Benchmarking Social Robot Navigation Across Academia and Industry

Anthony Francis,[1,*,†] **Claudia Pérez-D'Arpino,**[2,*] **Chengshu Li,**[3] **Fei Xia,**[4]
Alexandre Alahi,[5] **Aniket Bera,**[6] **Abhijat Biswas,**[7] **Joydeep Biswas,**[8] **Hao-Tien Lewis Chiang,**[4]
Michael Everett,[9] **Sehoon Ha,**[10] **Justin Hart,**[8] **Haresh Karnan,**[8] **Tsang-Wei Edward Lee,**[4]
Luis J. Manso,[11] **Reuth Mirksy,**[12] **Sören Pirk,**[13] **Phani Teja Singamaneni,**[14]
Peter Stone,[8,15] **Ada V. Taylor,**[7] **Peter Trautman,**[16] **Nathan Tsoi,**[17]
Marynel Vázquez,[17] **Xuesu Xiao,**[18] **Peng Xu,**[4] **Naoki Yokoyama,**[10]
Roberto Martín-Martín,[8‡] **Alexander Toshev**[19‡]

[1]Logical Robotics [2]NVIDIA [3]Stanford [4]Google [5]EPFL [6]Purdue [7]CMU [8]UT Austin [9]Northeastern [10]Georgia Tech [11]Aston
[12]Bar Ilan [13]Adobe [14]LAAS-CNRS [15]Sony AI [16]Honda [17]Yale [18]GMU [19]Apple
Corresponding Author: centaur@logicalrobotics.com

Abstract

A challenge to deploying robots widely is navigation in human-populated environments, or *social robot navigation*. Fairly evaluating social navigation algorithms is hard because it involves not simply robotic agents in static environments but also dynamic human agents. Further, human perceptions of robot behavior are important when evaluating social robot navigation. While reliable benchmarks have accelerated progress in fields like computer vision and language understanding, much work remains to be done to effectively benchmark social navigation due to several open challenges. For example, in social robot navigation, traditional quantitative measures of navigation algorithm performance must be augmented with user studies of human reactions to robot behaviors. Also, work in human-robot interaction (HRI) in industry and academia have different emphases, often leading to different evaluation techniques. In this paper, we review ongoing efforts to develop principles and guidelines for benchmarking social navigation and use these to analyze existing benchmarks with respect to HRI concerns.

The goal of social robot navigation research is to improve how robots behave when moving around people. However, a crisp definition of what makes navigation "social" is elusive. To address this gap, the authors - a workgroup of researchers brought together by the Social Navigation Symposium (Google, Stanford 2022) - have been developing principles and guidelines for benchmarking social robot navigation algorithms. Our discussions have clarified the social navigation problem (Sec. 1) and how it is analyzed scientifically (Sec. 2), developed a taxonomy of social navigation experimental setups, metrics, simulators, datasets, and deployment environments (Sec. 3), and are converging on recommendations to make evaluations more comparable (Sec. 4).

[*]These authors contributed equally.
[†]Anthony Francis's contributions began while at Google.
[‡]Equal advising contribution.

Emerging from these discussions is an understanding that, unlike traditional robot navigation, social navigation needs human-robot interaction (HRI) methods to analyze human behavior. We review existing benchmarks including academic benchmarks that enable detailed tests of behaviors in simulation, and industry benchmarks that analyze human reactions to social robot behavior (Sec. 5). Finally, we conclude by arguing social robot navigation benchmarks should incorporate elements of both of these approaches (Sec. 6).

1 Towards a Definition of Social Navigation

Social navigation has referred to a range of behaviors from simple navigation around dynamic obstacles, to complying with complex social norms, up to navigating with communicative intent. To define "social" more precisely, we examined the terms social and antisocial for humans. Social sometimes means participating in society, i.e., participating in an interacting group whose individuals modify their behavior to accommodate the needs of others while achieving their own. But social has a second meaning: a "social" individual has outstanding skills to work with others, based on an understanding of their feelings and needs and adapting to them. Antisocial individuals fail to follow the customs of society or live without consideration for others. Inspired by these terms when applied to humans, we offer this definition:

> A *socially navigating robot* is *a robot that acts and interacts with humans or other robots, achieving its navigation goals while modifying its behavior to enable the other agents to better achieve theirs*.

