Inferring pedestrian decision-making through inverse reinforcement learning

Xiangmin Yang[1][0009-0004-6618-1331], Liu Yang[2][0000-0002-3363-8620], Arnab Majumdar[1][0000-0002-6332-7858], Washington Ochieng[1]

[1] Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, UK
[2] School of Architecture, Southeast University, Nanjing, 210096, China
xiangmin.yang18@imperial.ac.uk
yangliu2020@seu.edu.cn

Abstract. To effectively simulate crowd behavior, understanding the decision-making processes of pedestrians is paramount. This paper proposes a novel method for deducing pedestrians' decision-making by conceptualizing the sum of their reward function along the trajectory as a utility function. While inverse reinforcement learning has been successfully used to retrieve the reward function of pedestrians, the outcome of advanced training algorithms is in the format of neural networks. Due to the black-box nature of neural networks, model interpretability methods are utilized to extract attributions of each input feature. This paper introduces a coupled method of inverse reinforcement learning and model interpretability to infer pedestrians' decision-making on urban sidewalks based on the trajectory data collected in previous experiments. Furthermore, a preliminary test in a classical reinforcement learning environment cart pole is included to demonstrate the viability of the proposed method.

Keywords: Pedestrian decision-making, pedestrian behavior simulation, inverse reinforcement learning, urban street

1 Introduction

Simulating crowd behavior through analytical models or simulation tools has proved its worth in the design of public spaces and planning for large gatherings [1]. Most advanced approaches for modelling multi-agent interactions within a crowd include direct trajectory prediction methods such as deep learning algorithms [2, 3], as well as deriving the movement preference of a crowd through reinforcement learning [4, 5].

Furthermore, a good practical performance of reinforcement learning greatly depends on the design of the reward function [6], which can be thought of as ranking various behaviors [7] as a response to environmental attributes, i.e., a decision-making process [8, 9]. Consequently, a comprehensive understanding of decision-making is vital, and this can also facilitate the provision of a more comfortable walking environment [8] and assist in developing efficient crowd management strategies [10].

While applying reinforcement learning in practice, designing a reward function that is capable of encouraging socially acceptable behavior remains a significant challenge [6]. Inverse reinforcement learning (IRL), in contrast, is capable for deducting such reward functions by learning from the input data. Since the outcome of adversarial IRL method is in the form of neural network, interpretation methods are still needed to quantitatively assess the contribution of each environmental attributes on decision-making. Moreover, the explained reward function allows urban researchers to understand the mechanism between built environment design and pedestrian behavior, which, in turn, can improve the design of public spaces and provide reference for agent modelling within similar scenario.

Therefore, this paper aims to: (1) propose a framework for inferring pedestrians' decision-making in urban sidewalks, and (2) validate the usage of model interpretability in evaluating attributions of each input for the reward function neural network retrieved through IRL.

2 Literature review

2.1 Pedestrian decision-making

The mechanism of decision-making for pedestrians, as identified by [8], involves a process of making trade-offs between desirable and undesirable attributes in the surrounding environments. Such a process is known as utility theory, which assumes that pedestrians assign a quantitative value β_k (also known as utility) to each attribute X_n. The utilities reflect the degree to which the corresponding attributes contribute to the actions made and are combined in a way commonly known as utility function G_n (see Equation 1 as an example). Under utility theory, each action pedestrians make could be viewed as an outcome of optimizing the utility function, as it stands for the mechanism of decision-making.

$$G_n = \beta_1 X_{n_1} + \beta_2 X_{n_2} + \cdots + \beta_k X_{n_k} \quad (1)$$

Two approaches have been adopted widely to infer the relative utility (as we are only interested in their comparative contribution) for different attributes encountered during movement: scores given by participants using a questionnaire [11, 12, 13] and by means of a regression model based on experimental data [14, 15, 16]. However both approaches suffer from drawbacks: pedestrians do not always perform in the way they declare in the questionnaire [17] and the regression model can only be used to decide the contribution of each factor towards action choice at a single timestep, often represented as discrete choices (e.g., choice of exit) [18].

