future internet

MDPI

Article

Pedestrian Simulation with Reinforcement Learning: A Curriculum-Based Approach

Giuseppe Vizzari *,† **and Thomas Cecconello** †

Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336/14, 20126 Milano, Italy

* Correspondence: giuseppe.vizzari@unimib.it

† These authors contributed equally to this work.

Abstract: Pedestrian simulation is a consolidated but still lively area of research. State of the art models mostly take an agent-based perspective, in which pedestrian decisions are made according to a manually defined model. Reinforcement learning (RL), on the other hand, is used to train an agent situated in an environment how to act so as to maximize an accumulated numerical reward signal (a feedback provided by the environment to every chosen action). We explored the possibility of applying RL to pedestrian simulation. We carefully defined a reward function combining elements related to goal orientation, basic proxemics, and basic way-finding considerations. The proposed approach employs a particular training *curriculum*, a set of scenarios growing in difficulty supporting an incremental acquisition of general movement competences such as orientation, walking, and pedestrian interaction. The learned pedestrian behavioral model is applicable to situations not presented to the agents in the training phase, and seems therefore reasonably general. This paper describes the basic elements of the approach, the training procedure, and an experimentation within a software framework employing Unity and ML-Agents.

Keywords: pedestrian simulation; multiagent systems; reinforcement learning

Citation: Vizzari, G.; Cecconello, T. Pedestrian Simulation with Reinforcement Learning: A Curriculum-Based Approach. *Future Internet* **2023**, *15*, 12. https://doi.org/10.3390/fi15010012

Academic Editors: Agostino Poggi, Martin Kenyeres, Ivana Budinská and Ladislav Hluchy

Received: 12 November 2022
Revised: 18 December 2022
Accepted: 21 December 2022
Published: 27 December 2022

1. Introduction

Architects, designers, and planners dealing with decisions about the structuring of environments subject to potential crowding employ commercial, off-the-shelf tools for the simulation of pedestrian and crowd dynamics on an everyday basis, especially in collective transportation facilities and in the urban context in general. Decisions related to the spatial arrangement, dimensioning, and even maintenance of specific facilities in which congested situations can arise call for the elaboration of what-if scenarios, indicating what would plausibly happen within a given geometry subject to certain levels of demand. Crowd managers (a relatively new kind of role) are called to plan situations in which existing facilities must be used for hosting large numbers of pedestrians for abnormal functions (e.g., concerts and fairs). With growing frequency, they use these tools to evaluate the crowd management procedures before they are enacted. The results of research on pedestrian and crowd simulation have thus lead to successful technology transfer, but the overall scenario still presents open challenges for researchers in different fields and disciplines to improve model expressiveness (i.e., simplifying the modeling activity or introducing the possibility of representing phenomena that have not yet been considered), the adequacy of the instruments in properly supporting the activity of modelers, and the efficiency of the simulators based on those approaches.

Despite the substantial effort and significant results achieved, the need to properly hand-craft simulation models for a given situation still represents a serious issue: on one hand, the modeler needs to have both serious competence on the topic of pedestrian and crowd dynamics (in addition to other more technical abilities, such as the ability to manage

computer-aided design (CAD) files representing the planned or analyzed environment); on the other hand, their activities often involve arbitrary decisions (e.g., about the extent of the influence of a structural element of the environment over pedestrians). Therefore, the scenarios produced by two different modelers might present differences that, in some cases, might even be potentially relevant. Generally, expert modelers would reach a consensus on most (if not all) modeling decisions; however, what they would call a "modeling mistake" could instead suggest a limit of or a problem with the underlying modeling approach.

Abstracting away from the specific case at hand, this kind of pattern can be found in different areas of application of informatics to computer-supported analysis, control, and expert decision support in even moderately complex sociotechnical systems: while pedestrians and crowd studies can be considered as a specific case of very current traffic and transportation [1], the same paradigm can be suitably applied to the investigation of future scenarios in the areas of the Internet of Things [2], digital twins [3], and smart cities [4]. Once again, we are facing a "knowledge bottleneck": in this particular case, it is about guiding the decisions on how to act in a given situation, but generally within a complex systems perspective, making it very difficult to effectively tackle the problem through optimization techniques.

In the last years, we have also witnessed an evolution in machine learning (ML) approaches, which are being employed with ever growing frequency, even for supporting scientific research in almost every context. The growing availability of data describing pedestrian and crowd behavior (see in particular https://ped.fz-juelich.de/da/doku.php, accessed on 11 November 2022) is motivating researchers to evaluate if this area of application can also see a proper application of these approaches, under which assumptions and conditions, and with what kind of performance, especially compared with existing approaches.

This study represents a contribution in this direction: in particular, we adopted a reinforcement learning (RL) (see the foundational book [5]) approach to the definition of a model for pedestrian locomotion in a built environment. RL represents a type of machine learning technique that is increasingly bing investigated for the implementation of autonomous agents, in particular when the acceptance of the term "autonomous" is strong and closer to the definition provided by [6] ("A system is autonomous to the extent that its behavior is determined by its own experience".) than the most widely adopted definitions in agent computing. RL describes how to train an agent situated in an environment in order to maximize an accumulated numerical reward signal (received by the environment as a feedback to every chosen action). The goal of a simulation should therefore be expressed in terms of a reward function, for which a higher accumulated value should be associated with a higher quality in the simulation dynamics. An RL agent is provided with a model of perception and action, but in addition to these modeling elements and the reward function, the approach can autonomously explore the space of potential agent behaviors and converge to a policy (i.e., a function mapping the state and perception to an appropriate action to be carried out in that context). Although the approach can exploit a certain amount of initial knowledge (analogous to reflexes in animals and humans, or internalized norms, rules, and shared ways to evaluate the acceptability of a given state of affairs), the overall goal is to grant the agent the ability to learn so it can adjust its behavior to improve its performance.

RL approaches, as with most areas of the ML landscape, has been strongly reinvigorated by the energy, efforts, promises, and results brought by the deep learning revolution, and it seems one of the most promising ways to investigate how to provide an agent higher levels of autonomy. On a more pragmatic level, recent developments and results in the RL area suggest that this approach may be an alternative to current agent-based approaches to the modeling of complex systems (see, for instance, the introduction by [7]). Currently, behavioral models for agents require human modelers to manually define agents' behavioral rules, often requiring a complicated interdisciplinary effort, as well as validation processes based on the acquisition and analysis of data describing the studied phenomenon. RL can partly automate this kind of process, focusing on the definition of an environment

representation, the definition of a model for agent perception and action, and defining a reward function and training procedure. The learning process is, in theory, able to explore the potential space of the policies (i.e., agent behavioral specifications) and converge to the desired decision-making model. While defining a model of the environment, as well as agent perception and action, the definition of a reward function and overall training procedure are tasks requiring substantial knowledge about the studied domain and phenomenon, the learning process may significantly simplifythe modeler's work, while solving issues related to model calibration. Although a few examples of applications of this approach can be found in the literature (a relevant selection is discussed in Section 2), the results achieved so far highlight significant limitations, especially in the capability of the generalization of the training phase. Learned behavioral models are sometimes very specific to the types of environments adopted within the training procedure, and this represents a serious problem because (i) it is inconvenient to pay computational costs for performing training in every scenario to be analyzed; (ii) the results of each training process are essentially different behavioral models, trained in different situations, and they are thus not comparable.

This paper presents an experimentation of this approach to pedestrian modeling, trying to start from the last considerations. Whereas RL agents learn how to behave to optimize their expected cumulative reward, pedestrians generally do not exhibit optimal behavior. Therefore, we carefully defined a reward function (combining contributions related to proxemics, goal orientation, and basic way-finding considerations). The most important aspect of this study, however, is the adopted learning process: we defined a particular training curriculum (a concept introduced by [8]), a set of scenarios growing in difficulty supporting the incremental acquisition of proper orientation, walking, and pedestrian interaction competences. Curriculum learning is a general approach not specifically related to RL, but it has been considered as a promising transfer learning approach for RL (as discussed, for instance, by [9]). We considered it particularly well suited for adoption in the RL context: the necessary reflection on the shaping of a reward function seems compatible with the formulation of a *structure* to be given to the overall learning process, leading to a general pedestrian decision model that can be employed in a wide variety of situations (hopefully most of the plausible ones considered in this line of work). The goal of this study was not to systematically investigate all the different alternatives in every single modeling choice (the RL approach is potentially powerful but it is also quite complicated, with a many choices for the different involved concepts and tasks, and several alternative for each of them), but rather to perform a first investigation, trying to achieve results that can be analyzed, especially considering experimental situations in which some data about pedestrian movements are available at least to perform the first steps toward a validation, to evaluate if the overall approach can be really promising, and possibly identifying some criticalities, benefits, and limitations.

After setting the present study within the relevant research landscape, we describe the fundamental elements of the approach, its implementation within a software framework (we are in the process of preparation of a software repository in which the framework will be made available for download) employing Unity (https://unity.com) and ML-Agents (https://github.com/Unity-Technologies/ml-agents, accessed on 11 November 2022), describing the achieved simulation results: in particular, we compare the achieved pedestrian model with both the basic pedestrian agents made available by Unity and with results from the literature in simple benchmark scenarios. We finally discuss the current limits of the approach and our current implementation, as well as ongoing future developments.

