Toy Story[1] Method for Multi-Policy Social Navigation

Abhisek Konar
Samsung AI Center
Montreal
abhisek.k@partner.samsung.com

Bobak H. Baghi
Samsung AI Center
Montreal
bobak.h@samsung.com

Francois R. Hogan
Samsung AI Center
Montreal
f.hogan@samsung.com

Gregory Dudek
Samsung AI Center
Montreal
greg.dudek@samsung.com

***Abstract*—In this work, we present a method for effective social navigation by selecting navigation policies based on local social context. Learning robotic navigation policies that are consistent with inferred social norms is challenging as these norms are often subjective, culturally dependent, task specific, and context sensitive. While learning-based approaches to social navigation have shown success, they must also be able to compete with established classical algorithms that confer theoretical and practical advantages in simpler scenarios. Therefore, it is beneficial to use the appropriate navigation algorithm depending on the context. In this paper, we devise a hybrid method that combines the strengths of both learned and analytical controllers to achieve the best performance in scenarios with variable social complexity. We show that such a combination strikes a favorable balance in some performance metrics and even surpasses both individual methods in others.**

I. Introduction

This paper develops social navigation policies that allow a robot to move through crowds in a manner that is consistent with normal and polite human behavior. While there exists effective strategies that allow robots to optimally navigate spaces while avoiding obstacles [19], these policies can lead to interactions which most humans would consider impolite and disagreeable. Social navigation, which seeks to address this problem, is traditionally treated as finding a single policy that is appropriate for all situations. In practice, however, navigation policies can vary by cultural region, social context, task, and activity. It is reasonable to expect that people walking out of a subway use different norms for speed, straightness-of-path, and socially appropriate distance from those walking down a warm beach.

In this paper, we investigate a multi-policy framework for social navigation and show that contextual switching between them improves performance. The title of this paper alludes to the movie *Toy Story* [15]. Just as the toys in the film change their behavioral policy when people are around, likewise we suggest robotic agents should adapt their policies to the presence of people and other robots. This paper examines the development of context-dependent behavioral policies. Leveraging the strengths of such contextually-conditioned policies leads to improved rate of successful navigation, while striking a favorable trade-off in several navigation metrics.

[1] *Toy Story* [15] is a popular animated movie.

II. Background

Humans incorporate contextual cues and local information into their decision making process when navigating an environment. This is done with such effectiveness that navigation in social contexts in the presence of other pedestrians seldom requires our conscious consideration. By incorporating such social and local information, social navigation algorithms seek to imbue robotic agents with the same capabilities.

An early model for social navigation was the social forces model developed by Helbing and Molnar [10]. Inspired by physical systems, it models human crowds as force-exerting particles. These forces can then be used to predict pedestrian motion or to inform the motion of an artificial agent, and have been used to study crowd dynamics [11]. This model, while elegant, is often too simple to capture the most complex pedestrian interactions that are not based on relative distances.

In the realm of robot motion planning, the method of potential fields by Khatib [12] constructs a potential field where repulsive forces from obstacles, both dynamic and static, and an attractive force from the destination dictates the motion of a robotic agent. Since this method naturally adapts to dynamic obstacles, it has seen use for the social navigation problem. One such instance is Svenstrup et al. [20] who use rapidly exploring random trees (RRT) [16] to find socially compliant navigation plans through a potential field.

Recently, deep reinforcement learning (DRL) has been successful in solving high dimensional control tasks [7, 8]. An important factor in the success of DRL methods is the reward function, which must be precisely engineered for most complex tasks in order to achieve desired results. Chen et al. [3] use DRL and one such engineered reward function to learn a policy for social navigation. This method, however, relies on the quality of the reward function, and the incorporation of social navigation heuristics within. Such heuristics, however, often defy codification attempts.

Inverse reinforcement learning (IRL) methods, on the other hand, focus on learning such reward functions from demonstration, and can potentially alleviate the reward engineering dilemma. Kretzschmar et al. [14] apply such an IRL approach to social navigation, where trajectories are parameterized by splines and a trajectory distribution is fit to the demonstrations.

