A Protocol for Validating Social Navigation Policies

Sören Pirk[1] Edward Lee[1] Xuesu Xiao[2,3] Leila Takayama[1,4] Anthony Francis[1] Alexander Toshev[5]

***Abstract*— Enabling socially acceptable behavior for situated agents is a major goal of recent robotics research. Robots should not only operate safely around humans, but also abide by complex social norms. A key challenge for developing socially-compliant policies is measuring the quality of their behavior. Social behavior is enormously complex, making it difficult to create reliable metrics to gauge the performance of algorithms. In this paper, we propose a protocol for social navigation benchmarking that defines a set of canonical social navigation scenarios and an in-situ metric for evaluating performance on these scenarios using questionnaires. Our experiments show this protocol is realistic, scalable, and repeatable across runs and physical spaces. Our protocol can be replicated verbatim or it can be used to define a social navigation benchmark for novel scenarios. Our goal is to introduce a protocol for benchmarking social scenarios that is homogeneous and comparable.**

I. Introduction

One of the main prerequisite of making robots ubiquitous and generally applicable is to endow them with the ability to move around people in a socially acceptable manner. A robot must be able to accomplish navigation tasks while adhering to social norms in shared spaces and respecting human actions and behaviors. We refer to such type of navigation as Social Navigation. Recently, the robotics community has witnessed an increased interest in Social Navigation. Among many other directions, researchers investigate the importance of respecting personal space [1], maintaining social dynamics [23] and velocities [9], socially-acceptable approaching behavior [6], and navigation in the presence of groups of people [10].

While these approaches are a testament for the rapid progress in this direction, the evaluation and comparison of algorithms has proven to be difficult for Social Navigation research. To facilitate progress in the community it is of paramount importance to have shared, realistic, repeatable, and scalable benchmarks. If one is to define such a benchmark, then simulation is one tool of choice – we have seen an increase in robotic simulation environments that focus on physics and visual realism [19], [24], [14]; simulation is scalable and repeatable across labs. However, simulating the intricacies of human behavior at different levels of abstraction, ranging from atomic actions and motion dynamics to more complex activities and behavior, has proven to be difficult. Therefore, simulation of humans falls short of providing a realistic medium for social navigation benchmarking.

[1]Robotics at Google
[2]Everyday Robots, Alphabet, www.everydayrobots.com
[3]The University of Texas at Austin
[4]Hoku Labs
[5]Apple ML Research; work done at Robotics at Google

Fig. 1. Frontal approach scenario: human and robot interact by moving along a straight trajectory in opposite directions (left). The robot yields early on to not block the human from walking along their path (right).

A different approach for evaluation is to perform demonstrations and studies in uncontrolled real settings in the wild, e.g. accessible public spaces such as university campuses. While such evaluations are by definition realistic, they face several limitations. For one, they are not guaranteed to be behaviorally natural as one has to enact social interactions in real studies, which may lead to undesired patterns in the observations (e.g. such as teetering motions). More importantly, though, real experiments are difficult to repeat and to conduct at scale. Every run, even under controlled conditions, will differ from prior runs and running experiments repeatedly – with humans in the loop – can be extremely costly. Finally, social interactions in real environments are defined by a wide range of variables (e.g. differences in human behavior and appearance, environmental settings, etc.) that make obtaining meaningful measurements difficult.

To address these challenges, we aim to propose a protocol for establishing a social navigation benchmark. The desired properties of our benchmark are:

- **Realism**: The benchmark is implemented in a real environment with real robots and real humans;
- **Scalability**: The benchmark allows for testing on a diverse set of social situations, with a cost which allows for frequent evaluations;
- **Repeatability**: The benchmark is repeatable across different runs and instantiations in different physical spaces.

To achieve the above properties we propose a real benchmark based on a predefined set of social scenarios evaluated using user surveys. In more detail, we introduce a set of social navigation scenarios (Figure 2) implemented in real-world settings (Figure 1). Each scenario is a canonical example of a common human-robot interaction that can occur when performing navigation tasks. The idea is that reducing a human-robot interaction to its essence allows us to better understand and validate social behavior of humans and robots. To this end, we define the scenarios in a way that they can be replicated by different labs with low effort and so as to avoid high variance in the human-robot interaction.