2. Related Literature

Pedestrian and crowd dynamics, as suggested in the Introduction, represents an area in which scientific research has produced valuable results, which are now being practically employed by off-the-shelf tools: PTV Viswalk (https://www.ptvgroup.com/en/solutions/products/ptv-viswalk/, accessed on 11 November 2022) officially states that it employs mechanisms based on the *social force model* introduced by [10]. An interesting and compact

discussion of the field from a research oriented standpoint was presented by [11], although it is difficult to provide a compact and yet substantial and comprehensive introduction to the field. This is particularly due to the fact that human decisions related to locomotion can refer to several aspects and areas of knowledge (ranging from environmental and social psychology, to geography, and even anthropology, to mention some of the most apparent) and, despite the fact that technology transfer was successfully carried out, there are still decision-making tasks that are actively being investigated. Way-finding and path-planning activities, for instance, are objects of recent intense research, and researchers have tried to consider factors such as partial or imprecise knowledge of an environment (as discussed by [12]), its dynamic level of congestion, and human factors such as imitation (as proposed by [13]), which can influence overall observed system dynamics.

Machine learning approaches have not yet delivered results able to substitute the traditional hand-crafted models adopted in commercial simulators, and they are still in the stage of active research. One of the first approaches was designed by [14], who investigated both RL techniques (Q learning) and a classification approach to basically choose an action among a small set of available alternatives, based on a description of the current situation and employing a decision tree.

More recently, different authors tried to frame the problem so that *regression* techniques could be employed, either to predict the scalar value of the pedestrian's velocity vector (see, in particular, the study by [15]) or to predict the both the walking speed and direction to be employed (as presented by [16]) considering the current perceived situation. The basic idea is that, owing to the growing availability of raw data describing pedestrian experiments (see the above mentioned website gathering and making available videos and tracking data about pedestrian and crowd experiments (https://ped.fz-juelich.de/da/doku.php, accessed on 11 November 2022)), we could simply devise a deep neural network to be trained according to the contextual situation perceived by a pedestrian and the velocity actually adopted in the next frame of the video. While this approach is relatively straightforward, it is quite limited in terms of the actual possibility to produce a *general model* of pedestrian behavior: even when the whole process should lead to the successful training of the network and to achieving very good results in the specific situations documented in the dataset, there is no guarantee that the network would produce plausible movement predictions in different situations not covered by the experiments.

There is also a trend of research not really working toward the achievement of fully-fledged pedestrian models, but rather focusing on *trajectory forecasting*. Quoting [17], this task can be defined as: "given the past trajectories of all humans in a scene, forecast the future trajectories which conform to the social norms". The temporal window associated with the prediction is understandably leaned toward the *short term*: generally, these studies focused on a *scene*, representing a relatively small area, in which relatively few pedestrians move, and the horizon of the prediction is limited to few seconds. Most recent studies employed deep learning techniques that are to a certain extent related to the above-mentioned approaches with regression.

The RL approach has recently been applied again to the problem of pedestrian behavioral modeling and simulation [18]: the authors clearly and honestly discussed the limits of the achieved model. In particular, although trained agents achieve encouraging quantitative results, from the perspective of the capability of the model to generalize and face potentially complicated environments, social interaction situations, and movement patterns, in some situations, they actually cannot complete the movement they wanted to perform. We emphasize that this is completely understandable when applying an approach that basically explores the space of potential policies for identifying reasonable behavioral specifications in a complicated situation, but this testifies that there is still a need to perform further investigations to fully evaluate the adequacy of the RL approach to the problem of pedestrians and crowd simulation.

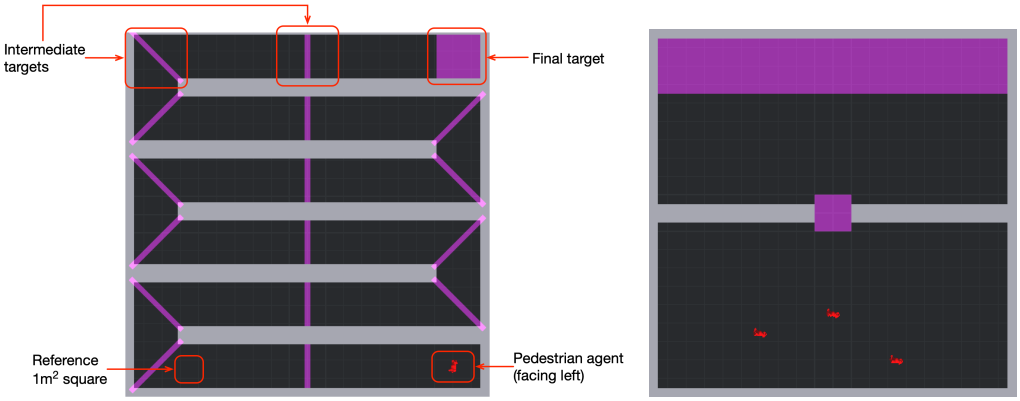
A general consideration of RL compared with other ML approaches is that, on the one hand, RL requires the modelers to provide a set of assumptions, not just about the model

of perception and action of the agent. This is a cost, but it also means that the model can embed (i) concepts about how the environment is actually conceived and interpreted by the agent in relation to its goal oriented behavior, and (ii) an idea of what should be considered desirable behavior (and what should be considered bad choices), and this can represent a way of guiding the learning process in a large space of potential policies. From this perspective, the presented approach is in tune with recent methods on heuristics-guided RL [19] (although we did not technically employ the techniques and framework proposed by the authors), not just for accelerating the training process, but also to achieve a more generally applicable behavioral model.

3. Proposed RL Approach

3.1. Representation of the Environment

For sake of simplicity, in this experimental study, we considered square environments of 20 × 20 m surrounded by walls, as depicted in Figure 1. We anticipated that the framework and especially the learned policies work in larger areas, but the reference overall scenario for this work is represented by indoor movements in everyday buildings and facilities or outdoor environments with structures (e.g., barriers) constraining pedestrian movements. Environments that we employed for both training and testing the model respected this size constraint.



(**a**) (**b**)

Figure 1. Example environments and annotations: (**a**) 'turns' environment and annotations; (**b**) 'unidirectional door' environment and annotations.

The smaller squares (of 1 × 1 m) in Figure 1 are depicted to allow a simpler appraisal of distances. Gray objects are walls, obstacles, and anything that agents perceive as a 'wall'. Violet rectangles are intermediate and final goals. These markers (in the vein of what was proposed by [20]) do not hinder the possibility of moving through them, and they are essentially a modeling tool to support agent's navigation in the environment. One of our goals in the study was to provide an alternative to Unity's path finding and (more generally) pedestrian agent control mechanisms. The agent perception model is introduced below, but we anticipated that they were able to perceive these markers and to select intermediate or final movement targets; we also show that reaching intermediate or final targets influences the agent's reward.

Environments must therefore undergo a preparation phase before being used in the proposed approach: while this kind of activity is difficult to automate and it requires manual effort, all commercial simulation platforms that we are aware of have an analogous requirement. An example of an environment annotated with this rationale is shown in Figure 1a: in this case, the targets in the middle of the horizontal corridors create an affordance: the incentive for agents to move toward that direction although the actual bend at the end of the corridor is still fairly distant. Moreover, oblique intermediate targets in the bends guide agents in the change of direction, also helping them to achieve

a plausible trajectory, avoiding taking trajectories excessively biased toward the internal part of the bend (see, e.g., the considerations proposed by [21] or [22]). Figure 1b shows an environment in which a door (an open one) is present: in this case, the target is used to guide agents in passing through the opening, because the final target is obstructed and not perceivable from positions inside a large portion of the southern room.

3.2. Agent Perception Model

Agents are provided with a limited set of *projectors* of rays, each extending up to a certain distance (10 m in these experiments) and providing specific information about what is "hit" by the ray and the associated distance from the agent.

Projectors (and therefore rays) are not uniformly distributed around the agent; they are more densely present in front of the pedestrian to loosely resemble real human visual perception.

The angle between the rays and the direction an agent is facing (both positive and negative) follows the rule described in Equation (1):

$$\alpha_i = Min(\alpha_{i-1} + \delta * i,\ max\_vision) \tag{1}$$

where δ was set to 1.5, *max_vision* to 90 and α_0 to 0. As a consequence, projectors emit rays at 0°, ±1.5°, ±4.5°, ±9°, ±15°, ±22.5°, ±31.5°, ±42°, ±54°, ±67.5°, ±82.5°, and ±90°. Figure 2 graphically depicts this distribution.