Fahad et al. [5] make use of maximum entropy deep IRL (MEDIRL) Wulfmeier et al. [21] using velocity-agumented social affinity map (SAM) features Alahi et al. [1] to capture position and motion information of surrounding pedestrians. Konar et al. [13] use risk-features and a sampling-based MEDIRL procedure to learn navigation policies. Baghi and Dudek [2] use an efficient IRL algorithm, guided cost learning (GCL) Finn et al. [6], in conjunction with a replay buffer to simultaneously learn a reward function and policy. The efficiencies obtained in the latter approach allows for the learning of high quality policies and reward functions in a sample efficient manner.

A. Multi-Policy Control

It is often the case that analytical and learned policies perform best in specific contexts. Analytical methods often require lower computational resources and benefit from theoretic guarantees, while policies trained using learning methods can navigate complex environments but require intensive training. It is natural, then, to carefully apply each method to the appropriate context. For such context-aware policy selection, Cunningham et al. [4] develop a framework for selecting the most effective closed-loop policy for autonomous vehicle control based on simulated horizons. Mehta et al. [18] apply a similar framework for social navigation in crowded scenarios.

III. Methodology

A. Inverse Reinforcement Learning

Inverse reinforcement learning aims to learn the optimal reward function from expert demonstration. By training on the learned reward function, the expert behavior can be replicated. This approach is particularly attractive for social navigation as codifying it in a limited set of rules is difficult.

In this section, we provide an overview of ReplayIRL Baghi and Dudek [2] which our method uses to obtain the socially compliant policy it uses. We model pedestrians using a Markov decision process (MDP). We consider the following MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}\}$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{T} = P(s_{t+1}|s_t, a_t)$ is the transition function, and $\mathcal{R}$ is the reward function.

IRL considers the MDP without rewards, denoted $\mathcal{M}\backslash\mathcal{R}$, and aims to recover the rewards $\mathcal{R}$ from a set of expert demonstrations $D_E = \{\tau_1, \tau_2, \dots\}$. In this work, we consider trajectories to be ordered sequences of states $\tau = \{s_1, s_2, \dots, s_T\}$.

To recover the rewards, Ziebart et al. [22] consider a maximum entropy distribution over trajectories

$$P(\tau) = \frac{\exp\left(\sum_{s\in\tau} r_\theta(s)\right)}{Z_\theta}, \tag{1}$$

where $Z_\theta = \int_\tau \exp\left(\sum_{s\in\tau} r_\theta(s)\right)$ is the partition function which normalizes the above probability distribution and θ are the parameters of the reward function $r_\theta : \mathcal{S} \mapsto \mathbb{R}$. The optimal parameters θ^* can then be found through maximum likelihood estimation (MLE). Wulfmeier et al. [21] use a neural network to parameterize the reward function r_θ, where θ now become the weights of the neural network. This enables learning of rich, non-linear reward functions.

For large state spaces, the partition function Z_θ quickly becomes intractable to compute as the number of trajectories increase and thus must be approximated. Guided cost learning (GCL), proposed by Finn et al. [6], adopts an importance sampling approximation of the partition function

$$Z_\theta = \mathbb{E}_{\tau\sim q}\, \frac{\exp\left(\sum_{s\in\tau} r_\theta(s)\right)}{q(\tau)}. \tag{2}$$

Additionally, GCL trains a policy $\pi(a_t|s_t) : \mathcal{S} \mapsto \mathcal{A}$ at each optimization iteration to use as the sampling distribution $q(\tau)$. This leads to lower variance estimates of the partition function [6]. Baghi and Dudek [2] show that by sharing the replay buffer of an off-policy RL algorithm (e.g. soft actor critic (SAC) [8]) with the IRL optimization procedure of GCL, social navigation policies can be efficiently trained from demonstration. In this work, we obtain our socially compliant policies through the training procedure described in Baghi and Dudek [2].