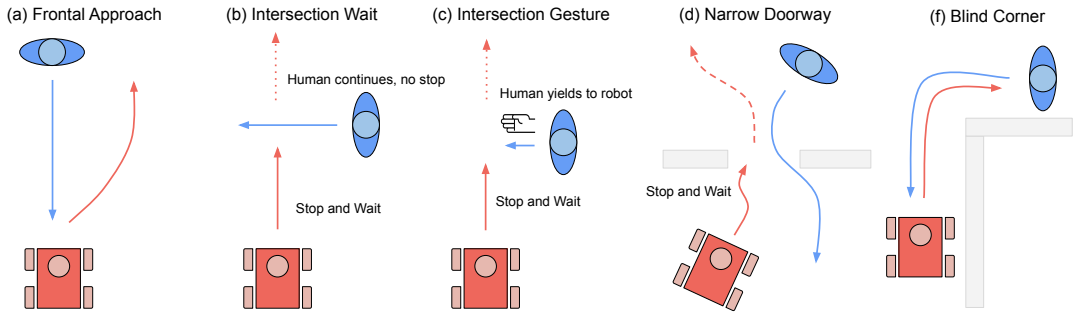

Fig. 2. The five social navigation scenarios of our benchmark: frontal approach (a), intersection wait (b), intersection gesture (c), narrow doorway (d), and blind corner (e).

This addresses the challenge of repeatability as enacting the canonical scenarios will be comparable.

Second, we propose an in-situ metric based on questionnaires to obtain ratings of humans who experienced the interaction with the robot. Unlike other metrics, this allows us to validate the performance of navigation policies w.r.t. social expectations. We show initial results of validating different policies and that questionnaires can be used to measure meaningful gradients for validating social navigation policies. Third, we present guidelines and good practices for defining social navigation benchmarks for other scenarios and environments. Altogether, we hope that our protocol will prove useful for the community to converge to more standardized validation setups.

II. Related Work

Driven by its importance to robotics, social navigation has become the focus of a growing body of research. Because of social navigation's complexity and growth, we cannot comprehensively discuss all related work. For an overview touching on the validation of social navigation, the interested reader is referred to the recent survey papers of Gao and Huang [5], Charalampous et al. [4], Mavrogiannis et al. [15], Kruse et al. [12], Rios-Martinez [21], and more recently, Xiao et al. [26] and Mirsky et al. [16]. Closest to our work are methods that use metrics to measure human discomfort [20], [11], [17] or sociability [18], [22], [2], datasets for social navigation [8], as well as approaches that employ questionnaires for validation purposes [3], [7], [25].

III. Benchmark for Social Navigation

To establish a social navigation benchmark our goal is to define a set of canonical social navigation scenarios. Each scenario represents a common interaction between a human and a robot performing a navigation task. Specifically, we define *frontal approach*, *intersection wait*, *intersection gesture*, *narrow doorway*, and *blind corner*. For each scenario we define start and end points and task the robot to navigate along the trajectory between the two points (see Fig. 3). Each scenario is then enacted by a human, who is provided with a short description of what is expected to happen. For example, for our frontal approach scenario, we simply say "Please walk along this trajectory, start walking when the robot is here, the robot is expected to yield." The human is then walking in the opposite direction of the robot, while

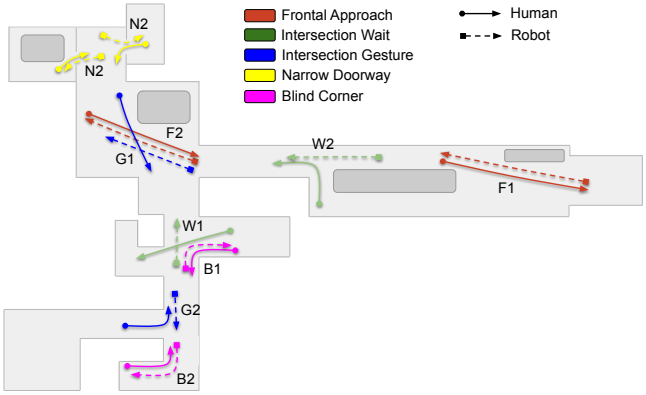

Fig. 3. Illustration of the locations we used for each of the social scenarios: frontal approach (red), intersection wait (green), intersection gesture (blue), narrow doorway (yellow), and blind corner (magenta). Grey round boxes represent obstacles present in the building (e.g., such as chairs or desks). Arrows with round ends and solid lines indicate human trajectories and arrows with square ends and dashed lines robot trajectories.