This social quality may be reflected through overt behavior changes, such as respecting social norms, or through an understanding of other agents' needs, feelings, and/or communicating capabilities. However, social norms are often not verbalized, and what other agents need to achieve, what they feel, or what they like can be unclear. To operationalize these concerns, we identified *aspects of social navigation* that can be used to evaluate the quality of social behavior, including safety, comfort, legibility, politeness, social competency, understanding other agents, and responding to context:

- **Safety:** A minimal requirement for robots and human sociality is not harming others in the course of business. Avoiding collisions with humans is important but is not the only safety concern (Mavrogiannis et al. 2021). For instance, while it might be acceptable for a factory robot to bump a guardrail defining the edge of its workspace, social robots should avoid damaging human environments, which often contain important objects that can be damaged or wall coverings whose visual appearance is important. Robots should also avoid behaving in a way that induces a human to injure themselves.
- **Comfort:** Humans should also feel comfortable around robots, defined in (Kruse et al. 2013) as *the absence of annoyance and stress for humans in interaction with robots*. Many features contribute to comfort, including maintaining human-robot distance, not cutting humans off, and naturalness of motion. Unacceptable robot speed, navigation jitter and unexpected head movements are factors that degrade humans' perception of comfort.
- **Legibility:** Legibility refers to the property of an agent's behavior that makes it possible for other agents to infer their goals (Dragan, Lee, and Srinivasa 2013). This includes not only the robot's goal but also incidental interactions when performing other tasks, e.g., moving to the right or left when passing in a hallway. (Dragan, Lee, and Srinivasa 2013) suggests that legibility involves relaxing constraints such as predictability of trajectories.
- **Politeness:** Politeness refers to behavior that is respectful and considerate of people. There are at least two dimensions: physical politeness (how robots navigate around people, such as not cutting people off) and communicative politeness (gestures or verbal signals, such as saying "excuse me"). Politeness can have a strong effect on people's perception of robots (Inbar and Meyer 2019).
- **Social Competency:** Robots should comply with social, political, and legal norms for sharing space. Many social competencies are matters of following conventions rather than optimizing performance (e.g., driving on the left or right). Some social competencies, like turn-taking, can emerge naturally (Kose-Bagci, Dautenhahn, and Nehaniv 2008), whereas others must be engineered (or simulated with wizard-of-oz studies (Kanda and Ishiguro 2017)).
- **Understanding Other Agents:** Accommodating other agents requires understanding what they want, what they're doing and where they're going, so robots don't thwart them by crossing paths or blocking goals. Understanding when the *interaction potential* - the likelihood of robots entering human personal space - should be minimized (Trautman and Krause 2010) or maximized (Mead and Matarić 2017) depends on the task.
- **Responding Appropriately to Context:** Social navigation should be evaluated within the context that it is to be deployed. Context helps us understand the relative importance of the previous objectives. An example shared in the symposium was a robot in a hospital bringing an emergency drug to a doctor: politeness is less important than task success. We identified several forms of context, including cultural context, environmental context, diversity, task context, and interpersonal context, all of which can change which response is right in a given situation.

These aspects of social navigation can be seen as additional objectives that the agent needs to optimize for while still achieving its main objective, and context can be seen as weighting which aspects are most important at any given time. These aspects are not completely orthogonal: improving legibility might improve safety and even comfort, whereas nonverbal politeness depends on understanding other agents' trajectories. As another example, what is considered appropriate or polite behavior depends on the cultural context (Recchiuto and Sgorbissa 2022).

Properly studying these aspects of social navigation directly impacts which metrics to measure (Gao and Huang 2021), what datasets to collect, how to build simulators, and how to structure benchmarks. However, another key factor is understanding what kind of social navigation research is being conducted, for algorithm development in academia and product development in the industry have different concerns.

2 Scientific Questions of Social Navigation

Benchmarks require measures by which we compare and an evaluation methodology for making the comparison. Because social robot navigation is concerned with methods for controlling mobile robots to operate effectively around people, many argue social robot evaluation methodologies should involve the collection of human perception of robot behavior. However, the scientific questions a benchmark asks are equally important in deciding its scope and content.