In summary, a better approach for inferring the utility function of value associated with pedestrian decision-making is needed, that can both i) quantify contributions of each attribute towards the function decision reflecting pedestrian's movement preferences in real life, ii) comprehend the movement preference behind the entire trajectory.

As summarized by previous studies [19], common factors affecting pedestrian trajectory include the: movement direction, and speed and density of neighboring pedestrians. Additionally, built environment factors have also been identified as being associated with pedestrians' movement preference [20]. For example, [21] highlighted that the design of street features, including width of sidewalks, transparency of street walls, green spaces (including trees), and street furniture impacts on pedestrians' movements.

2.2 Inverse reinforcement learning (IRL)

As an essential element of reinforcement learning, the reward function R can be used to infer an agent's intentions [6] and is often used to provide a reward signal during each transition along the trajectory. The sum of rewards along the trajectory τ could be viewed as a utility function [22]:

$$G(\tau) = \sum_{t=1}^{\tau} R_t \qquad (2)$$

while optimizing G, the optimal policy π^* is achieved, which stands for the preference of actions over the entire trajectory.

Inverse reinforcement learning [23], on the other hand, has been proposed to infer the reward function being optimized by expert agents. Such a method has been considered necessary as it facilitates the determination of the relative weights of the multi-attribute reward function and constructs an intelligent agent capable of performing specific tasks akin to real-life experts.

In the field of pedestrian dynamics, IRL has been successful in: inferring a navigation policy for robots in human crowds [24, 25, 26], modelling interactions between pedestrians with vehicles [27, 28] and cyclists [29], and predicting future trajectories [30]. This has exploited the capacity of IRL for constructing intelligent agents but left the reward function unexplored.

Adversarial IRL method [31, 6] is a crucial framework proposed by [32] to infer reward function based on generative adversarial network. Such method involves the training of a generator neural network G, together with a discriminator neural network D. The reward function can then be retrieved from the discriminator D in a form of neural network [6]. However, analyzing the reward function still poses challenges, as the neural network decision-making mechanism is not explicitly accessible [33].

In this research, we propose a new method to infer a pedestrian's decision-making process through combining inverse reinforcement learning and model interpretability, which enables the analysis of thein which the outcome is a reward neural network analyzed with the assistance of model interpretability. The result shall present the relative contribution from each attribute within the environment to the decision making of pedestrians, which can not only shed light on the understanding of interactions between environment and pedestrians, but also can be used for a more accurate multiagent modelling.

3 Methodology

3.1 Model framework

Model formation. The entire trajectory of pedestrians could be viewed as a Markov Decision Process (MDP) denoted by the tuple $M = (S, A, T, R, \gamma)$, where S is the state space; A is the action space; $T(s, s', a)$ is the transition probability from state s to state s' under action a determined by the environment dynamics; R is the reward function we are trying to recover; γ is the discount factor ranged on $[0,1)$, indicating the level of preference for immediate reward over future reward.

As summarized above, common impact factors of pedestrian decision-making include pedestrian density, moving direction, and pedestrian speed. From the perspective of each pedestrian, the perceived information only contains the density and speeds of N pedestrians within their eyesight in the direction of moving h, normally represented by a sector with a radius of 4 meters and a central angle of 180 degrees [34]. The density is expressed in terms of the number of pedestrians within eyesight and categorized into two levels of speed: slow and fast, i.e.,

$$N_{slow} + N_{fast} = N_{in\ sight} \quad (3)$$

This forms the foundation of agent state space S together with the current position X and destination D_d. As this research focuses on the urban sidewalk design, the width of the sidewalk W, the shortest distance with the green space D_g, and transparency of the street wall T (denoted by a binary number that 1 stands for transparent) are also combined to reflect the environmental impact. In summary, the state space S will be expressed as:

$$S = [X, D_d, h, N_{slow}, N_{fast}, W, D_g, T\ ] \quad (4)$$

The action space A is denoted as:

$$A = [V, \omega] \quad (5)$$

where V stands for velocity, and ω is the change of direction.