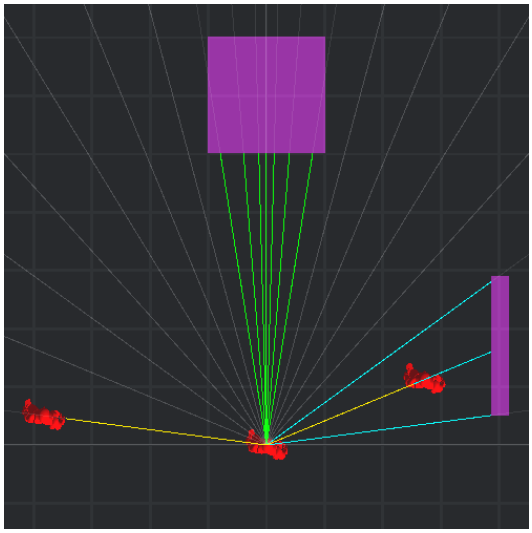


Figure 2. Rays and provided information: yellow = agent, cyan = intermediate target, green = final target, and transparent = wall or none of the others. Pedestrian agents are depicted in red.

The overall number of projectors and rays is therefore 23, but the information associated with and conveyed by rays is different for different objects:

- *Walls and targets* are associated with four observations that are intended to support basic navigation of the environment (i.e., choice of the direction to be followed): (i) a distance (a numeric value normalized employing a ReLU-inspired function to be between 0 and 1, with distances above 10 m capped at 1), and a one hot encoding indicating if the perceived entity is a wall (ii), a target still not visited (iii), or an already visited one (iv); as shown in Figure 2 this type of ray is not blocked by agents.
- *Agents and walls* are associated with additional information, whose intended meaning is instead to support the regulation of distance between the agent and nearby heterogeneous entities that may generate an impact (which should be avoided): rays bringing this kind of information about agents and walls are therefore associated with a distance (i) (analogous to the one in the previous type of ray), a Boolean type (agent or wall) (ii), and optional information about direction of movement (iii) and walking

speed (only relevant for agents) (iv); unlike the previous type of ray, this one is blocked by the first entity that would cause a collision.

Some of the acquired information is therefore potentially doubled: in case a wall is hit by the ray, its presence and distance are highlighted by the two different ray types. In the future, we can attempt to remove references to walls in the second type of ray, but the present results show that this redundancy does not seem to cause training convergence issues.

Other relevant information for the agent observation is its own current walking speed. To improve the performance of the neural networks typically employed in recent RL algorithms all observations, in addition to those for which normalization has already been introduced, current walking speed was normalized in the interval [0,1]. For the normalization of walking speed, we set the maximum velocity for agents to be 1.7 m/s. The overall agent's observation is summarized in Table 1. Essentially, the overall number of observations for each RL pedestrian agent was 185 (1 due to its own previous walking speed, 23 rays each associated with 4 observations for walls and targets, and 23 rays associated with 4 observations for agents and walls).

Table 1. Summary of agent observations.

Type of Observation	Observation	Value
Intrinsic	Own speed	Number
Walls and Targets	Distance	Number
	Type/Tag	One Hot Encoding
Agents and Walls	Distance	Number
	Type/Tag	Boolean
	Direction	Number
	Speed	Number

The overall agent perception model is therefore a simplification of real human perceptive capabilities: the discrete nature of the projected rays makes it possible that, in some situations, objects (obstacles or other agents) might be not immediately perceived, especially when they are not in the center of the field of view. Nonetheless, this definition represents a good balance between plausibility and performance, limiting the computational cost of agent perception as well as the structure of agent observation for the RL algorithm.

3.3. Agent Action Model

The agent action in this model is essentially a change in its speed vector; this translates into two potential changes, for each decision, essentially related to the magnitude of the vector, i.e., *walking velocity* and *direction*.

Each agent is triggered by the overall execution engine to take a decision about if and how to change its own actual speed vector three times per second, in line with [23], in an attempt to consider cognitive plausibility, quality of the achieved results, and computational cost.

The agent's action space was therefore modeled as the choice of two (conceptually) continuous values in the $[-1,1]$ interval that were used to determine a change in velocity vector for magnitude and direction.

The first element, a_0, causes a change in the walking speed defined by Equation (2):

$$speed_t = Max\left(speed_{min}, Min\left(speed_{t-1} + \frac{speed_{max} * a_0}{2}, speed_{max}\right)\right) \quad (2)$$

where $speed_{min}$ was set to 0, and $speed_{max}$ was set to 1.7 m/s. According to this equation, the agent is able to reach a complete stop, or the maximum velocity is two actions (i.e., about 0.66 s). To account for the heterogeneity within the simulated pedestrian population,

each agent is provided with an individual desired walking velocity that is drawn from a normal distribution with average of 1.5 m/s and a standard deviation of 0.2 m/s; so, for each agent, the actual $speed_{max}$ does differ.

The second element of the decision, a_1, determines the change in the agent's direction according to Equation (3):

$$\alpha_t = \alpha_{t-1} + a_1 * 20 \tag{3}$$

The walking direction can therefore change 20° each 0.33 s, which is plausible for normal pedestrian walking, but would be probably not be reasonable for modeling running and/or sport related movements.

For the perception model, the model associated with the agent's action presents limits: for instance, it is not suited to situations in which an agent can choose to jog or run or to perform sudden and significant changes in the walking direction (such as basketball players trying to dribble around opponents). Normal walking scenarios, not emergency evacuation situations, seem compatible with this setting.

3.4. Reward Function

As briefly discussed in the Introduction, RL employs a feedback signal from the environment to the trained agents to guide their learning process as a form of weaker substitute for labels in supervised learning approaches. This feedback signal is defined as a *reward function*, which represents a central element of an RL approach, because agents are trained to maximize the accumulated instantaneous rewards associated with their actions. Pedestrian decision-making activities are fairly complex: conflicting tendencies are evaluated and sometimes reconciled quickly, almost unconsciously, while we walk. Sometimes, individual and collective actions are reasonable or at least explainable in retrospective, in a combination of individual and collective intelligence, that however leads to suboptimal overall performance (see, for instance, the above-cited studies by [13,24]).

Given the above considerations and exploiting the available knowledge on this topic, we hand-crafted a reward function. Initially, we defined basic components, i.e., factors that are generally agreed upon as elements influencing pedestrian behavior. In a second phase, we performed a sort of initial tuning of the related weights, defining the relative importance of the different factors. A fully fledged sensitivity analysis was not performed, whichwill be object of future studies.

The overall reward function is defined in Equation (4)

$$Reward: \begin{cases} +6 & \text{Final target reached} \\ +0.5 & \text{Intermediate target reached} \\ -1 & \text{Reached a previously reached intermediate target} \\ -0.5 & \text{No target in sights} \\ -0.5 & \text{Agent in very close proximity} < 0.6 \text{ m} \\ -0.005 & \text{Agent in close proximity} < 1 \text{ m} \\ -0.001 & \text{Agent in proximity} < 1.4 \text{ m} \\ -0.5 & \text{Wall in proximity} < 0.6 \text{ m} \\ -0.0001 & \text{Each step complete} \\ -6 & \text{Reached the end of steps per episode} \end{cases} \tag{4}$$

The only way to increase the cumulative reward is therefore the reaching of intermediate or final targets. It is not uncommon in RL settings to have a *single* source of a positive reward, i.e., the achievement of the final goal. Let us remind the reader that our goal was to achieve a *generally and directly applicable, yet plausible*, pedestrian behavioral model. In most situations, pedestrians move within an environment structured in several interconnected rooms, whose overall structure is known (e.g., the building hosting their office or workplace, a school or university they attend, or transport stations they use every day). Using just a single positive reward associated with the achievement of the final goal would

require the agent to explore the environment through a training process at the end of which the model would have internalized the environment. By instead allowing the annotation of the environment, there are passages and waypoints that are intermediate targets to be pursued, and this allows agents, through training a macro behavioral specification basically guiding them, to reach intermediate targets until the final one is in sight, and then it should be pursued.

However, reaching targets that have been previously visited has a negative reward, because it implies moving back away from the final goal, and it makes it much less reasonable to try to "exploit" the reward to reach a formally reasonable but totally implausible policy (i.e., reach as many intermediate targets before reaching the final one before the end of the episode). Negative rewards thus are used to suggest that some actions should not be chosen unless they eventually lead to the final goal (and unless better alternatives do the same): a small negative reward granted due to the simple passage of time is usual, which pushes agents to avoid standing still and to actively look for solutions, but we also have negative rewards due to proxemics (as introduced in the foundational work [25]), and to penalize walking too close to walls (again, unless necessary). Finally, the penalization of actions leading to a position from which no target (either intermediate or final) can be seen stimulates agents to pursue the goals.

It must be stressed that the definition of this function is both crucial and hard to complete; moreover, the definition of the reward function is related to the adopted training process, in our case, a curriculum-based approach, so we needed to anticipate some elements in these considerations that we more thoroughly introduce shortly. Let us consider, for instance, the last point, where the penalization to actions bring an agent to a position from which no target can be perceived. We can wonder if having a small bonus for actually seeing a target instead would work analogously: all positive rewards, however, should be defined carefully, because they can lead to pathological behaviors. In this case, in complex scenarios, the training process can converge to a local stationary point in the policy space associated with a behavior for which an agent finds an intermediate target and stands still, achieving a relatively small bonus for each subsequent decision of the episode, rather than trying to reach the final target. This would imply receiving a long sequence of negative rewards. In turn, the small bonus for actually having a target in sight is used in one of the scenarios included in the curriculum (in particular, observe), whose goal is to lead agents to learn that, in certain situations, they should simply stand still, for instance while queuing and waiting for a bottleneck to become reachable, instead of performing overly frequent and essentially useless small turns at very low velocity (or while standing still).