The key scientific question asked in benchmarking social robot navigation is how different social navigation approaches compare to each other. Given the complexity of social navigation, different benchmarks often focus on different aspects of the problem and thus different scientific questions. Some of these questions arise from traditional robot navigation research and can arguably be evaluated using traditional methods, with adjustments for human participants:

- **Method Evaluation:** How do methods compare with each other against baselines? Some aspects of method evaluation involve quantitative metrics and can be performed in simulation, such as revealing problems in a robot's safety layer as it faces increasing obstacle densities. However, when human evaluations are required, these are typically conducted in the real world, though toolkits are now coming into use that enable labeling simulated trajectories as well (Baghel et al. 2021).
- **Ablation Studies:** How do components of a method affect its overall performance? These are generally conducted by turning method components off. While in theory ablation studies could be conducted on-robot, in practice these studies are often only conducted in simulation, as real human participant time can be wasted on variants of the algorithm expected to perform poorly (or known to perform poorly in simulation).

- **Method Generalization:** How do policy behaviors generalize to different environments? Benchmarks can test methods under different conditions to evaluate this, a task which is easier (though less realistic) in simulation.

However, while physical safety may be measurable quantitatively, politeness seems to imply human evaluation, and comfort is often defined in terms of human reactions. Scientific questions involving these aspects involve human perceptions and reactions to robot behaviors, and should arguably be investigated through HRI studies:

- **Human Ratings:** How do humans rate the socialness of social navigation methods, either intrinsically or in comparison to a baseline? For some researchers, human ratings of policy behavior in real contexts is the gold standard for policy performance, but for these ratings to be effective, studies must follow proper HRI protocols and use validated survey instruments.
- **Behavior Analysis:** How does human behavior change when exposed to different robot navigation policies? While ratings are explicit, behavior change is implicit or even unconscious, and studies must be conducted according to HRI guidelines that ensure conditions are appropriately blinded so participant and rater reactions are valid.
- **Issue Discovery:** Benchmarks can also be used to find out the frequency of encounter types between humans and robots, as well as the frequency of problems that affect a given policy, to guide research in the direction of problems that occur in the wild. These types of studies must be conducted on robot in a live deployment.

Many benchmarks focus on a subset of the questions above. However, in Section 6 we argue that because social navigation involves understanding both how robots affect other agents and which methods are effective, benchmarks will benefit from incorporating both HRI components that evaluate human reactions in the real, as well as ablation studies even if those cannot be collected in the real.

3 A Taxonomy of Social Navigation

Our workgroup does not aim to provide a single comprehensive evaluation protocol to be used for all social navigation research, because social navigation research spans a wide range of domains with different goals and priorities. Instead, we are developing principles and guidelines to guide the development of protocols that enable fair comparisons of methods. The first step in this process involves scoring social navigation approaches along a formal set of axes including factors of analysis, metrics collected, dataset types, simulator platforms, social scenarios, and overall benchmarks.

- **Factors:** Benchmarks, datasets, and simulators for social navigation have similar properties, and we have identified common factors to create a vocabulary for analysis. The social navigation scenario includes the context of the navigation, the physical environment where robots and humans operate, the human behaviors expected, the robot task being performed, and the role of the robot. The robot platform is another key factor, including its morphology, sensors, actuators, and communication modalities. Data and metrics collected are additional factors, along with methods for authoring human and robot behaviors, and whether experiments are conducted in simulation or real.
- **Metrics:** Recent surveys of social navigation metrics have uncovered close to a hundred different metrics in use (see (Gao and Huang 2021; Mirsky et al. 2021) for recent reviews). However, these metrics can be grouped by broader features, such as algorithmically computed (such as Success-weighted Path Length (SPL) (Anderson et al. 2018)) or gathered by surveys of humans (such as comfort ratings). Surveyed metrics can further be divided into those collected explicitly via questionnaires (e.g., about perceived comfort) or implicitly through sensors (such as stress in human facial expressions). Algorithmic metrics in turn can be hand-crafted or learned from data gathered from surveys. Other axes of metrics include whether metrics cover moment-to-moment behavior or entire episodes, or whether metrics model individual features (e.g., success rate or politeness) or holistic performance (e.g., ratings of "socialness").
- **Datasets:** We have used these factors to analyze datasets such as JRDB (Martín-Martín et al. 2019), THOR (Rudenko et al. 2020), TRAJNET++ (Kothari, Kreiss, and Alahi 2021), UCY (Lerner, Chrysanthou, and Lischinski 2007), ETH (Pellegrini et al. 2009), EPID (Majecka 2009), SDD (Robicquet et al. 2016), EFL) (Park et al. 2016), WILDTRACK (Chavdarova et al. 2018), SCAND (Karnan et al. 2022a,b). We will not try to summarize our findings, but note datasets require additional factors for analysis such as coverage, sampling distribution, annotations, and privacy and fairness handling.
- **Simulators:** Social navigation simulators include iGibson (Li et al. 2021), SEAN (Tsoi et al.), and SocNavBench (Biswas et al. 2022). Many simulators have different APIs and metrics. To enable clearer comparison across simulation environments, we are working to create a common API shared across all simulators.
- **Scenarios:** Social navigation studies include field studies of behavior in the wild, long-term robot deployments at particular sites, controlled laboratory experiments, social navigation scenarios that aim to create a particular in-the-wild behavior, and "staged" scenarios that attempt to recreate the chaos of crowd scenarios. We have developed a "scenario card" which enables us to compare social navigation scenarios, such as frontal approach, overtaking, intersections, and blind corner.
- **Benchmarks:** Benchmarks involve an evaluation protocol for collecting metrics for social robot navigation methods in social navigation scenarios. Current benchmarks include SEANavBench,[1] iGibson (Li et al. 2021; Shen et al.), and the Social Scenarios Protocol (Pirk et al. 2022), discussed in further depth in Section 5.