Training Algorithm. Within the framework of Adversarial IRL method**,** the generator G serves as the policy $\pi(a|s)$, which aims to generate occupancy measure $\rho_\pi = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi)$, i.e., the distribution of state action pair under policy $\pi(a|s)$, similar to that of expert policy π_E, which is often provided as a set of trajectories sampled by executing π_E in the environment named expert demonstration.

The discriminator D, on the other side, aims to differentiate the generated state action pair with expert demonstration through minimizing cross-entropy loss: $\log(1 - D(\tau)) - \log D(\tau)$. Adversarial Inverse Reinforcement Learning (AIRL) proposed by [6] is utilized in this research due to its capability to recover reward function that is robust to changes in environment dynamics. Compared with other adversarial IRL methods, such as GAIL [31], AIRL gives a special structure to the discriminator D:

$$D_{\theta,\emptyset}(s,a,s') = \frac{\exp\{f_{\theta,\emptyset}(s,a,s')\}}{\exp\{f_{\theta,\emptyset}(s,a,s')\} + \pi(a|s)} \quad (6)$$

Where $f_{\theta,\emptyset}$ is the learnt reward neural network parametrized by θ and $\emptyset$, restricted to a reward approximator g_θ and a shaping term $h_\emptyset$:

$$f_{\theta,\emptyset}(s,a,s') = g_\theta(s,a) + \gamma h_\emptyset(s') - h_\emptyset(s) \quad (7)$$

Model interpretability. Three model interpretability methods, namely feature ablation, Shapley value sampling (SVP), and Kernel SHAP are explored in this research for their usage in explaining the reward function $f_{\theta,\emptyset}(s,a,s')$. All methods are aimed at evaluating the attribution of each input to the output of the neural network.

In feature ablation, each input is replaced with a baseline, and the attribution is calculated based on the impact of replacement on the output [35]. The concept of Shapley value is utilized in the other two methods, and it is based on concepts from cooperative game theory, in which all the input features collaborate together to accomplish one task (output). The attribution of each input feature is then calculated based on its impact on output when it is added to permutations of other input features [36]. As it is computationally intensive to calculate all the permutations when the dimension of input features is large, alternatives have been proposed that SVP uses random sampling of permutation and Kernel SHAP uses a weighting kernel to perform linear approximation [37].

In this research, a Python package named "Captum" [38] is used to implement these model interpretability methods.

3.2 Data input

The source of data used in the model training process was collected through an experiment led by Liu Yang and her colleagues, which took place from December 15 to 22, 2021, in Nanjing, China. A total of 34 healthy participants aged 14–55 were recruited through social media, and informed consent was obtained from adults and guardians of any teenagers. The experiment adhered to the Declaration of Helsinki guidelines.

Located at Xinjiekou Subway Station in Nanjing's commercial centre, the chosen ground space featured four-story commercial buildings on one side and a sidewalk with trees on the other. A preliminary test on December 16, 2021, with four participants validated the experimental design. The remaining 30 participants, 14 males and 16 females, were grouped into six groups and joined the experiment on the days between December 17 to 22.

The participants were briefed on the experiment's purpose and equipped with portable wireless physiological recording devices in order to collect trajectory data upon arrival at the subway station. Subsequently, participants were guided to the station exit and instructed to go to the next road crossing within 10 minutes. During the walk, participants could freely explore the sidewalk and open spaces but were not allowed to go into buildings. **Table 1** shows a sample of the collected trajectory data and the resultant trajectories are demonstrated in **Fig. 1**.