the robot is driving toward its goal position. By defining a specific scenario with a constrained trajectory we define a canonical example of a social interaction.

Note that we keep the definition of social scenarios as lightweight as possible – we only define the scenario along with a brief description and the start and end points of the trajectory for the robot. We do not constrain the environment, human behavior, human appearance, or the desired robot trajectory. This allows us to implement and validate social scenarios in a variety of environments, as well as repeatedly measuring the performance of a policy on a defined social scenario. To establish a benchmark we define each scenario two times in different locations of our building (Fig. 3).

A. Social Navigation Scenarios

For our current benchmark we have selected five social scenarios of human-robot interactions as detailed below. For most of our scenarios we focus on observant and passive robot behavior; i.e. the robot is expected to yield and make room for the human. Consequently, our ideal social navigation policy would generate robot behavior in a way that the human would almost not notice the robot. An illustration of the social scenarios of our benchmark is shown in Fig. 2. Examples of our real setup for each social scenario are shown in Fig. 1, 4, and 5.

Frontal: Robot and human are approaching each other from two ends of a straight trajectory; enough space is

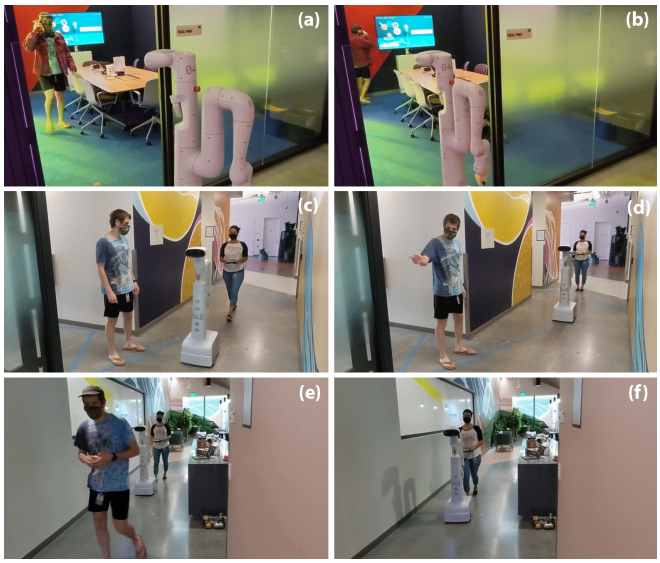

Fig. 4. Examples of the social scenarios: narrow doorway (a, b), intersection gesture (c, d), and intersection wait (e, f). Each scenario is defined by enacting the social interaction with a human and a robot that is manually controlled by a human operator.

provided for the robot to yield. Robot and human are walking toward each other and the robot is expected to yield early on to avoid socially-intimidating behavior. Human and robot are alternating their start and end positions.

Intersection Wait: Robot and human approach each other on perpendicular trajectories. The human does not stop walking down their path. The robot is expected to drive slowly when it approaches the human and it has to come to a complete stop to yield to the human. After the human is out of sight, the robot continues on its trajectory.

Intersection Gesture: Robot and human approach each other on perpendicular trajectories. Human and robot come to a complete stop. The human recognizes the robot and then gestures – with a waving hand motion along the trajectory of the robot – that they yield to the robot. The robot interprets the gesture and continues its path.

Narrow Doorway: Human and robot cross each other's paths by moving through a narrow doorway. Robot and human alternately start inside or outside a room and try to get in or out. In this scenario, the human or the robot has to yield to the respective other agent. If the robot arrives at the door before the human it is allowed to continue on its path. If the human arrives at the door first, the robot has to wait outside the door and yield to the human.

