot interaction, 2016.

[220] Paul Robinette, Alan R Wagner, and Ayanna M Howard. Assessment of robot guidance modalities conveying instructions to humans in emergency situations. In *RO-MAN*, 2014.

[221] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[222] Stephanie Rosenthal, Joydeep Biswas, and Manuela M Veloso. An effective personal mobile robot agent through symbiotic human-robot interaction. In *AAMAS*, volume 10, pages 915–922, 2010.

[223] Matteo Rubagotti, Inara Tusseyeva, Sara Baltabayeva, Danna Summers, and Anara Sandygulova. Perceived safety in physical human–robot interaction—a survey. *Robotics and Autonomous Systems*, 151:104047, 2022.

[224] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *Int J Rob Res*, 2020.

[225] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. pearson, 2016.

[226] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and systems*, volume 2, pages 1–9. Ann Arbor, MI, USA, 2016.

[227] Nicole Salomons, Tom Wallenstein, Debasmita Ghose, and Brian Scassellati. The impact of an in-home co-located robotic coach in helping people make fewer exercise mistakes. In *RO-MAN*, 2022.

[228] Satoru Satake, Takayuki Kanda, Dylan F Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How to approach humans? strategies for social robots to initiate interaction. In *HRI*, 2009.

[229] Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine*, 37(1):22–29, 2018.

[230] Gary W Selnow. Using interactive computer to communicate scientific information. *American Behavioral Scientist*, 1988.

[231] Stela H Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E Young. Poor thing! would you feel sorry for a simulated robot? a comparison of empathy toward a physical and a simulated robot. In *HRI*. IEEE, 2015.

[232] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017.

[233] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. A larger audience, please!—encouraging people to listen to a guide robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 31–38. IEEE, 2010.

[234] Samuel S. Sohn, Honglu Zhou, Seonghyeon Moon, Sejong Yoon, Vladimir Pavlovic, , and Mubbasir Kapadia. Laying the foundations of deep long-term crowd flow prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[235] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. In *HRI*, 2006.

[236] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. The oz of wizard: simulating the human for interaction research. In *HRI*, 2009.

[237] Krzysztof Stencel and Patrycja Węgrzynowicz. Implementation variants of the singleton design pattern. In *OTM*, 2008.

[238] Maia Stiber. Effective human-robot collaboration via generalized robot error management using natural human responses. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 673–678, 2022.

[239] Maia Stiber, Russell Taylor, and Chien-Ming Huang. Modeling human response

to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 676–683. IEEE, 2022.

[240] Maia Stiber, Russell H. Taylor, and Chien-Ming Huang. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 222–230, New York, NY, USA, 2023. Association for Computing Machinery. URL `https://doi.org/10.1145/3568162.3576990`.

[241] Megan Strait, Cody Canning, and Matthias Scheutz. Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 479–486, 2014.

[242] Walter W Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012.

[243] Aamodh Suresh, Angelique Taylor, Laurel D Riek, and Sonia Martinez. Robot navigation in risky, crowded environments: Understanding human preferences. *arXiv preprint arXiv:2303.08284*, 2023.

[244] Nilesh Suriyarachchi, Rohan Chandra, John S Baras, and Dinesh Manocha. Gameopt: Optimal real-time multi-agent planning and control for dynamic intersections. In *2022 IEEE 25th Intl. Conf. on Intelligent Transportation Systems (ITSC)*, pages 2599–2606. IEEE, 2022.

[245] Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. Improving social awareness through dante: Deep affinity network for clustering conversational interactants. *Proc. ACM Hum.-Comput. Interact.*, 4, May 2020. URL `https://doi.org/10.1145/3392824`.

[246] Lei Tai, Jingwei Zhang, Ming Liu, and Wolfram Burgard. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1111–1117. IEEE, 2018.

[247] Antero Taivalsaari, Tommi Mikkonen, Dan Ingalls, and Krzysztof Palacz. Web browser as an application platform. In *SEAA*, 2008.

[248] Kenta Takaya, Toshinori Asai, Valeri Kroumov, and Florentin Smarandache. Simulation environment for mobile robots testing using ros and gazebo. In *ICSTCC*, 2016.

[249] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *HRI*, 2011.

[250] Ben Talbot, David Hall, Haoyang Zhang, Suman Raj Bista, Rohan Smith, Feras Dayoub, and Niko Sünderhauf. Benchbot: Evaluating robotics research in photorealistic 3d simulation and on real robots. *arXiv preprint arXiv.2008.00635*, 2020.

[251] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 114–122. IEEE, 2019.

[252] Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International journal of medical education*, 2:53, 2011.

[253] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. Physical vs. virtual agent embodiment and effects on social interaction. In *IVA*, 2016.

[254] Andrea L Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, 2008.

[255] Sydney Thompson, Abhijit Gupta, Anjali W Gupta, Austin Chen, and Marynel

Vázquez. Conversational group detection with graph neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 248–252, 2021.