We decided to set the duration of an episode to a very high number of turns. Having adopted a curriculum-based approach, episodes are strongly related to the steps of the curriculum (each episode belongs to a step of the curriculum), and they are generally to be solved (termination conditions are given in Section 4.1) before moving forward to the next stage of the curriculum. As shown in Section 4.5, the overall approach has good convergence properties.

3.5. Adopted RL Algorithm

For this study and experimentation. we adopted the state-of-the-art deep RL algorithm provided by ML-Agents and, in particular, Proximal Policy Optimization (PPO), initially introduced by [26]). PPO is a policy gradient algorithm whose goal is directly learning the policy function π by calculating the gradient of the return achieved as a result of the action choice. Methods of this kind have better convergence properties than dynamic programming methods, but they need a more abundant set of training samples.

Policy gradients function by learning the policy's parameters through a policy score function, $J(\Theta)$, through which it is possible to apply gradient ascent to maximize the score

of the policy with respect to the policy's parameters, Θ. A common way to define the policy score function is through a loss function:

$$L^{PG}(\Theta) = E_t[log\pi_\Theta(a_t|s_t)]A_t \quad (5)$$

which is the expected value of the log probability of taking action a_t at state s_t times the advantage function A_t, representing an estimate of the relative value of the taken action. As such, when the advantage estimate is positive, the gradient is positive as well. By means of gradient ascent, the probability of taking the correct action increases, while the probabilities of the actions associated with negative advantage instead decrease in the other case.

The goal of this study was essentially to evaluate the adequacy of the overall approach to the problem of achieving a proper pedestrian simulation model, without introducing novel RL algorithms. We did not compare the performance of different RL algorithms on the same problem, which will be the object of future studies.

4. Curriculum Learning

4.1. Rationale of the Approach

Curriculum learning, introduced by [8], represents a strategy within machine learning initially devised with the aim of reducing the training times by presenting examples in a specific order of increasing difficulty during training, illustrating gradually more concepts and more complications to the overall decision. Curriculum learning was later employed more specifically as a *transfer* learning technique in RL and multiagent RL, as discussed by [9]. The agent can exploit experiences acquired by carrying out simpler tasks while training to solve more complex ones, in an *intra-agent* transfer learning scheme. In some situation, it was also reported to support a better generalization of the overall training process (see, for instance, [27]). The capability of an agent to generalize the experience and be able to face situations not already experienced during training by leveraging elements of similarity with past experiences is also extremely important for our problem. Pedestrian simulation generally implies analyzing the implications of different, alternative designs on the same crowding condition, without having to perform training for every specific design (which would lead to incomparable results, because they would be achieved by means of different pedestrian models). Within our specific context, in particular, we verified that agents can be trained to "solve" individual scenarios that are present in our curriculum, and, in some cases, the training would even be shorter than the overall curriculum based training process. However, the achieved pedestrian model was not able to produce plausible results in all of the scenarios included in the curriculum, which were assembled as a reasonable representation of a wide class of indoor environments.

A naive application of a curriculum approach, however, initially led to issues somewhat resembling the *vanishing gradient* problem (as discussed by [28]). Technically, here we do not have a recurrent neural network (or an extremely deep one such as those employed for classification of images trained on huge annotated datasets) but, as we show, the overall training process is relatively long and the "oldest experiences" are overridden by the more recent ones.

Within each step of the curriculum execution, a number of parallel executions (for the proposed results 16) of scenarios associated with the specific curriculum steps is carried out. Each execution (representing an episode) can be completed by the agents (every agent reaches the goal) or when a specified duration is reached (some agents have not yet reached the goal); then, it is repeated, unless a successful termination condition for the curriculum step is verified. A step of the curriculum is considered successfully completed whenever two conditions are met: (i) a sufficiently high number of agents have been trained in the scenario (a fixed number, manually established considering the level of crowding in the scenario), and (ii) the average cumulative reward for trained agents in the last episode, excluding the top and bottom 10% (for avoiding being excessively influenced by a small number of outliers), exceeds a given threshold (specifically configured for every step of the curriculum, being dependent on the configuration of the environment). These termination

conditions are important, and they can probably be improved and generalized, but, for the time being, we accepted the limit of a manual- and expert-based definition.

The finally adopted approach, therefore, proceeds training agents in a set of scenarios in growing complexity, one at a time, and provides a final retraining in a selected number of earlier scenarios before the end of the overall training to refresh previously acquired competences.

4.2. Details of the Curriculum

Starting from the above considerations, we defined a specific curriculum for RL pedestrian agents based on this sequence of tasks of increasing complexity that were subgoals of the overall training:

- Steer and walk toward a target;
- Steer to face target;
- Reach the target in narrow corridors;
- Walk through bends avoiding walking too close to walls;
- Avoid collisions with agents walking in the same direction;
- Avoid collisions with agents walking in conflicting directions;
- Combine all behaviors.

We defined this sequence through our expertise in the context of pedestrian simulation, as well as according to a preliminary experimental phase in which we identified issues and difficulties in achieving the desired behavior. For instance, the second step—steering to face a target—was introduced after initial experiments in which we realized that as a consequence of training in more geometrically complex scenarios, agents sometimes had difficulties in finding their targets when the environment was not essentially "guiding them" toward the final target.

It would be interesting to evaluate to what extent this sequence is robust, if it can be improved, or if it is close to the optimum, but such an analysis was not performed at this stage of the research, because we were interested in evaluating the adequacy of the approach and the possibility of achieving promising results in the domain of pedestrian simulation. This kind of analysis (also of ablative nature: is this curriculum minimal or can some steps be safely removed?) on the structure and content of the curriculum will be the object of future studies.

We also included specific test scenarios, that is, environments that were not included in the training curriculum but that were used to evaluate the ability of agents to exhibit plausible behaviors in scenarios that were not experienced in the training phase, rather than just showing that they memorized the environments they had seen.

4.3. Training Environments

Table 2 reports the different environments that were defined for each of the subgoals of the overall training. It also shows which environment was included in the final retraining phase that had to be carried out before using the trained agents for simulation in new environments.

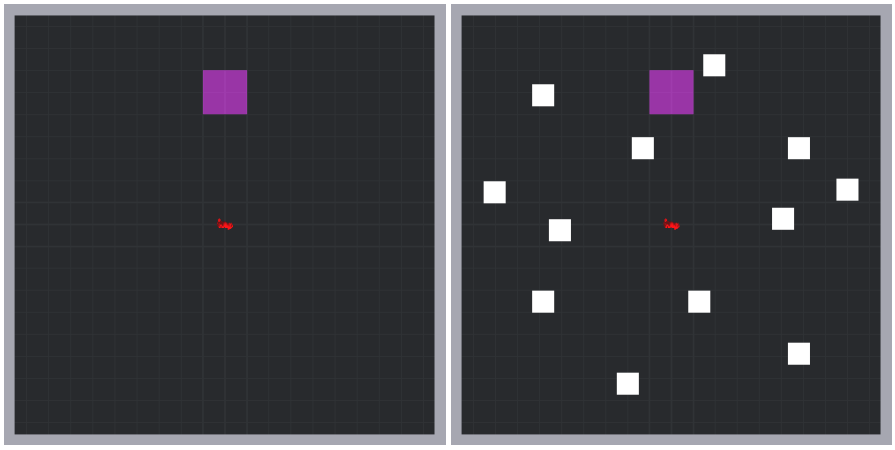
Ti save space, we do not describe every environment and scenario included in the curriculum; instead, we provide a selection of these training environments in Figure 3. Several of these scenarios replicate experiments that were carried out with real pedestrians to study specific behaviors, such as the analyses by [29] or by [30], although we currently did not investigate high-density situations that seem difficult to simulate with a tool such as Unity (which includes 3D models for pedestrians and components for the management of physics that should be overridden for managing significant levels of density, e.g., higher than 1 pedestrian per square meter).

Table 2. Training environment curriculum.