4 Criteria for Good Benchmarks

Based on our analysis of social navigation aspects, factors, datasets, simulators, and metrics, we argue there are 6 key

[1] https://seanavbench.interactive-machines.com/

factors to make good social navigation benchmarks:

1. **Evaluate Social Behavior:** A good benchmark should evaluate the properties of algorithms in social scenarios which involve humans and robots interacting. Therefore, a social benchmark should have metrics related to social behavior and not just contain pure navigation metrics such as Success-weighted by Path Length (Anderson et al. 2018) or pure task metrics such as success rates.
2. **Include Quantitative Metrics:** Benchmarks should provide quantitative metrics on a variety of dimensions of interest, enabling researchers with different goals to use them to evaluate algorithms with respect to their context while comparing with other approaches in the literature.
3. **Provide Baselines for Comparison:** Benchmarks should include baseline policies that show worst-case performance; comparisons with upper-bound oracle performance or state-of-the-art policies are desirable as well.
4. **Be efficient, repeatable, and scalable:** To allow for better democratization and productive competition and collaboration amongst different scientists, efficient, repeatable, and scalable benchmarks are preferable.
5. **Ground Human Evaluations in Human Data:** Many researchers agree that we do not have a good predictive model of how humans will react to or rate robot behavior; therefore, benchmarks should measure socialness based on data gathered from human evaluations.
6. **Use well-validated evaluation instruments:** These human evaluations should be based on metrics that are validated, which is an iterative process that involves proposing metrics, conducting studies, statistically analyzing responses, and exposing metrics to peer review.

5 Current Social Navigation Benchmarks

Existing benchmarks often focus on one or more of the aspects of social navigation outlined above, and perform better or worse on the criteria we have outlined. In this section, we review benchmarks in use in academia and industry.

Academic Benchmarks

SEANavBench is an academic benchmark created for the SEANavBench workshop and competition held at ICRA'22. SEANavBench combines SocNavBench (Biswas et al. 2022) and SEAN 2.0 (Tsoi et al.), enabling algorithms to run on simulated robots via ROS in environments via Unity. Algorithms can be evaluated on simulated scenes across environment sizes, crowd densities, and pedestrian behavior, including simulated pedestrians and replay of pedestrian datasets. This enables the analysis of how algorithms can succeed or fail as environmental conditions change and the measurement of performance using a variety of metrics. SEANavBench uses SEAN-EP (Tsoi et al. 2021) to run the SEAN 2.0 simulation environment on the web, which could be used to collect human feedback.

iGibson (Li et al. 2021; Shen et al.) is a simulation environment for navigation and manipulation tasks in household scenes, used to create the iGibson Challenge 2021 social navigation benchmark [2] at the CVPR 2021 Embodied AI Workshop. In this benchmark, robots must navigate to targets without collision among pedestrians (Pérez-D'Arpino et al. 2021), which are simulated via the ORCA model (van den Berg et al. 2011) in fifteen interactive indoor household scenes. Evaluation metrics include STL (Success weighted by Time Length) for reaching the goal quickly, and PSC (Personal Space Compliance) for maintaining a comfortable distance from all pedestrians. This benchmark enabled quantitative comparison of approaches from over a dozen teams, including methods based on techniques like DD-PPO (Wijmans et al. 2019), PPO (Schulman et al. 2017), SAC (Haarnoja et al. 2018), and so on, providing a clear picture of which algorithms were superior for the task. However, the benchmark does not include human ratings, and in 2021 did not include on-robot tests.