[256] Sebastian Thrun, Maren Bennewitz, Wolfram Burgard, Armin B Cremers, Frank Dellaert, Dieter Fox, Dirk Hähnel, Charles Rosenberg, Nicholas Roy, Jamieson Schulte, et al. Minerva: A tour-guide robot that learns. In *AAAI*, 1999.

[257] Leimin Tian and Sharon Oviatt. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2): 1–32, 2021.

[258] Russell Toris, David Kent, and Sonia Chernova. The robot management system: A framework for conducting human-robot interaction studies through crowdsourcing. *JHRI*, 2014.

[259] Russell Toris, Julius Kammerl, David V Lu, Jihoon Lee, Odest Chadwicke Jenkins, Sarah Osentoski, Mitchell Wills, and Sonia Chernova. Robot web tools: Efficient messaging for cloud robotics. In *IROS*, 2015.

[260] Elena Torta, Raymond H Cuijpers, and James F Juola. Design of a parametric model of personal space for robotic social navigation. *IJSR*, 2013.

[261] Pete Trautman, Jeremy Ma, Richard M Murray, and Andreas Krause. Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. *The International Journal of Robotics Research*, 34(3):335–356, 2015.

[262] Peter Trautman and Andreas Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 797–803. IEEE, 2010.

[263] Peter Trautman and Karankumar Patel. Real time crowd navigation from first principles of probability theory. In *ICAPS*, 2020.

[264] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore, et al. Spencer: A socially aware service robot for passenger guidance and help in busy airports. In *FSR*, 2016.

[265] Joanne Truong, Max Rudolph, Naoki Yokoyama, Sonia Chernova, Dhruv Batra, and Akshara Rai. Rethinking sim2real: Lower fidelity simulation leads to higher sim2real transfer in navigation. *arXiv preprint arXiv:2207.10821*, 2022.

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

[267] Nathan Tsoi, Mohamed Hussein, Jeacy Espinoza, Xavier Ruiz, and Marynel Vázquez. Sean: Social environment for autonomous navigation. In *HAI*, 2020.

[268] Nathan Tsoi, Mohamed Hussein, Jeacy Espinoza, Xavier Ruiz, and Marynel Vázquez. Sean: Social environment for autonomous navigation. In *HAI*, 2020.

[269] Nathan Tsoi, Mohamed Hussein, Olivia Fugikawa, J. D. Zhao, and Marynel Vázquez. An approach to deploy interactive robotic simulators on the web for hri experiments: Results in social robot navigation. In *IROS*, 2021.

[270] Nathan Tsoi, Mohamed Hussein, Olivia Fugikawa, J. D. Zhao, and Marynel Vazquez. An approach to deploy interactive robotic simulators on the web for hri experiments: Results in social robot navigation. In *IROS*, 2021.

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

[272] Nathan Tsoi, Kate Candon, Deyuan Li, Yofti Milkessa, and Marynel Vázquez.

Bridging the gap: Unifying the training and evaluation of neural network binary classifiers. *NeurIPS*, 2022.

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

[274] Nathan Tsoi, Alec Xiang, Peter Yu, Samuel S. Sohn, Greg Schwartz, Subashri Ramesh, Mohamed Hussein, Anjali W. Gupta, Mubbasir Kapadia, and Marynel Vázquez. Sean 2.0: Formalizing and generating social situations for robot navigation. *RA-L*, 2022.

[275] Nathan Tsoi, Jessica Romero, and Marynel Vázquez. How do robot experts measure the success of social robot navigation? In *HRI Companion*, 2024.

[276] Nathan Tsoi, Rachel Sterneck, Xuan Zhao, and Marynel Vázquez. Influence of simulation and interactivity on human perceptions of a robot during navigation tasks. *THRI*, 2024.

[277] Jur Van den Berg, Ming Lin, and Dinesh Manocha. Reciprocal velocity obstacles for real-time multi-agent navigation. In *ICRA*, 2008.

[278] Dizan Vasquez, Billy Okal, and Kai O Arras. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *IROS*, 2014.

[279] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[280] Marynel Vázquez, Yofti Milkessa, Michelle M Li, and Neha Govil. Gaze by semi-virtual robotic heads: Effects of eye and head motion. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11065–11071. IEEE, 2020.

[281] Gentiane Venture and Dana Kulić. Robot expressive motions: a survey of generation and evaluation methods. *THRI*, 2019.

[282] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image Vis. Comput.*, 2009.

[283] Stanford Vision and Learning Lab. igibsonchallenge2021. `https://github.com/StanfordVL/iGibsonChallenge2021`, 2022.

[284] Lennart Wachowiak, Peter Tisnikar, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan. Analysing eye gaze patterns during confusion and errors in human–agent collaborations. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 224–229. IEEE, 2022.