B. Potential Fields Implementation

The implementation of the potential field controller is based on [12]. The motion of the agent is influenced by an attractive force from the goal and repulsive forces from nearby pedestrians. The attractive force is given by

$$\vec{f}_{attr} = -k_p(\vec{x} - \vec{x}_{goal}) - k_v\dot{\vec{x}} \tag{3}$$

where $\vec{x}$ and $\vec{x}_{goal}$ are the positions of the agent, and the goal respectively. $\dot{\vec{x}}$ is the velocity of the agent and k_p, k_v are the hyperparameters, position gain, and velocity gain. The repulsive force is given by

$$\vec{f}_{rep_i} = \begin{cases} \frac{1}{2}\eta\left(\frac{1}{\rho} - \frac{1}{\rho_0}\right)\frac{1}{\rho^2}\frac{\partial\rho}{\partial x} & \text{if } \rho \le \rho_0 \\ 0 & \text{if } \rho > \rho_0 \end{cases} \tag{4}$$

$$\vec{f}_{rep} = \sum_i^{\mathcal{O}} \vec{f}_{rep_i} \tag{5}$$

where ρ is the shortest distance between the agent and the obstacle i, ρ_0 is the threshold distance within which obstacles start exerting repulsive force on the agent, $\mathcal{O}$ is the set of all obstacles and η is a hyperparameter that scales the repulsive force.

The resultant force acting on the agent, given by

$$\vec{f}_{total} = \vec{f}_{attr} + \vec{f}_{rep} \tag{6}$$

guides the agent in a collision-free path towards the goal.

While the original method [12] assumes a holonomic robot, our agents are non-holonomic and are best approximated by a robot whose action space consists of forward speed and angular velocity. Possible forward velocities are in the range $[0, 1]$ m/s (no reverse motion allowed) and angular velocities fall in the range $[-\pi/6, \pi/6]$. Additionally, we restrict turning at high speed, so the desired speed is inversely proportional to

the change in orientation. The resultant force, f_{total}, obtained from (6), is mapped to the agent's action space by equations

$$s_f = \frac{\vec{f}_{total} \cdot \vec{\dot{x}}}{\|\vec{\dot{x}}\|} \frac{1}{\theta_f} \tag{7}$$

$$\theta_f = \frac{\theta_{fx}}{\pi} \tag{8}$$

where, s_f is the desired speed, θ_f is the change in orientation, θ_{fx} is the signed angle between $\vec{f}_{total}$ and $\vec{\dot{x}}$.

C. Multi-Policy Control

Our goal is to investigate the efficacy of the contextual application both socially-conscious and potential fields policies for navigation. To this end, we employ a straightforward hybrid policy. In this approach, the euclidean distance between the agent and the nearest pedestrian determines whether an IRL or potential field policy is used to determine the action. We use a threshold based-approach, whose value can be determined experimentally for best results.

IV. Experiments and Results

To evaluate our approach, we construct a multi-policy and multi-agent simulation environment. The simulator represents a top-down view of an open space that stages the movements of a crowd. We use the publicly available UCY pedestrian dataset [17]. Specifically, we use the `student003` subset. The starting positions of each pedestrian is dictated by the dataset, and the goal position for each pedestrian is their final dataset position. A subset of pedestrians are chosen to be controlled by an external evaluation policy, while the remainder simply follow the trajectories found in the dataset. We evaluate our approach in scenarios with different ratios of controlled to non-controlled pedestrians. We experiment on 4 scenarios (1-4) where 10%, 30%, 70%, 90% of the original pedestrians are replaced by agents respectively. These represent a broad spectrum of possible situations, from scenarios with a few robots mixed in a crowd of people (house party with serving robots) to ones where a large number of robots operate in the vicinity of a few humans (warehouse robots with a few supervisors). As the policy selection threshold, we use 2 meters, below which the IRL policy for that agent is engaged.

For each scenario, we test on a set of metrics that capture both standard navigational capabilities and social compliance.

1) Successful completion: Fraction of runs that reach the designated goal position without any collisions.
2) Completion time: Number of time-steps required by the policies to reach the goal position.
3) Distance to displacement ratio: Ratio of the length of the trajectory traced by an agent to the distance of the goal from the starting position. This metric measures the directness of the path taken by an agent to reach its goal.
4) Intimate intrusions: Number of time-steps in which an agent violated the intimate space of a pedestrian, defined to be 1.2 meters, which is informed by the study of proxemics [9].