Table 1. Data sample of participant's trajectories

Start Time(s)	X(°)	Y(°)
0	118.779357	32.0441735
1	118.779357	32.0441735
2	118.779357	32.0441755
3	118.779356	32.0441778
4	118.779361	32.0441805
5	118.779366	32.044178

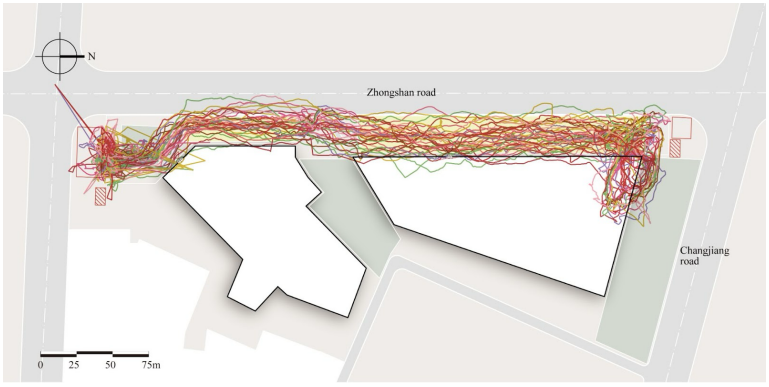

Fig. 1. Trajectories mapped on the site plan

3.3 Data preprocessing

One landmark statue on Zhongshan road is picked as the origin point, with X axis and Y axis extend to the north and east respectively. The geodesic distance is then calculated based between collected trajectory and the proposed reference system so that the positions of the participants are expressed in the unit of meters.

Trajectories are then put into groups based on the allocation on the day of the experiment. At each time step, the information of N_{slow}, N_{fast} is available to the agent within the same group.

Each trajectory is then divided into adjacent intervals of 5 seconds by convention [25]. As this research focuses on the distance with green space D_g and the effect of the street wall, any interval that does not include a position within 2 m range of green space or street wall is eliminated. The destination D_d of this interval is set as the position at the last time step.

4 Preliminary Results

The proposed method, which uses model interpretability methods to understand the reward network trained by AIRL, is further tested based on a classical control task, cart pole, in reinforcement learning [39]. The goal of a cart pole is to keep a pole upright as long as possible by moving the cart left and right, which comprise the action space. The state space includes the position and velocity of the cart, and the angle and angular velocity of the pole with respect to the cart.

A trained model provided by [40] is imported into the test. This model has demonstrated successfully keeping the pole upright for the duration of the episode (500 seconds), and it can be observed from the video that the cart is primarily moving to the right. AIRL is used to retrieve the reward network of this model and three model interpretability methods are then used to infer the relative attribution of each input feature.

The result is presented in **Fig. 2.** . All three methods demonstrate consistency across all the input features besides position in the current state, as feature ablation shows a small positive contribution. Meanwhile, SVP and Kernel SHAP present almost no correlation. Velocity in the current state has been identified as the major negative contributor to the reward, together with angular velocity which has only 1/2 to 1/3 of its contribution. The biggest positive contributor to reward is the left action, and the right action, position and velocity in the next state have almost the same positive attribution.