Blind Corner: Human and robot cross each other's paths at a blind corner. Human and robot move down a hallway toward the corner and 'surprise each other' by meeting at the corner at the same time. Both agents have to come to a complete stop to then resolve the situation. The robot is expected to either yield to the human after the collision or to avoid the collision by anticipating the situation.

IV. In-Situ Validation

To validate the social-compliance of a policy we define an in-situ metric based on a questionnaire for each social scenario. We use a five-level Likert scale [13] to define

Frontal Approach	
1	The robot moved to avoid me.
2*	The robot obstructed my path.
3	The robot maintained a safe and comfortable distance at all times.
4*	The robot nearly collided with me.
5	It was clear what the robot wanted to do.
Intersection Wait	
6	The robot let me cross the intersection by maintaining a safe and comfortable distance.
7	The robot changed course to let me pass.
8	The robot paid attention to what I was doing.
9	The robot slowed down and stopped to let me pass.
Intersection Gesture	
10	The robot maintained a safe and comfortable distance at all times.
11	The robot slowed down and stopped.
12	The robot followed my command.
13	I felt the robot paid attention to what I was doing.
Narrow Doorway	
14*	The robot got in my way.
15	The robot moved to avoid me.
16	The robot made room for me to enter or exit.
17*	It was clear what the robot wanted to do.
Blind Corner	
18	The robot moved to avoid me.
19	The robot stopped to let me pass.
20*	I had to move around the robot.
21*	The robot nearly collided with me head-on.

TABLE I

Questions for each social scenario.

4 - 5 questions for each scenario and ask participants to rate their agreement toward these questions based on the following scale: 1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree. Additionally, we allow for the rating 0=cannot tell/something went wrong. The questions for each scenario are provided in Table I: "How much do you agree with the statement ... ?". We run the scenario with human and robot and immediately ask the participant to provide the rating before performing another scenario. This allows us to get reliable in-experience ratings.

V. Experiments

To begin validating our benchmark, we have conducted experiments demonstrating that our setup is scalable, reliable and repeatable. For most of the experiments we use a simple iLQR-based model predictive controller (MPC) to generate linear and angular velocity commands for our robot (provided by Everyday Robots, Alphabet). We repeatedly run the policy and ask the human participants to answer the questionnaire (Table I) after each run.

In Table II we show the results of running the MPC policy against all of our five social scenarios in both of the defined locations (Fig. 3). For this experiment, we recorded 10 runs for each social scenario in both of the defined locations. We report the per question average (QAVG), as well as the overall average (SAVG) along with the standard deviation for each scenario. For negatively formulated questions (labeled with a '*' in Table I) we reverse coded the ratings to make them comparable to the positively formulated ones. Across the different scenarios we obtain similar average ratings. This suggests that our questionnaire based metric can be used to obtain meaningful results for different social scenarios.

	Frontal Approach					Intersection Wait				Intersection Gesture				Narrow Doorway				Blind Corner			
Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Location	F1					W1				G1				N1				B1			
QAVG	3.1	2.6	3.2	2.7	2.7	2.2	2.4	2.4	1.3	2.1	3.8	2.2	2.1	4.1	2.2	2.0	4.4	3.4	2.0	2.9	2.0
SAVG	2.9					2.1				2.6				1.9				3.1			
STD	1.7					1.4				0.6				1.4				1.5			
Location	F2					W2				G2				N2				B2			
QAVG	2.7	3.6	3.3	2.7	3.0	2.5	2.8	2.2	1.2	3.2	1.8	1.4	2.0	3.7	1.6	1.3	4.3	1.9	2.6	3.2	2.7
SAVG	2.9					1.3				2.1				1.7				2.7			
STD	1.7					1.3				1.2				1.2				1.5			

TABLE II

FIVE SOCIAL SCENARIOS EACH DEFINED FOR TWO LOCATIONS.