Industry Benchmarks

The Social Navigation Scenarios Protocol (Pirk et al. 2022) is an industry benchmark proposed by Robotics at Google (Google 2023) and used in (Pirk et al. 2022), (Xiao et al. 2022), and (Cuan et al. 2022) to evaluate the performance of a series of learning-based model predictive control policies (MPC) for social robot navigation. This benchmark protocol involves selecting social navigation scenarios of interest, such as Frontal Approach, Blind Corner, Corridor Intersection, and so on. Each scenario's human-robot interaction is defined by the start and end of the robot trajectory and a short description of what is expected to happen for the human. This serves two purposes: enabling the collection of expert human trajectories for training social navigiation policies, and evaluating policies on the same scenarios with relatively low variability. Over the course of (Pirk et al. 2022), (Xiao et al. 2022), and (Cuan et al. 2022), the protocol was iteratively improved. For example, the questionnaire proposed in (Pirk et al. 2022) was analyzed in (Xiao et al. 2022) to identify reliable factors according to Cronbach's alpha, which were used to update the questionnaire for (Cuan et al. 2022), which enabled more extensive analysis.

6 HRI and Social Navigation

HRI provides a wide array of experimental protocols and tools for validating survey instruments, which can help the social navigation community to improve its tools. However, social navigation has its own unique challenges. In this section, we review techniques typically used in industry and academia, with the caveat that this characterization of what is "industrial" or "academic" is really a caricature, and many researchers cross what we are presenting as a divide.

HRI for Social Navigation in Industry

HRI in social robotics for industry often refers to studying human-robot interactions to improve outcomes of robots interacting with people. HRI studies help develop research ideas into products, guiding the development of "experiences" that focus less on improving performance and more

[2] https://svl.stanford.edu/igibson/challenge2021.html

on creating interactions that enable robots and humans to accomplish more together. Industry can deploy robots at scales difficult for academia, but deployments have legal, privacy, and safety issues as serious as the constraints of IRB review. Industry can even conduct large-scale studies to validate survey instruments, but cannot always control potential confounds like the robot platform or deployment environment.

Still, techniques from academic HRI are useful to industry. HRI helps design benchmarks and A/B tests built around questionnaires so these instruments are valid. For example, industry often measures success by A/B tests of social navigation approaches evaluated by raters. HRI studies show, however, that ratings provided by roboticists differ from those of laypeople, so drawing raters from the expected user pool is critical, and proper study protocols must be followed.

HRI and Social Navigation in Academia

Like their counterparts in industry, HRI researchers in academia study how to make humans and robots interact better. Academia often focuses more closely on developing reliable, validated scientific techniques for analyzing how to benchmark those interactions. For this reason, industry often relies on academic researchers to advise its HRI efforts.

But academic social robotics is more than just HRI research, as many social navigation researchers are interested in understanding the performance of methods to improve them. Academia often focuses on studies that vary environment properties like crowd density or algorithm properties via ablations to illuminate the sources of power. Industry deployments can challenge these improvements in the real.

Bridging Academic and Industrial HRI

While some techniques for social robot navigation benchmarking are more frequently used in industry and others more frequently in academia, we argue ideal benchmarks should build on techniques from both traditions.

Industry already benefits from academia, using social navigation methods from academia as building blocks for social navigation solutions. However, even if a method does well on a human-rated A/B test, without ablation studies that reveal how these methods cause good (or bad) social behavior, it's hard to know how these methods could be improved.

Conversely, academic social robotics might benefit from more focus on the human part of HRI. While studying the sources of power of algorithms with ablation studies and a diversity of environments is important, arguably the final arbiter of good performance is how real humans react to the methods when deployed in real on-robot deployments.

Another tradeoff exists between controlled studies and the wild. To create repeatable studies, many academic researchers simplify problems until they can be tested in a controlled laboratory setting. By design, this reduces the chance of discovering interesting behaviors in the wild. Conversely, deployments in the wild can discover unexpected outcomes, but require a large number of trials to function.