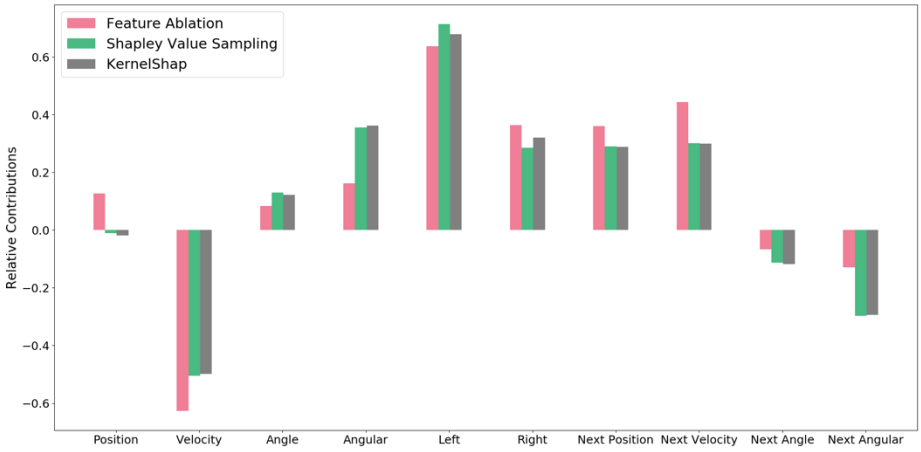

Fig. 2. Relative attribution of reward network input features

As the cart is primarily moving towards the right, it could be argued that the right action is mainly causing the imbalance of the pole (tilt to the left), and thus, the left action is counteracting to regain balance. This is consistent with the result of the test that the attribution of left action is approximately twice that of the right action. Consequently, positive velocity, negative angle and angular velocity, i.e., indicators that the cart is moving to the right, are negatively correlated with the reward. The only

exception lies in the current position, which is mostly irrelevant to the direction of the movement when it changes so rapidly, and thus is consistent with the results of SVP and Kernel SHAP.

However, due to the rightward bias of the cart movement, its position is mostly positive (i.e., on the right side of the origin) during the entire trajectory. Hence a negative value in the next state indicates a series of left movements, which leads to a strong imbalance of the pole. This explains the high positive correlation between the next position and reward. Additionally, the system is preferred to transition to a balanced state, i.e., small angle (low correlation), while moving to the right, i.e., positive velocity and negative angular velocity.

5 Discussions and conclusion

This paper introduces an innovative approach to understand pedestrians' decision-making processes by framing the sum of their reward function along the trajectory as a utility function. Despite the conventional use of IRL for retrieving pedestrians' reward functions, recent progress in training algorithms has shifted towards neural networks. The opaque nature of neural networks necessitates the application of model interpretability techniques to discern the contributions of individual input features. This study advocates a combined method involving AIRL and model interpretability to infer pedestrians' decision-making on urban sidewalks, utilizing trajectory data from prior experiments.

The result of the preliminary test has demonstrated the usage of model interpretability upon explaining the reward network retrieved by AIRL. It can be summarized that velocity and angular velocity should be the key elements when the agent is making action choices.

Similarly, it is reasonable to deduce that an analysis of the pedestrian movement reward net could provide insights into their decision-making process. A figure similar to **Fig. 2** can be achieved with the relative contribution from each attribute to the decision making of pedestrian, e.g., the pedestrian values more on the transparency of the street wall over their distance with the green space.

Consequently, the obtained rewards offer valuable insights for urban designers in order to influence pedestrian movements by refining the characteristics of urban streets, with a particular focus on sidewalks. These rewards can be incorporated into pedestrian models, enabling predictions of walking behavior across various urban design scenarios. This integration facilitates decision-making support for designers seeking to enhance the overall pedestrian experience in urban environments.

It is also useful for the obtained reward to be used in multiagent modelling. This result not only enhances traditional social force modelling [41] by providing a reference for the magnitude of the social force, but also serves as a valuable starting point for reward engineering in various scenarios.

In a subsequent step, the construction of pedestrian decision-making on urban sidewalks will be performed, based on the model proposed in 3.1 with the data demonstrated in 3.2. Different scenarios such as subway station or large event gathering will

also be examined to explore the proposed method's scalability. Furthermore, as the research at the current stage assumes homogeneity among all the pedestrians, categorizing different types of pedestrians based on the occupancy measure and comparing the focus of their decision-making will also be explored.

Acknowledgments. Liu Yang is funded by the China Postdoctoral Science Foundation (No. 2023M740601), the Natural Science Foundation of Jiangsu Province (No. BK20210260), and the National Natural Science Foundation China (No. 52378009).

References

[1] M. Haghani and M. Sarvi, "Pedestrian crowd tactical-level decision making during emergency evacuations," *Journal of Advanced Transportation,* 2016.

[2] Y. Liu, Q. Yan and A. Alahi, "Social nce: Contrastive learning of socially-aware motion representations," 2021.