[285] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. The role of physical embodiment in human-robot interaction. In *RO-MAN*, 2006.

[286] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. Embodiment and human-robot interaction: A task-based perspective. In *RO-MAN*, 2007.

[287] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafir. Communicating robot motion intent with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 316–324, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349536. doi: 10.1145/3171221.3171253. URL `https://doi.org/10.1145/3171221.3171253`.

[288] Michael Walker, Thao Phung, Tathagata Chakraborti, Tom Williams, and Daniel Szafir. Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy. *arXiv preprint arXiv:2202.11249*, 2022.

[289] Michael L Walters, Kerstin Dautenhahn, René Te Boekhorst, Kheng Lee Koay,

Christina Kaouri, Sarah Woods, Chrystopher Nehaniv, David Lee, and Iain Werry. The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In *ROMAN*, 2005.

[290] Junxian Wang, Wesley P Chan, Pamela Carreno-Medrano, Akansel Cosgun, and Elizabeth Croft. Metrics for evaluating social conformity of crowd navigation algorithms. In *2022 IEEE Intl. Conf. on Advanced Robotics and Its Social Impacts (ARSO)*, pages 1–6. IEEE, 2022.

[291] Manhua Wang, Seul Chan Lee, Harsh Kamalesh Sanghavi, Megan Eskew, Bo Zhou, and Myounghoon Jeon. In-vehicle intelligent agents in fully autonomous driving: The effects of speech style and embodiment together and separately. In *AutomotiveUI*, 2021.

[292] Ning Wang, David V Pynadath, and Susan G Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *HRI*. IEEE, 2016.

[293] Jan M Wiener, Simon J Büchner, and Christoph Hölscher. Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spat. Cogn. Comput.*, 2009.

[294] Luc Wijnen, Séverin Lemaignan, and Paul Bremner. Towards using virtual reality for replicating hri studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 514–516, 2020.

[295] Sarah N Woods, Michael L Walters, Kheng Lee Koay, and Kerstin Dautenhahn. Methodological issues in hri: A comparison of live and video-based methods in robot to human approach direction trials. In *RO-MAN*, 2006.

[296] Sarah N. Woods, Michael L. Walters, Kheng Lee Koay, and Kerstin Dautenhahn. Methodological issues in hri: A comparison of live and video-based methods in robot to human approach direction trials. In *RO-MAN*, 2006.

[297] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer*

Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on. IEEE, 2018.

[298] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 5(2):713–720, 2020.

[299] Xuesu Xiao, Tingnan Zhang, Krzysztof Marcin Choromanski, Tsang-Wei Edward Lee, Anthony Francis, Jake Varley, Stephen Tu, Sumeet Singh, Peng Xu, Fei Xia, Sven Mikael Persson, Leila Takayama, Roy Frostig, Jie Tan, Carolina Parada, and Vikas Sindhwani. Learning model predictive controllers with real-time attention for real-world navigation. In *Conf. on Robot Learning*. PMLR, 2022.

[300] Jin Xu and Ayanna Howard. How much do you trust your self-driving car? exploring human-robot trust in high-risk scenarios. In *SMC*, 2020.

[301] Qianli Xu, Jamie Ng, Odelia Tan, Zhiyong Huang, Benedict Tay, and Taezoon Park. Methodological issues in scenario-based evaluation of human–robot interaction. *IJSR*, 2015.

[302] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. Frownonerror: Interrupting responses from smart speakers by facial expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

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

[304] Mohammad Abu Yousuf, Yoshinori Kobayashi, Yoshinori Kuno, Akiko Yamazaki, and Keiichi Yamazaki. Development of a mobile museum guide robot that can configure spatial formation with visitors. In *ICIC*. Springer, 2012.

[305] Myroslava Zaiets. Factors influencing the mass adoption of vr video platforms. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2021.

[306] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. Space speaks: towards socially and personality aware visual surveillance. In *MPVA*, 2010.

[307] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. Self-annotation methods for aligning implicit and explicit human feedback in human-robot interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 398–407, 2023.

[308] Qiping Zhang, Nathan Tsoi, and Marynel Vázquez. Sean-vr: An immersive virtual reality experience for evaluating social robot navigation. In *HRI Companion*, 2023.

[309] Qiping Zhang, Nathan Tsoi, Mofeed Nagib, Booyeon Choi, Jie Tan, Hao-Tien Lewis Chiang, and Marynel Vázquez. Predicting human perceptions of robot performance during navigation tasks. *ACM Transactions on Human-Robot Interaction*, 2025.

[310] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019.

[311] Jakub A Złotowski, Astrid Weiss, and Manfred Tscheligi. Navigating in public space: participants' evaluation of a robot's approach behavior. In *HRI*, 2012.

[312] Jakub Złotowski, Astrid Weiss, and Manfred Tscheligi. Navigating in public space: Participants' evaluation of a robot's approach behavior. In *HRI*, 2012.