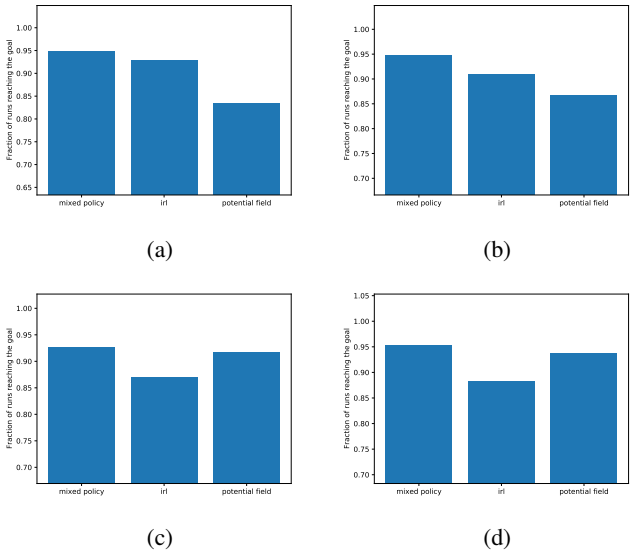

Fig. 1: Fraction of trajectories from different policies successfully reaching the goal. Sub-figures (a)-(d) denote results from scenarios 1-4. Across all the scenarios the mixed policy outperforms the individual policies.

Our results suggest that across all the scenarios, the mixed policy outperforms the individual policies in successfully reaching the goal (Fig. 1). A plausible explanation for this phenomenon is that while the socially trained IRL model is better at avoiding pedestrians (outperforms potential fields at higher densities of pedestrians), the potential field-based agents are better at avoiding each other (outperforms IRL model at higher densities of agents), and by combining them the mixed policy enjoys the best of both worlds.

The potential field agents exhibit more aggressive movements with higher speeds (figure 2) and more direct routes towards the goal (figure 3). This leads to a faster completion time (5), but also registers a higher number of violations of the personal space (figure 4), which is not acceptable in a social setting. In general, ignoring the rules often allows for faster task execution. The mixed policy agent takes the middle ground. In all the scenarios, it maintains social intrusion levels comparable to the IRL agent and performs better at non-social metrics like completion time and directness of route.

V. Conclusion and Future Work

In this work, we investigate the efficacy of context-dependant policy selection in an effort to maximize the performance of both socially compliant and non-compliant policies. As a proof of concept, we combine a socially compliant IRL agent with a potential field controller and compare its performance across different scenarios. We find that the combined policy maintains a good balance of being both a social and an efficient navigator. Possible future work includes further consideration of complimentary policies and data-driven methods for policy selection and application.

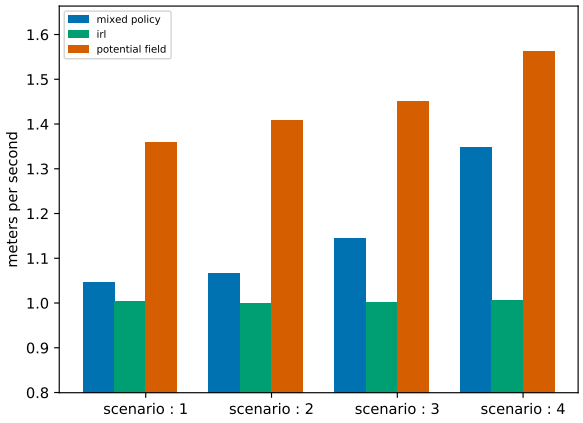

Fig. 2: Average speed of agents controlled by different policies across different scenarios. Agents controlled by the potential field method exhibit higher speeds as compared to the IRL agents, while the mixed policy speeds up in inverse proportion to the number of pedestrians. That is, as the number of agents in the scenario increases, the number of encounters with pedestrians decreases and the average speed of the mixed policy approaches that of the potential field method.