[3] A. Mohamed, K. Qian, M. Elhoseiny and C. Claudel, "Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction," 2020.

[4] M. Everett, Y. F. Chen and J. P. How, "Collision Avoidance in Pedestrian-Rich Environments With Deep Reinforcement Learning," *IEEE Access,* vol. 9, pp. 10357-10377, 2021.

[5] C. Pérez-D'Arpino, C. Liu, P. Goebel, R. Martín-Martín and S. Savarese, "Robot Navigation in Constrained Pedestrian Environments using Reinforcement Learning," XI'AN, 2021.

[6] J. Fu, K. Luo and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," Vancouver, 2018.

[7] J. Eschmann, "Reward Function Design in Reinforcement Learning," in *Reinforcement Learning Algorithms: Analysis and Applications*, B. Belousov, H. Abdulsamad, P. Klink, S. Parisi and J. Peters, Eds., Springer, 2021, pp. 25-33.

[8] Y. Tong and N. W. F. Bode, "The principles of pedestrian route choice," *Journal of the Royal Society Interface,,* vol. 19, no. 189, 2022.

[9] S. V. Albrecht, F. Christianos and L. Schäfer, Multi-Agent Reinforcement Learning: Foundations and Modern Approaches, Cambridge, Massachusetts: The MIT Press, 2024.

[10] M. Haghani and M. Sarvi, "Imitative (herd) behaviour in direction decision-making hinders efficiency of crowd evacuation processes," *Safety Science,* vol. 114, pp. 49-60, 2019.

[11] Y. Yang and J. Sun, "Study on Pedestrian Red-Time Crossing Behavior: Integrated Field Observation and Questionnaire Data," *Transportation*

Research Record: Journal of the Transportation Research Board, vol. 2393, no. 1, 2013.

[12] G. R. Bivina and M. Parida, "Prioritizing pedestrian needs using a multi-criteria decision approach for a sustainable built environment in the Indian context," *Environment, development and sustainability,* vol. 22, pp. 4929-4950, 2020.

[13] E. Papadimitriou, S. Lassarre and G. Yannis, "Human factors of pedestrian walking and crossing behaviour," *Transportation Research Procedia,* vol. 25, pp. 2002-2015, 2017.

[14] M. Haghani and M. Sarvi, "Stated and revealed exit choices of pedestrian crowd evacuees," *Transportation Research Part B: Methodological,* vol. 95, pp. 238-259, 2017.

[15] R. Lovreglio, E. Ronchi and D. Nilsson, "A model of the decision-making process during pre-evacuation," *Fire Safety,* vol. 78, pp. 168-179, 2015.

[16] F. Soares, E. Silva, F. Pereira, C. Silva, E. Sousa and E. Freitas, "To cross or not to cross: Impact of visual and auditory cues on pedestrians' crossing decision-making," *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 82, pp. 202-220, 2021.

[17] E. Papadimitriou, S. Lassarre and G. Yannis, "Pedestrian risk taking while road crossing: a comparison of," *Transportation Research Procedia,* vol. 14, pp. 4354-4363, 2016.

[18] D. M. P. Wedagamaa, S. Bennettb and D. Dissanayake, "Analyzing Pedestrian Perceptions towards Traffic Safety Using Discrete Choice Models," *International Journal on Advanced Science Engineering Information Technology,* vol. 10, no. 6, 2020.

[19] J. Kim, S. Tak, M. Bierlaire and H. Yeo, "Trajectory Data Analysis on the Spatial and Temporal Influence of Pedestrian Flow on Path Planning Decision," *Sustainability,* vol. 12, no. 24, 2020.

[20] N. Basu, M. M. Haque, M. King, M. Kamruzzaman and O. Oviedo-Trespalacios, "A systematic review of the factors associated with pedestrian route choice," *Transport Reviews,* vol. 42, no. 5, pp. 672-594, 2022.