	Participant 1					Participant 2					Participant 3				
Question	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
QAVG	2.8	3.8	2.0	2.8	3.4	2.6	3.3	2.0	3.5	2.2	2.6	2.9	2.6	2.9	2.5
SAVG	2.5					2.7					2.8				
STD	1.3					1.1					1.1				

TABLE III

FRONTAL APPROACH: THREE DIFFERENT PARTICIPANTS.

	Set 1					Set 2					Set 3				
Question	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
QAVG	2.9	2.2	2.0	2.0	3.4	2.6	3.1	2.6	1.0	2.5	2.3	3.0	2.4	3.0	2.3
SAVG	2.5					2.4					2.6				
STD	1.3					1.1					1.0				

TABLE IV

FRONTAL APPROACH: THREE SETS.

	Policy 1 (MPC)					Policy 2 (MPC + BC)				
Question	1	2	3	4	5	1	2	3	4	5
QAVG	2.3	4.3	2.0	3.9	2.6	4.6	1.7	4.3	1.6	4.3
SAVG	2.2					4.3				
STD	1.1					1.2				

TABLE V

FRONTAL APPROACH: TWO POLICIES.

Moreover, we observed that the measurements for the same social scenario in different locations corresponded with each other. This indicates that our setup can be replicated for the same social scenarios in other locations (e.g. other labs) or for other social scenarios. One run of a social scenario commonly takes 15-45 seconds, while answering the 4-5 questions for each scenario requires 20-40 seconds.

To test our setup for individual human biases, we ran the same Location (F2) frontal approach scenario 30 times with three different participants. Table III shows that for each participant we obtained similar average ratings for each question and for the entire social scenario. This suggests that our questionnaire-based metric provides good inter-rater reliability.

To measure the variance of our setup when running the same experiment at different dates and times, we ran the Location (F2) frontal approach scenario 90 times with the same human participant. Table IV shows similar average ratings and standard deviations across the different runs of the same social scenario. This suggests that our questionnaire based metric can be used to obtain reliable measurements across different validation runs. Furthermore, this suggests that it may not be necessary to capture large quantities of runs to obtain reliable results but that batches of up to 30 runs already provide meaningful results.

To test our benchmark for different polices we compare the MPC policy with a behavioral cloning (BC) policy. Here we captured 300 trajectories of expert data of the frontal approach scenario and trained a convolutional neural network to predict intermediate waypoints for the MPC policy. We use this policy to generate socially-compliant behavior. Table V shows the average questionnaire ratings of 20 runs of each policy. These results indicate that we are able to measure the capabilities of this more advanced policy compared to the common MPC policy (indicated by the higher averages).

Fig. 5. Validation run of the blind corner scenario. Robot and human are moving in opposite directions around a blind corner on a colliding path. The MPC policy is not able to anticipate the human interaction causing a collision – the robot drove onto the person's foot. Consequently the rating for this social interaction was: Q18=strongly disagree, Q19=disagree, Q20=neutral, Q21=strongly agree.

In Fig. 5 we show a validation run of our blind corner scenario that highlights the advantages of obtaining in-situ ratings from human participants. For this run, the robot briefly collided with the human, touching the person's foot. While this event is hardly noticeable for an external observer or in an ex-situ setting (Fig. 5, right) it generated a clear negative response when the participant provided their ratings.

VI. CONCLUSION

We have introduced a novel protocol for establishing a benchmark for social navigation scenarios. The proposed approach is based on defining a set of common social interactions that occur for navigation tasks. Each social scenario is defined in a canonical manner to support the repeated validation of policies. Additionally, we have proposed and piloted a questionnaire-based metric to obtain in-situ ratings of human participants that allow us to assess the social compliance of navigation policies. We only rely on a lightweight specification for each social scenario. Therefore, our benchmark and our questionnaire can readily be extended by additional scenarios. As future work, we plan to extend our benchmark and to use it to validate existing and novel socially-compliant navigation policies.

VII. ACKNOWLEDGMENTS

We thank our robot operators April Zitkovich, Jake Lee, Khem Holden, Rosario Jauregui Ruano, and Diego Reyes for their diligent work collecting all social navigation data samples reported in the paper.