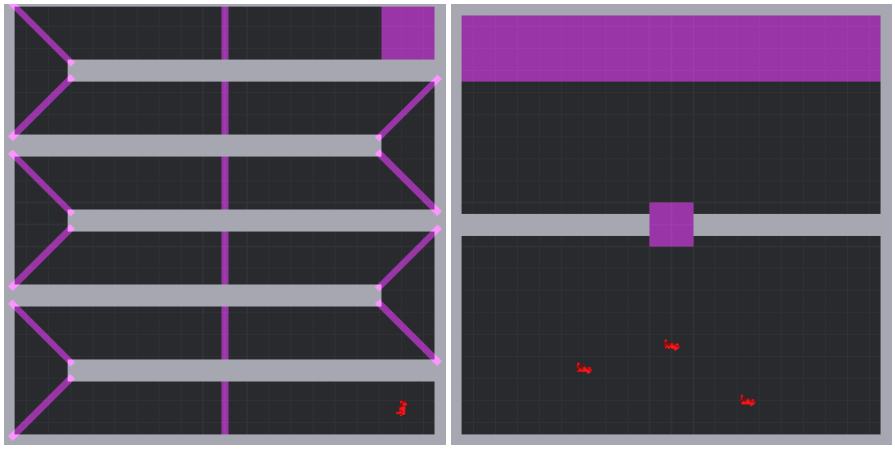
Behavior	Environment	Retraining
Steer and walk toward a target	StartEz	×
	Start	✓
Steer to face target	Observe	✓
Reach the target in narrow corridors	Easy corridor	×
Walk through bends avoiding walking too close to walls	Turns	×
	Turns with obstacles	✓
Avoid collisions with agents walking in the same direction	Unidirectional door	✓
Avoid collisions with agents walking in conflicting directions	Corridor	✓
	Intersection	✓
	T Junction	✓
Combine all behaviors	Crowded Bidirectional Door	✓

We also do not have the space for commenting on the training in all of these scenarios; however, we can highlight some stylized findings that we did observe:

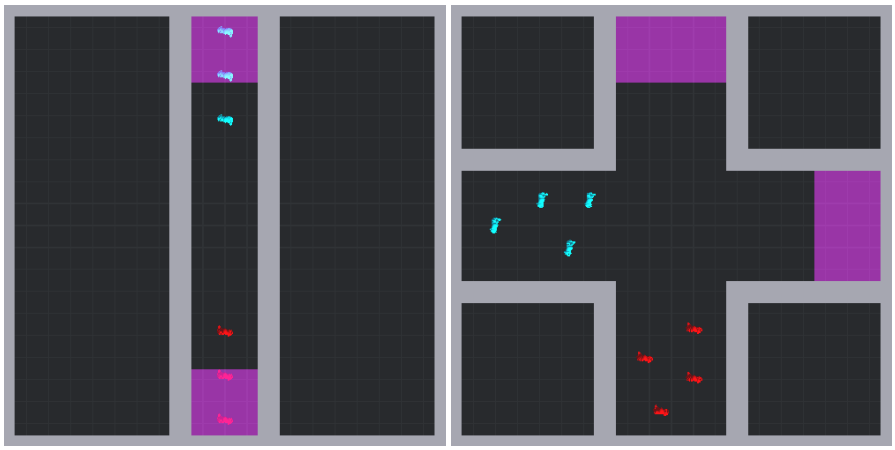
- Within the corridor environment, agents learn to walk in lanes that, due to the low density, are quite stable.
- The turns and turns with obstacles environments produce plausible results in terms of trajectories, but this is mostly due to the placement of intermediate target helping agents in having smooth and plausible paths (as suggested in Section 3.1); once again, similar issues with the management of pedestrian trajectories in bends is present are model-based simulation approaches, as discussed by [21], and in more general situations, as discussed by [31].
- All the environments in which agents had to face narrow passages were crucial in leading them to accept the trade off between choosing some actions leading to an immediate negative reward (i.e., passing close to a wall) and achieving a longer-term positive reward (i.e., reaching the final target).
- All the environments in which agents had to interact with others were analogously decisive, but with a different role: they helped agents understand how to balance the need to slow down, and sometimes even stopping and waiting (when steering is simply not possible or not sufficient) to avoid collisions, with the overall necessary intermediate and final goal orientation. In some situations, at least within our training process, something similar to the "faster is slower" effect [32] was present, because without proper motivations, agents would have inevitably ended up pushing others, and they would not have learned some respectful and collaborative behavior, which is essential to queuing processes.



(**a**) Start environment. (**b**) Observe environment.



(**c**) Turn environment. (**d**) Unidirectional door environment.



(**e**) Corridor environment. (**f**) Intersection environment.

Figure 3. A selection of training environments: white blocks are obstacles, and agents can be red or blue to indicate that they belong to groups having different goals in the environment (e.g., the eastern or northern exits in the intersection environment).

4.4. Training Configuration

Listing 1 reports the defined training configuration file (detailed descriptions of different fields are reported at https://github.com/Unity-Technologies/ml-agents/blob/release_16_docs/docs/Training-Configuration-File.md, accessed on 11 November 2022). The employed ML-Agents version we adopted was 0.25.1 for Python and 1.0.7 for Unity.

Listing 1. Training configuration file.

```
ehaviors:
edestrian:
rainer_type: ppo
yperparameters:
atch_size: 512
uffer_size: 5120
earning_rate: 0.003
eta: 0.01
earning_rate_schedule: constant
etwork_settings:
idden_units: 256
um_layers: 2
eward_signals:
xtrinsic:
amma: 0.99
trength: 1.0
ax_steps: 100000000000000000
ime_horizon: 64
```

Once again, we were interested in evaluating the adequacy of the approach, so we did not perform a systematic analysis of the effect of changing the different hyperparameters. This task will be object of future studies. We just comment here on some of the adopted choices:

- The neural network employed within the PPO algorithm is a fully connected network with 2 hidden layers of 256 nodes each (remember that the agent has 185 observation input signals, associated with inputs to this neural network); a bigger network leads to (sometimes much) longer training times, but it does not improve the quality of the achieved results, whereas a smaller network does not converge to a reasonable policy.
- We employed a basic PPO without curiosity mechanisms, as presented by [33]; therefore, we had essentially just extrinsic reward signals.
- We adopted a very high number for max_steps to let the curriculum guide the actual training, rather than predefined parameters. We also maintained the default value of `time_horizon`.

4.5. Reward Trend During Training

The preliminary tests that we conducted before reaching this configuration for the curriculum, which were based on a single scenario or were based on curricula significantly more compact than the one described above, were unsuccessful, or, at least, they did not produce good results within the same time frame associated with training with this configuration for the curriculum.

The overall training time with the defined curriculum varied according to different factors, but on a Windows-based PC with an Intel Core i7-6820HL @ 2.70 GHz, employing only the CPU (the adopted version of ML-Agents suggests doing so, becuase it would not properly exploit a GPU), would require around 37 min to reach the final retraining phase (which is significantly shorter). Technically, agents were trained in 9 equal environments at the same time, with a Unity velocity set to 100 (i.e., one second of simulation execution corresponded to 100 simulated seconds). The available hardware would not allow further compression of simulated time, but future developments in the ML-Agents framework could produce significant improvements (especially if they would fully exploit GPUs). On the other hand, such a possibility would call for some changes in the training phase workflow.

Figure 4 shows the trend in the cumulative reward. The Tensorboard average reward is the raw measure provided by Tensorboard, while the *average reward* is computed averaging

out the cumulative reward achieved by agents in 36 episodes within an environment. The *Trimmed average reward* actually removes the 10% top- and 10% bottom-performing episodes.

Figure 4. Trend in the reward throughout the training phase.

The different colors highlight the duration of the different scenarios of the curriculum. As expected, the reward dropped (sometimes dramatically) when agents changed the environment, but over time, the training converged. It also clearly shows that environments in which agents have more significant interactions are tougher for training the algorithm. A vanilla PPO was able to successfully converge in such situations, which are much closer to situations that call for specific multiagent RL algorithms. In these situations, the basic approaches often fail due instability in the reward trend that depends on more factors outside the scope of control of the trained agent; specific reward functions that balance individual and aggregated level evaluation of the situation and new algorithms are typically employed. We also conducted an analogous experimentation considering groups of pedestrians, a situation that makes pedestrian-to-pedestrian interaction both more complex and much more frequent (essentially uniformly present in each step of the training) than the type described in the present study, and PPO was not able to converge. The description of this additional experimentation was beyond the scope of the present study.

5. Analysis of Achieved Results

5.1. Qualitative Analysis of Generalization in Test Scenarios

After the training phase, we evaluated the ability of the trained agents to perform smooth and plausible movements in some specific environments that were not presented within the training phase. The goal was basically to understand if the approach was able to grant pedestrian agents, through the above-described training process, a general capability to produce realistic behaviors even in newly encountered situations.

In particular, we evaluated agents' behaviors in the environment depicted in Figure 5 (*"anchor"* environment), in which agents enter from the NE and NW corners make a sharp bend and move north (a movement pattern with a junction between two flows that is not that different from the T-junction environment). Trained agents do not have particular problems, although they might have a hesitation close to the point in which the flows merge due to the sudden necessity to interact and coordinate with other pedestrians coming from the two entrances (something that is also plausible and that can be qualitatively observed in real-world experiments).