Overall, benchmarks that combined these approaches would be ideal. For example a benchmark that included a simulated component which enabled collecting a broad variety of data on algorithm ablations and environment conditions would enable us to understand the behavior of algorithms, even if it was not feasible to collect in the wild - for example, in simulation one can test what would happen if one disabled the safety layer of an algorithm in a dense crowd. But hand in hand with this simulated component should be a real component which enables A/B testing of qualities which are highly performant enough to be deployed safely on-robot, using validated questionnaires that enable reliable collection of human ratings. With both sets of data, we can make predictions about how algorithm changes would make humans feel about the sociability of navigation behavior - predictions we cannot make given benchmarks that address only algorithmic or human data in isolation.

7 Conclusion

The benchmarks discussed reveal complementary strengths in academic and industry studies of social robot navigation. Academic social benchmarking has developed extensive techniques for analysis of social navigation policies, including environmental diversity to determine generalization and ablation studies to reveal sources of power; industry benchmarking has focused on on-robot testing which has required incorporating techniques from HRI to create validated instruments which can enable reliable comparisons of policies. The ideal social navigation benchmark would combine features of both approaches. For these reasons, our criteria for good benchmarks - evaluating social behavior, with quantitative metrics, with baselines for comparison, scalably, with human evaluations grounded in human data collected with validated survey instruments - are designed to encourage researchers to create the best of both worlds.

References

Anderson, P.; Chang, A.; Chaplot, D. S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; et al. 2018. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Baghel, R.; Kapoor, A.; Bachiller, P.; Jorvekar, R. R.; Rodriguez-Criado, D.; and Manso, L. J. 2021. A toolkit to generate social navigation datasets. In *Advances in Physical Agents II: Proc. of the 21st International Workshop of Physical Agents (WAF 2020)*, 180–193. Springer.

Biswas, A.; Wang, A.; Silvera, G.; Steinfeld, A.; and Admoni, H. 2022. Socnavbench: A grounded simulation testing framework for evaluating social navigation. *THRI*.

Chavdarova, T.; Baqué, P.; Bouquet, S.; Maksai, A.; Jose, C.; Bagautdinov, T.; Lettry, L.; Fua, P.; Van Gool, L.; and Fleuret, F. 2018. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *CVPR*.

Cuan, C.; Lee, E.; Fisher, E.; Francis, A.; Takayama, L.; Zhang, T.; Toshev, A.; and Pirk, S. 2022. Gesture2Path: Imitation Learning for Gesture-aware Navigation. *arXiv preprint arXiv:2209.09375*.

Dragan, A. D.; Lee, K. C.; and Srinivasa, S. S. 2013. Legibility and predictability of robot motion. In *HRI*. IEEE.

Gao, Y.; and Huang, C.-M. 2021. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI*, 420.

Google. 2023. Robotics @ Google website. https://research.google/teams/robotics/. [Online; accessed 29-Jan-2023].

Google, Stanford. 2022. Social Navigation Symposium. https://sites.google.com/view/socialnavigationsymposium/home. Accessed: 2023-01-28.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*. PMLR.

Inbar, O.; and Meyer, J. 2019. Politeness counts: Perceptions of peacekeeping robots. *IEEE Transactions on Human-Machine Systems*, 49(3): 232–240.

Kanda, T.; and Ishiguro, H. 2017. *Human-robot interaction in social robotics*. CRC Press.

Karnan, H.; Nair, A.; Xiao, X.; Warnell, G.; Pirk, S.; Toshev, A.; Hart, J.; Biswas, J.; and Stone, P. 2022a. Socially Compliant Navigation Dataset (SCAND).

Karnan, H.; Nair, A.; Xiao, X.; Warnell, G.; Pirk, S.; Toshev, A.; Hart, J.; Biswas, J.; and Stone, P. 2022b. Socially CompliAnt Navigation Dataset (SCAND): A Large-Scale Dataset Of Demonstrations For Social Navigation. *RA-L*.

Kose-Bagci, H.; Dautenhahn, K.; and Nehaniv, C. L. 2008. Emergent dynamics of turn-taking interaction in drumming games with a humanoid robot. In *RO-MAN*. IEEE.

Kothari, P.; Kreiss, S.; and Alahi, A. 2021. Human Trajectory Forecasting: A Deep Learning Perspective. *IEEE Transactions on Intelligent Transportation Systems*.

Kruse, T.; Pandey, A. K.; Alami, R.; and Kirsch, A. 2013. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12): 1726–1743.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer graphics forum*, volume 26, 655–664. Wiley Online Library.