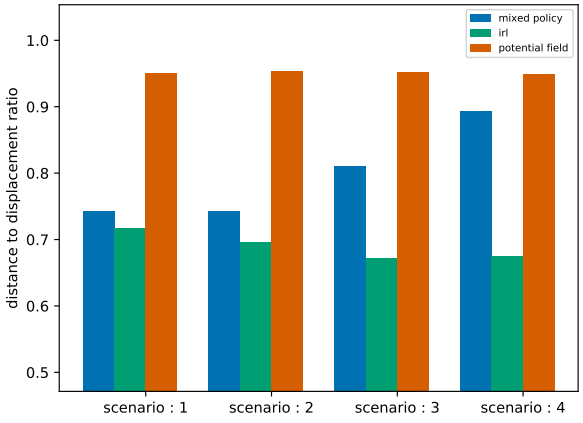

Fig. 3: Average distance to displacement ratio of different methods across different scenarios. While the potential field method prefers direct routes to the goal, the IRL method focuses on social compliance irrespective of the actual presence of pedestrians. In scenarios with higher pedestrian density, the mixed policy behaves more like the IRL method. As the number of pedestrians decreases, the need for social compliance reduces and the policies start opting for optimal navigation rather than social navigation.

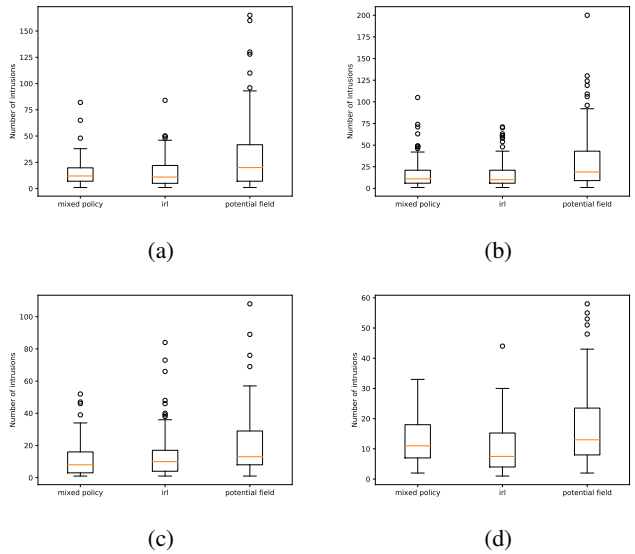

Fig. 4: Number of intimate intrusions incurred by different policies across different scenarios. Sub-figures (a)-(d) denote results from scenarios 1-4. Both the mixed policy and the IRL method commits significantly less intrusions in the personal space of pedestrians as compared to the potential field method.

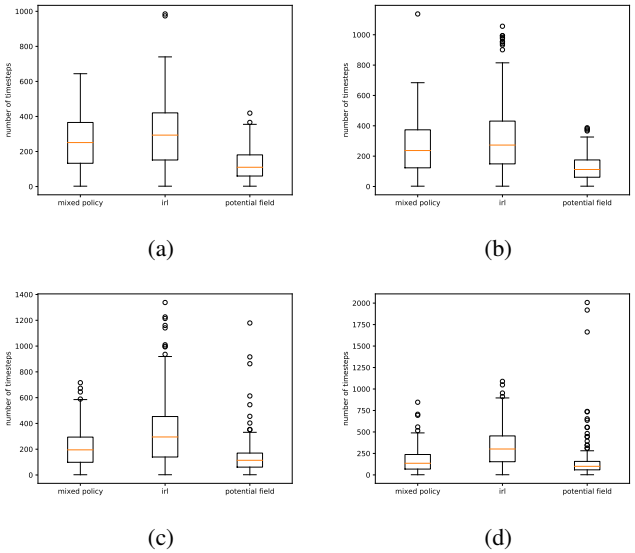

Fig. 5: completion time (in timesteps) by different policies across different scenarios. Sub-figures (a)-(d) denote results from scenarios 1-4. The potential field method has the lowest completion time in all scenarios. The competitiveness of the mixed policy is inversely proportional to the number of pedestrians present in the scenario.