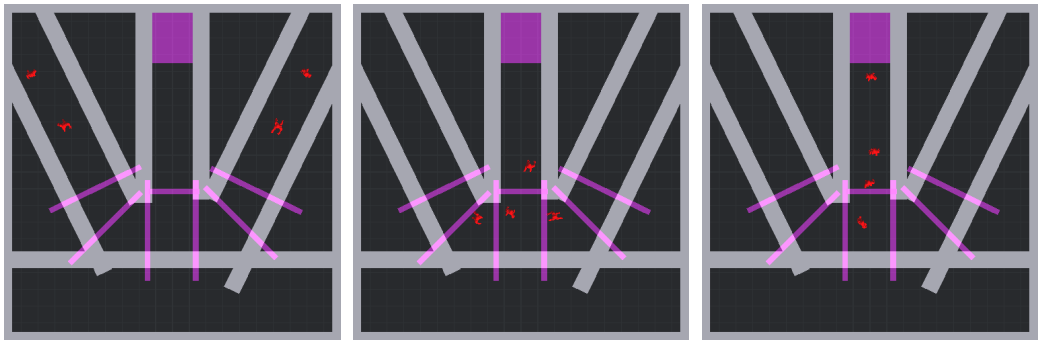


Figure 5. Anchor environment execution.

Figure 6 shows the *"omega"* environment, a maze-like structure in which 90° and U-turns to the right and left are present without choices among different passages (the flow is unique, and there are basically no choices for agents, which need to regulate the distances from other agents and obstacles). We emphasize that the training environments did not include all of these configurations for bends. Trained agents exhibited a reasonable behavior, slowing down before the bends to avoid collisions with walls.

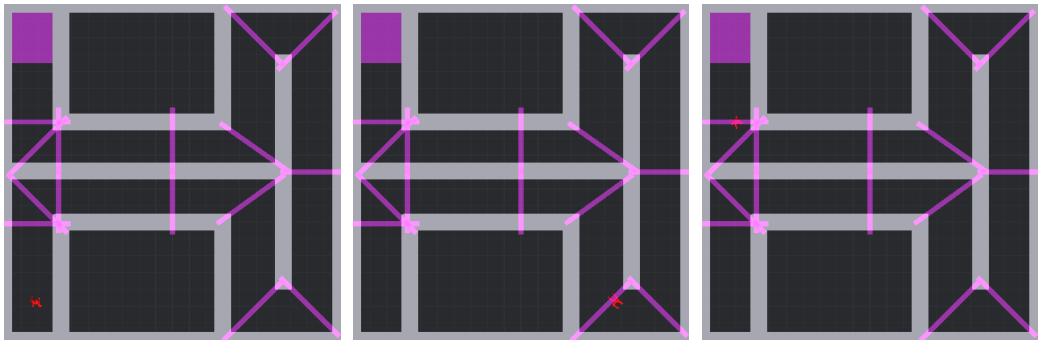


Figure 6. Omega environment execution.

The environment shown in Figure 7 (*"door choice"*) is a relatively simple situation that includes the choice of a passage between two alternatives leading from the southern to the central region, in addition to a single passage to the northern region that includes the final target. Within the training environments, agents never face a situation in which they have to choose among two or more intermediate targets, and we wanted to understand if instead it would be necessary to include this kind of situation in a proper curriculum for training pedestrian dynamics.

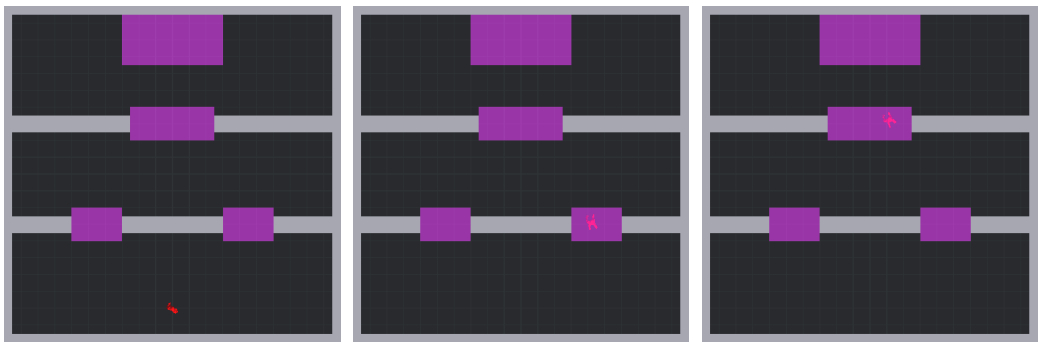


Figure 7. "Door choice" environment execution.

Trained agents did not have a problem in performing a plausible movement pattern in this scenario. They did not always choose the closest passage, but (i) real-world experiments showed that real pedestrians are not necessarily optimizing the expected

travel time (although this generally happens when additional factors to distance, such as congestion, influence their decisions); (ii) additional modifications to the model and to the training curriculum would be necessary to achieve a complete capability to perform way-finding in more complicated situations, especially to be competitive with hand-written and calibrated models.

Figure 8 finally depicts the *"bidirectional door"* environment, a variant of the "crowded bidirectional door" employed in the training. The situation in which the agent is trained includes a number of pedestrians trying to move from the northern to the southern room and vice versa. The lower number of pedestrians present in this situation, coupled with their random initial position, paradoxically can represent a problem for the agents, because they cannot perceive the potential conflict until the very last moments. This scenario was therefore aimed at finding out if the trained agents were able to move at free-flow speed and then slow down when they perceive a conflicting pedestrian, avoiding it while avoiding a complete disruption of the overall trajectory.

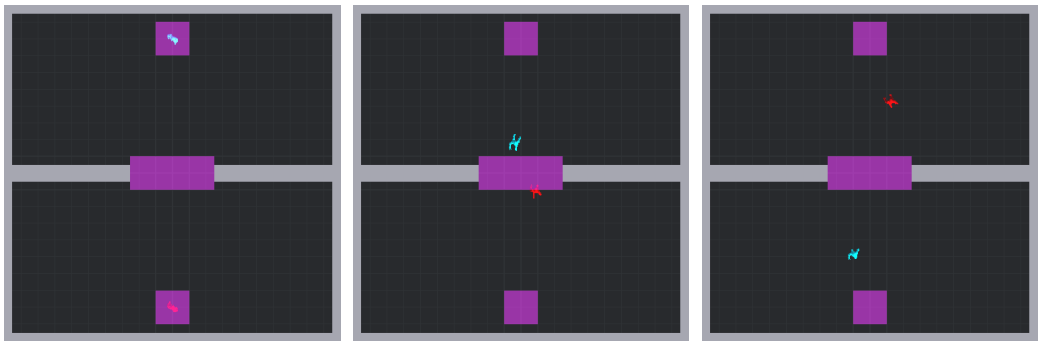


Figure 8. "Bidirectional door" environment execution.

When agents had initial positions granting them immediate mutual perception, they would start moving cautiously, and they crossed the door keeping their right, then moved to the final target. Otherwise, agents started moving at full velocity until they perceived each other, slowing down, and again changing position to avoid each other when passing through the door, generally keeping their right. Sometimes agents did not follow the most direct path to the final target after passing through the door, but the overall behavior was acceptable.

We did not test if the side preference changed or in what proportion, something that could be due to the randomness in the training process or to some systematic bias (maybe due to the spatial structure of the training environments) that leads to an uneven distribution of this preference.

5.2. Quantitative Analyses of Achieved Pedestrian Dynamics

While we mostly talked about the results from an RL perspective (training convergence and trend in the reward during training), and we qualitatively described the overall pedestrian behavior in test environments, we now show some quantitative results in one of the most complicated and challenging environments, the "crowded bidirectional door".

First, Figure 9 shows the distribution of the desired and actual (average) walking speed of pedestrians during a single execution, showing that pedestrian interactions coupled with the environmental structure played a significant role in shaping the overall system dynamics. Agents needed to negotiate who will pass first through the door, and this happened without direct forms of interaction, just through the mutual perception and the learned behavioral policy, which has embedded a sort of emergent norm (e.g., "cautiously approach a passage and pass without bumping other people, stopping and waiting if necessary"). We also appreciate the finding that basically all agents smoothly performed their movement. Some agents were very fast, but the slowest ones (due to the initial

position, they were bound to wait for the others to pass through the door) still moved at about 0.5 m/s.