Li, C.; Xia, F.; Martín-Martín, R.; Lingelbach, M.; Srivastava, S.; Shen, B.; Vainio, K.; Gokmen, C.; Dharan, G.; Jain, T.; et al. 2021. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*.

Majecka, B. 2009. Statistical models of pedestrian behaviour in the forum. *Master's thesis, School of Informatics, University of Edinburgh*.

Martín-Martín, R.; Rezatofighi, H.; Shenoi, A.; Patel, M.; Gwak, J.; Dass, N.; Federman, A.; Goebel, P.; and Savarese, S. 2019. Jrdb: A dataset and benchmark for visual perception for navigation in human environments. *arXiv preprint arXiv:1910.11792*.

Mavrogiannis, C.; Baldini, F.; Wang, A.; Zhao, D.; Trautman, P.; Steinfeld, A.; and Oh, J. 2021. Core challenges of social robot navigation: A survey. *arXiv preprint arXiv:2103.05668*.

Mead, R.; and Matarić, M. J. 2017. Autonomous human–robot proxemics: socially aware navigation based on interaction potential. *AuRo*, 41(5): 1189–1201.

Mirsky, R.; Xiao, X.; Hart, J.; and Stone, P. 2021. Prevention and Resolution of Conflicts in Social Navigation–a Survey. *arXiv preprint arXiv:2106.12113*.

Park, H. S.; Hwang, J.-J.; Niu, Y.; and Shi, J. 2016. Egocentric future localization. In *CVPR 2016*, 4697–4705.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, 261–268. IEEE.

Pérez-D'Arpino, C.; Liu, C.; Goebel, P.; Martín-Martín, R.; and Savarese, S. 2021. Robot Navigation in Constrained Pedestrian Environments using Reinforcement Learning. *ICRA 2021*.

Pirk, S.; Lee, E.; Xiao, X.; Takayama, L.; Francis, A.; and Toshev, A. 2022. A protocol for validating social navigation policies. *arXiv preprint arXiv:2204.05443*.

Recchiuto, C.; and Sgorbissa, A. 2022. Diversity-aware social robots meet people: beyond context-aware embodied AI. *arXiv preprint arXiv:2207.05372*.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, 549–565. Springer.

Rudenko, A.; Kucner, T. P.; Swaminathan, C. S.; Chadalavada, R. T.; Arras, K. O.; and Lilienthal, A. J. 2020. THÖR: Human-Robot Navigation Data Collection and Accurate Motion Trajectories Dataset. *RA-L*, 5(2): 676–682.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shen, B.; Xia, F.; Li, C.; Martín-Martín, R.; Fan, L.; Wang, G.; Pérez-D'Arpino, C.; Buch, S.; Srivastava, S.; Tchapmi, L.; et al. ???? iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IROS 2021*.

Trautman, P.; and Krause, A. 2010. Unfreezing the robot: Navigation in dense, interacting crowds. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 797–803. IEEE.

Tsoi, N.; Hussein, M.; Fugikawa, O.; Zhao, J. D.; and Vázquez, M. 2021. An Approach to Deploy Interactive Robotic Simulators on the Web for HRI Experiments: Results in Social Robot Navigation. In *IROS 2021*. IEEE.

Tsoi, N.; Xiang, A.; Yu, P.; Sohn, S. S.; Schwartz, G.; Ramesh, S.; Hussein, M.; Gupta, A. W.; Kapadia, M.; and Vázquez, M. ???? Sean 2.0: Formalizing and generating social situations for robot navigation. *RA-L*, 7(4).

van den Berg, J.; Guy, S. J.; Snape, J.; Lin, M. C.; and Manocha, D. 2011. Rvo2 library: Reciprocal collision avoidance for real-time multi-agent simulation. *See https://gamma. cs. unc. edu/RVO2*.

Wijmans, E.; Kadian, A.; Morcos, A.; Lee, S.; Essa, I.; Parikh, D.; Savva, M.; and Batra, D. 2019. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*.

Xiao, X.; Zhang, T.; Choromanski, K. M.; Lee, T.-W. E.; Francis, A.; Varley, J.; Tu, S.; Singh, S.; Xu, P.; Xia, F.; Persson, S. M.; Takayama, L.; Frostig, R.; Tan, J.; Parada, C.; and Sindhwani, V. 2022. Learning Model Predictive Controllers with Real-Time Attention for Real-World Navigation. In *CORL*. PMLR.