[21] R. Ewing, A. Hajrasouliha, K. M. Neckerman, M. Purciel-Hill and W. Greene, "Streetscape Features Related to Pedestrian Activity," *Journal of Planning Education and Research,* vol. 36, no. 1, pp. 5-15, 2015.

[22] W. B. K. Allievi, H. Banzhaf and F. S. Stone, "Reward (Mis)design for autonomous driving," *Artificial Intelligence,* vol. 316, 2023.

[23] A. Y. Ng and S. J. Russell, "Algorithms for Inverse Reinforcement Learning," San Francisco, 2000.

[24] M. Fahad, Z. Chen and Y. Guo, "Learning How Pedestrians Navigate: A Deep Inverse Reinforcement Learning Approach," Madrid, 2018.

[25] D. Gonon and A. Billard, "Inverse Reinforcement Learning of Pedestrian–Robot Coordination," *Robotics and Automation Letters,* vol. 8, no. 8, pp. 4815-4822, 2023.

[26] G. Chalvatzaki, X. S. Papageorgiou, P. Maragos and C. S. Tzafestas, "Learn to Adapt to Human Walking: A Model-Based Reinforcement Learning Approach for a Robotic Assistant Rollator," *Robotics and Automation Letters,* vol. 4, no. 4, pp. 3774-3781, 2019.

[27] P. Nasernejad, T. Sayed and R. Alsaleh, "Modeling pedestrian behavior in pedestrian-vehicle near misses: A continuous Gaussian Process Inverse Reinforcement Learning (GP-IRL) approach," *Accident Analysis & Prevention,* vol. 161, 2021.

[28] P. Nasernejad, T. Sayed and R. Alsaleh, "Multiagent modeling of pedestrian-vehicle conflicts using Adversarial Inverse Reinforcement Learning," *Transportmetrica A: Transport Science,* vol. 19, no. 3, 2023.

[29] R. Alsaleh and T. Sayed, "Modeling pedestrian-cyclist interactions in shared space using inverse reinforcement learning," *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 70, pp. 37-57, 2020.

[30] K. Saleh, M. Hossny and S. Nahavandi, "Long-Term Recurrent Predictive Model for Intent Prediction of Pedestrians via Inverse Reinforcement Learning," Canberra, 2018.

[31] J. Ho and S. Ermon, "Generative adversarial imitation learning," Montreal, 2016.

[32] P. C. ,. P. A. S. L. Chelsea Finn, "A Connection Between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models," *abs/1611.03852,* 2016.

[33] Y.-h. Sheu, "Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research," *Frontiers in Psychiatry ,* vol. 11, 2020.

[34] G. Zhangl, Z. Yu, D. Jin and Y. Li, "Physics-infused Machine Learning for Crowd Simulation," New York, 2022.

[35] K. Radha and M. Bansal, "Feature Fusion and Ablation Analysis in Gender Identification of Preschool Children from Spontaneous Speech," *Circuits Syst Signal Process ,* vol. 42, pp. 6228-6252, 2023.

[36] E. Strumbelj and I. Kononenko, "An Efficient Explanation of Individual Classifications using Game Theory," *Journal of Machine Learning Research,* vol. 11, pp. 1-18, 2010.

[37] S. M. Lundberg and S.-I. Lee., "A unified approach to interpreting model predictions.," *Advances in neural information processing systems,* vol. 30, 2017.

[38] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for PyTorch," *arXiv eprint,* vol. 2009.07896, 2020.

[39] A. G. Barto, R. S. Sutton and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *Transactions on Systems, Man, and Cybernetics,* Vols. SMC-13, no. 5, pp. 834-846, 1983.

[40] Hugging Face, "PPO Agent playing seals/CartPole-v0," [Online]. Available: https://huggingface.co/HumanCompatibleAI/ppo-seals-CartPole-v0. [Accessed 1 Feb 2024].

[41] D. Helbing and P. Molnar., "Social force model for pedestrian dynamics," *Physical review E,* vol. 51, no. 5, p. 4282, 1995.