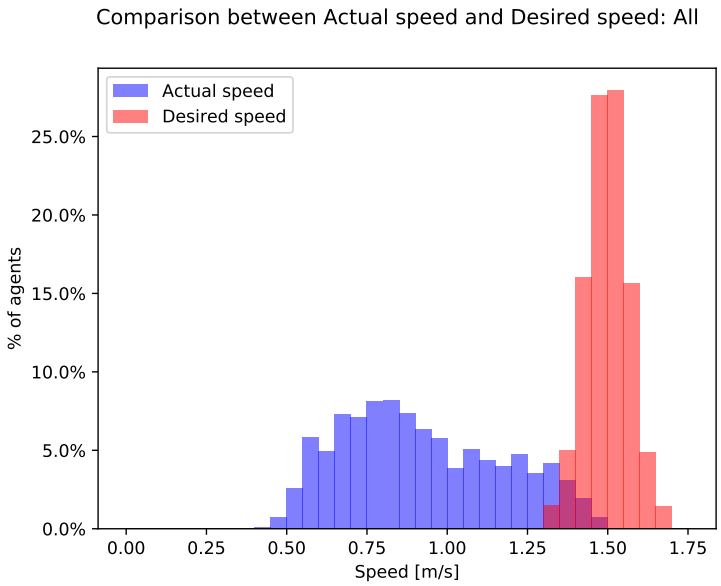
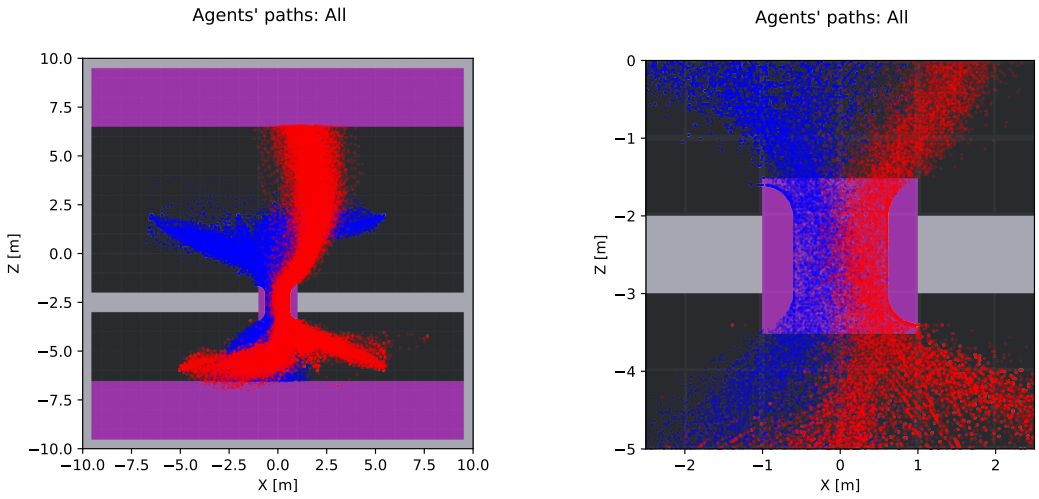


Figure 9. Desired and actual walking speeds in the "crowded bidirectional door" environment.

Figure 10a,b show agents' trajectories in the whole "crowded bidirectional door" environment and in a focused area centered on the door, respectively, to better highlight the movement patterns and some observable systematic implications of the model definition. Red trajectories are associated with pedestrians starting in the southern region, whereas the blue ones are associated with those starting from the northern part of the environment. Some relevant comments about this overall resulting dynamics can be provided:

- For the qualitative analysis of the "bidirectional door" environment, we observed a preference, because agents systematically used the right side of the passage, forming stable lanes, that were instrumental to the overall smoothness of the flow; the door was actually wide enough to accommodate the passage of two pedestrians, and in the presence of a narrower passage, the result would likely be different. This kind of phenomenon is observed in the real world, but it is not necessarily stable (especially at higher densities [34]). Our primary target, compatibly with the limits of the Unity framework and adopted 3D pedestrian models, was not to model high-density situations, and narrow passages essentially implied a locally high density, so we think that, for this model, this level of density is probably the limit.
- Although agents used most of the passage distributing relatively well, they systematically avoided points close to the walls. This was partly due to the physics of the Unity framework, which would not allow a collision between a pedestrian and a wall. However, the sharp and especially straight borders (in [35], real world trajectories have "sharp" borders, but they are "jagged") of the blue and red point clouds on the side of the walls (but not on the border between pedestrian flows) seem to suggest that the proxemic threshold for wall distance indicated in the reward function was perfectly (and maybe too systematically) internalized by the trained agents.



(**a**) (**b**)

Figure 10. Agent trajectories in the "crowded bidirectional door" environment. (**a**) Agent positions in the overall environment. (**b**) Agent positions in the door area.

Figure 11 shows a fundamental diagram, in particular the relationship between walking speed and density in a given area (a 5 m $\times$ 5 m square centered on the door connecting the northern and southern regions.) Velocity dropped, as expected, with the increase in the density (as extensively discussed, for instance, in [35]). A quantitative comparison with real-world data was however not reasonable at this stage of the research, because the measurement mechanism is quite basic and might need improvements (measuring density is still object of discussion and research, as discussed by [36]). However, the velocity levels are plausible, and the drop in velocity with the growth in the density was expected, although it was maybe a bit larger than what is observed in reality. The achieved movement pattern is explainable but probably not the one that was expected. Agents cautiously approached the passage, because the initial distribution and density were such that they almost immediately could perceive another pedestrian potentially causing a collision, plus several other pedestrians competing for using the door. When they finally moved through the door, they actually sped up and reach maximum velocity in the door area (also because the reward function tells them that being close to walls is unpleasant, so they try to minimize the time spent in this condition). However, they have to slow down to coordinate with agents after the door that let them pass. While this movement pattern seems to be not in conflict with available observations [30], additional comparisons are necessary for more serious validation. We also have to consider the fact that learning to balance goal-driven tendencies and collision avoidance within an RL approach to pedestrian simulation is not simple, and past attempts generally ended in failures to complete some movement patterns, as discussed by [18].

A systematic validation of the model is still probably not that important because the model is still the object of study and improvement; we provide some additional results that can be used as a form of comparison with data about real-world pedestrian dynamics. In particular, the T-junction environment was used for observing the pedestrian dynamics that were analyzed by [30]. Figure 12 shows where the pedestrians' positions, focused in the area where the flows originated from the right and left sides of the area, actually merge, and a heatmap shows pedestrians' average velocity in the different points of the environment, with a metric scale supporting interpretation and comparison with experimentally observed data. The results are interesting and promising: the density of our simulations was actually lower than that experienced by pedestrians in real-world experiments, but the movement patterns showed some stylized movements that are in good agreement with real-world data. (i) The lower part of the square in which the left and right corridors meet is not used much by pedestrians, who tend to have a smooth and more round trajectory within

the bend, (ii) pedestrians tend to keep to their side but some mixing between pedestrians coming from the left and right sides of the environment could be observed after the bend, when the pedestrians move toward the upper part of the environment. The level of density within this simulation was still too low to have a reasonable comparison with real-world experiments, and this represents one of the limits of the current simulation framework. However, the observed velocities also showed that interaction among agents, within the merge area and at the beginning of the north corridor, led to conflicts and a reduction in walking speed, in agreement with expectations and empirical evidence.

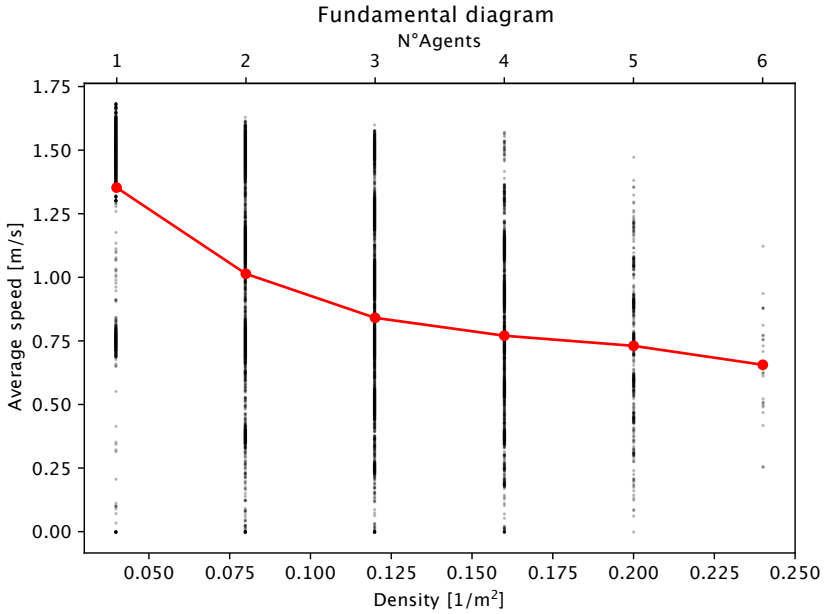
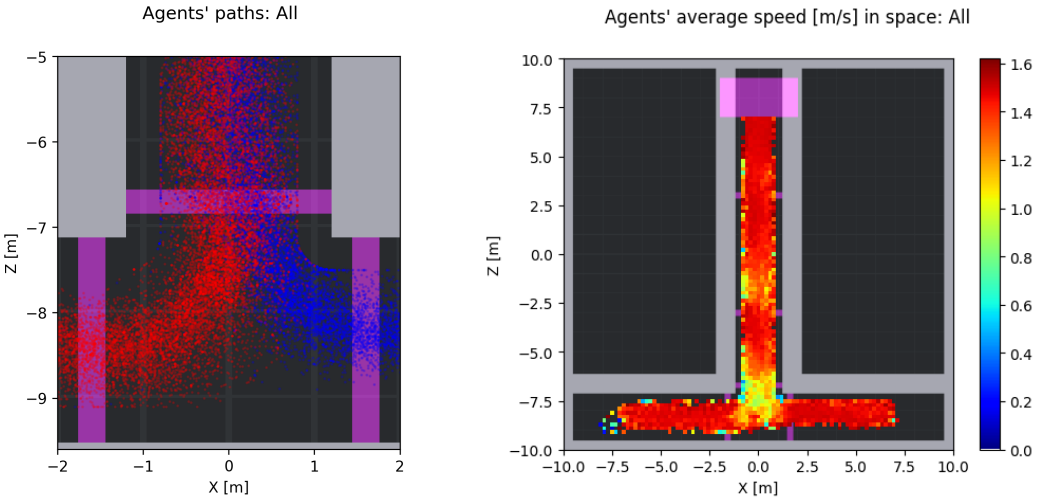


Figure 11. Fundamental diagram (speed vs. density) in a 25 m^2 (5 $\times$ 5 m) area around the door.