References

[1] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-Aware Large-Scale Crowd Forecasting. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2211–2218, June 2014. doi: 10.1109/CVPR.2014.283.

[2] Bobak H. Baghi and Gregory Dudek. Sample Efficient Social Navigation Using Inverse Reinforcement Learning. *arXiv:2106.10318 [cs]*, June 2021.

[3] Yu Fan Chen, Michael Everett, Miao Liu, and Jonathan P. How. Socially aware motion planning with deep reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1343–1350, September 2017. doi: 10.1109/IROS.2017.8202312.

[4] Alexander G. Cunningham, Enric Galceran, Ryan M. Eustice, and Edwin Olson. MPDM: Multipolicy decision-making in dynamic, uncertain environments for autonomous driving. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1670–1677, May 2015. doi: 10.1109/ICRA.2015.7139412.

[5] Muhammad Fahad, Zhuo Chen, and Yi Guo. Learning How Pedestrians Navigate: A Deep Inverse Reinforcement Learning Approach. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 819–826, October 2018. doi: 10.1109/IROS.2018.8593438.

[6] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization. In *International Conference on Machine Learning*, pages 49–58. PMLR, June 2016.

[7] Scott Fujimoto, Herke Hoof, and David Meger. Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, July 2018.

[8] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic Algorithms and Applications. *arXiv:1812.05905 [cs, stat]*, December 2018.

[9] Edward T. Hall. *Handbook for Proxemic Research*. Society for the Anthropology of Visual Communication, 1974.

[10] Dirk Helbing and Peter Molnar. Social Force Model for Pedestrian Dynamics. *Physical Review E*, 51, May 1998. doi: 10.1103/PhysRevE.51.4282.

[11] Dirk Helbing, Lubos Buzna, Anders Johansson, and Torsten Werner. Self-Organized Pedestrian Crowd Dynamics: Experiments, Simulations, and Design Solutions. *Transportation Science*, 39(1):1–24, February 2005. ISSN 0041-1655, 1526-5447. doi: 10.1287/trsc.1040.0108.

[12] Oussama Khatib. Real-Time Obstacle Avoidance for Manipulators and Mobile Robots. *The International Journal of Robotics Research*, 5(1):90–98, March 1986. ISSN 0278-3649. doi: 10.1177/027836498600500106.

[13] A. Konar, B. H. Baghi, and G. Dudek. Learning Goal Conditioned Socially Compliant Navigation From Demonstration Using Risk-Based Features. *IEEE Robotics and Automation Letters*, 6(2):651–658, April 2021. ISSN 2377-3766. doi: 10.1109/LRA.2020.3048657.

[14] Henrik Kretzschmar, Markus Spies, Christoph Sprunk, and Wolfram Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11):1289–1307, September 2016. ISSN 0278-3649, 1741-3176. doi: 10.1177/0278364915619772.

[15] John Lasseter. Toy Story, 1995.

[16] Steven M. LaValle. Rapidly-exploring random trees: A new tool for path planning. Technical Report 98-11, Computer Science Department, Iowa State University, 1998.

[17] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by Example. *Computer Graphics Forum*, 26 (3):655–664, 2007. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2007.01089.x.

[18] Dhanvin Mehta, Gonzalo Ferrer, and Edwin Olson. Autonomous navigation in dynamic social environments using Multi-Policy Decision Making. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1190–1197, October 2016. doi: 10.1109/IROS.2016.7759200.

[19] B. Oommen, S. Iyengar, N. Rao, and R. Kashyap. Robot navigation in unknown terrains using learned visibility graphs. part i: The disjoint convex obstacle case. *IEEE Journal on Robotics and Automation*, 3(6):672–681, 1987. doi: 10.1109/JRA.1987.1087133.

[20] Mikael Svenstrup, Thomas Bak, and Hans Jørgen Andersen. Trajectory planning for robots in dynamic human environments. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4293–4298, October 2010. doi: 10.1109/IROS.2010.5651531.

[21] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum Entropy Deep Inverse Reinforcement Learning. *CoRR*, abs/1507.04888, March 2016.

[22] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, Chicago, IL, USA, 2008.