REFERENCES

[1] P. Althaus, H. Ishiguro, T. Kanda, T. Miyashita, and H.I. Christensen. Navigation for human-robot interaction tasks. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, volume 2, pages 1894–1900 Vol.2, 2004.

[2] 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. *J. Hum.-Robot Interact.*, 9(4), sep 2020.

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, Jan 2009.

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

[5] Yuxiang Gao and Chien-Ming Huang. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI*, 8, 2022.

[6] Chien-Ming Huang, Takamasa Iio, Satoru Satake, and Takayuki Kanda. Modeling and controlling friendliness for an interactive museum robot. 07 2014.

[7] Michiel Joosse, Manja Lohse, Niels Van Berkel, Aziez Sardar, and Vanessa Evers. Making appearances: How robots should approach people. *J. Hum.-Robot Interact.*, 10(1), jan 2021.

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

[9] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. May i help you? design of human-like polite approaching behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 35–42, New York, NY, USA, 2015. Association for Computing Machinery.

[10] Kapil D. Katyal, Katie Popek, Gregory D. Hager, I-Jeng Wang, and Chien-Ming Huang. Prediction-based uncertainty estimation for adaptive crowd navigation. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings*, page 353–368, Berlin, Heidelberg, 2020. Springer-Verlag.

[11] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, PP:1–15, 04 2021.

[12] 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.

[13] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[14] Luis J Manso, Pedro Nuñez, Luis V Calderita, Diego R Faria, and Pilar Bachiller. Socnav1: A dataset to benchmark and learn social navigation conventions. *Data*, 5(1):7, 2020.

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

[16] Reuth Mirsky, Xuesu Xiao, Justin Hart, and Peter Stone. Prevention and resolution of conflicts in social navigation–a survey. *arXiv preprint arXiv:2106.12113*, 2021.

[17] Mehdi Moussaïd, Dirk Helbing, Simon Garnier, Anders Johansson, Maud Combe, and Guy Theraulaz. Experimental study of the behavioural mechanisms underlying self-organization in human crowds. *Proceedings of the Royal Society B: Biological Sciences*, 276(1668):2755–2762, 2009.

[18] Elena Pacchierotti, Henrik Christensen, and Patric Jensfelt. Embodied social interaction for service robots in hallway environments. volume 25, pages 293–304, 01 2005.

[19] Daniel Perille, Abigail Truong, Xuesu Xiao, and Peter Stone. Benchmarking metric ground navigation. In *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 116–121. IEEE, 2020.

[20] Jorge Rios-Martinez, Alessandro Renzaglia, Anne Spalanzani, Agostino Martinelli, and Christian Laugier. Navigating between people: A stochastic optimization approach. In *2012 IEEE International Conference on Robotics and Automation*, pages 2880–2885, 2012.

[21] J. Rios-Martinez, A. Spalanzani, and C. Laugier. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, 7(2):137–153, Apr 2015.

[22] Roya Salek Shahrezaie, Santosh Balajee Banisetty, Mohammadmahdi Mohammadi, and David Feil-Seifer. Towards deep reasoning on social rules for socially aware navigation. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 515–518, New York, NY, USA, 2021. Association for Computing Machinery.

[23] Xuan-Tung Truong and Trung-Dung Ngo. "to approach humans?": A unified framework for approaching pose prediction and socially aware robot navigation. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):557–572, 2018.

[24] Nathan Tsoi, Mohamed Hussein, Jeacy Espinoza, Xavier Ruiz, and Marynel Vázquez. Sean: Social environment for autonomous navigation. In *Proceedings of the 8th International Conference on Human-Agent Interaction*, pages 281–283, 2020.

[25] Araceli Vega, Luis J. Manso, Douglas G. Macharet, Pablo Bustos, and Pedro Núñez. Socially aware robot navigation system in human-populated and interactive environments based on an adaptive spatial density function and space affordances. *Pattern Recognition Letters*, 118:72–84, 2019. Cooperative and Social Robots: Understanding Human Activities and Intentions.

[26] Xuesu Xiao, Bo Liu, Garrett Warnell, and Peter Stone. Motion planning and control for mobile robot navigation using machine learning: a survey. *Autonomous Robots*, pages 1–29, 2022.