(**a**) (**b**)

Figure 12. Quantitative results describing pedestrians' dynamics in the T-junction environment. (**a**) Agents' positions in the merge area. (**b**) Agents' velocities in the whole environment.

In terms of the scalability of the approach, we were able to perform experiments with over two hundred agents employing the decision-making model trained through the described method. The overall *inference* scalability was therefore quite reasonable. On the other hand, the *training* scalability was much more limited (it would currently be impossible to perform the proposed training workflow with hundreds of agents), although,

as suggested above, the adoption of GPU support in the training phase would significantly improve the situation.

5.3. Limitations in Large Environments

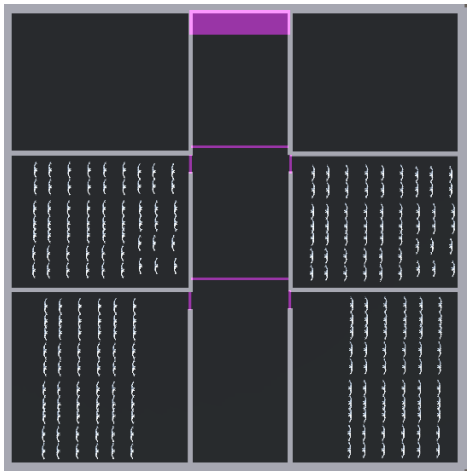
While the achieved results are generally encouraging, and they suggest that the overall goals of the research are plausibly within reach, there are a number of limitations, some of which have already been considered in defining the approach and for which we have some analyses and considerations.

First, although the proposed approach was set on a microscopic scale, the considered environments, despite being reasonably representative of human indoor environments, do not present relatively large halls, so one might wonder if the model achieved through the above-described RL procedure would be able to generate plausible behaviors.

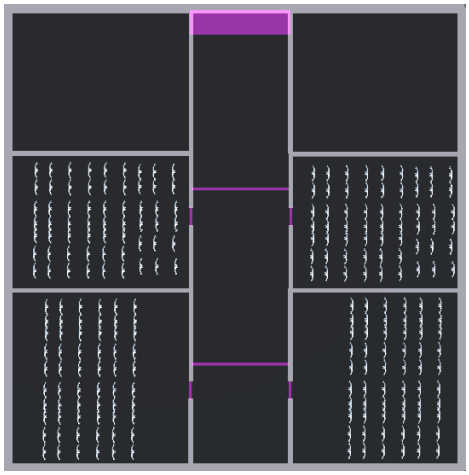
We defined a few additional environments in which we could perform experiments to acquire some insight into the above query. Figure 13 shows two environments lager than the ones in which the training took place (50 m by 50 m squares), structured into lecture-hall-like rooms in which a relatively large number of agents (60 per room) was initially positioned. Agents must exit these rooms, enter a common corridor, and move toward the northern exit. The two environments differed in the position of doors in the lecture halls, which were positioned in the corner of the room for the environment in Figure 13a and at a midwall position for the environment in Figure 13b.

We performed several experiments with agents whose behavior was based on the above-defined training procedure. While most of them were able to successfully vacate the room, some of them (in particular due to the fact that the single exits from the rooms represented bottleneck) took some time to vacate them); in almost every situation, a small number of agents was unable to complete the task within reasonable time. While most of the other agents were able to queue nearby the room exit and then move toward the final target and exit the environment, the failing agents remained in the room, effectively unable to locate the intermediate target, traveling around in circles. Our interpretation of this behavior is that the room size was large enough to make the combination of the defined perception mechanism and curriculum experience insufficient for robust environment navigation.

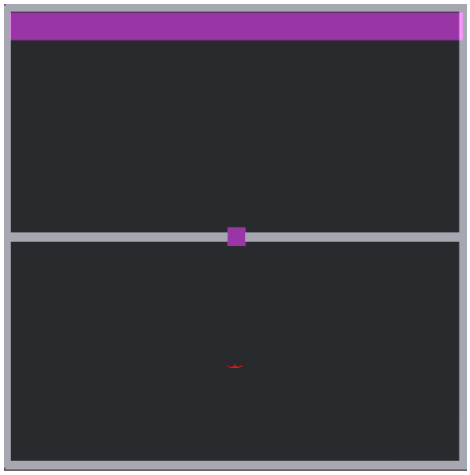
Therefore, and only with the aim of determining the possibility of adding experiences to agents' training by presenting them additional environments in the curriculum, we introduced two additional scenarios at the end of the above defined curriculum. In particular, we created two large environments, depicted in Figure 13c,d, to propose situations in which the agent needed to turn around to find an intermediate target that was both relatively far and not immediately in sight, and to further experience situations in which agents must find a way to coordinate actions, for instance, by queueing and waiting for other agents having a better initial position to vacate an area of shared interest (i.e., the bottleneck). These environments were similar, both in terms of structure and intended effects on training, to other ones that were already present in the curriculum, which could probably be substituted by these new ones, or we could consider randomizing not just the position of doors, but also the size of the environment (at least within a predefined range). For the sake of simplicity, however, we simply initially added these two steps to the curriculum and considered the effects on training and the final capability of agents to successfully perform the tasks. The achieved results were encouraging: situations of agents that lost track in the large environment with lecture halls were avoided, and agents were, in general less, inclined to move quickly toward the lecture hall exit, forming a jam; they were more respectful of personal distances. A more quantitative discussion of the results in this scenario would require a more thorough analysis, but the present results show that the overall curriculum approach represents a plausible step toward achieving if not an acceptable general model for pedestrian movement, at least a useful starting point for a short fine-tuning phase to be carried out in the specific studied situation.



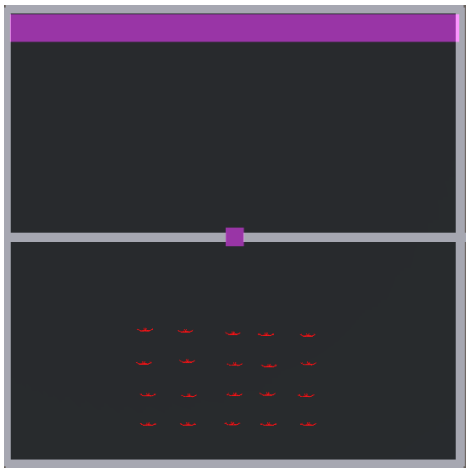
(**a**) Large environment (exit in the corner).



(**b**) Large environment (exit in midwall location).



(**c**) Additional bottleneck environment.



(**d**) Analogous crowded environment.

Figure 13. Additional large environments in which the model was tested (**a**,**b**), and those finally included in the curriculum (**c**,**d**).

6. Conclusions and Future Developments

This paper presented a study exploring the adequacy of applying RL techniques to pedestrian simulation, not just focusing on the possibility to train agents through RL techniques to move within a specific environment, but also considering the need to achieve general models, which are applicable to a wide range of situations. The results we achieved and described in the paper are promising and encouraging, and they show that developing pedestrians and crowd simulation with RL is feasible.

The present study has several limitations that essentially define different lines for future research in different contexts:

- Analysis is needed of the effects of changes in the RL algorithm, hyperparameters, configuration of the curriculum (in this last element, an ablation study in an attempt to identify a minimal curriculum). We reached the presented solution performing some comparisons with alternative settings, but a systematic analysis of each of these aspect would require a focused specific study.
- Validation and expressiveness: Additional quantitative experiments are necessary to improve the evaluation of the achieved results on the side of pedestrian simulation, toward a validation of the model or the acquisition of new objectives for model improvement. Moreover, there might be different situations (especially environmental geometries). For instance, none of the environments we described proposed round

walls; we just employed straight ones. Here, the agents trained through the approach described in this paper might not be able to produce realistic behaviors.
- Overcoming some current behavioral limits: (i) Modeling groups within the simulated pedestrian population is not possible, and preliminary work in this direction suggests that a change in the adopted RL algorithm will be necessary due to the more systematic presence of agent-to-agent interaction, and a multiagent reinforcement learning perspective is necessary [37]; (ii) dealing with high-density situations, which will also imply more fine tuning Unity-specific parameters for the management of the interaction of 3D articulated models such as those associated with pedestrians; (iii) further exploring the capability of the model to perform way-finding, possibly achieving the capability to adapt to the perceived level of congestion, as discussed by [13].

Beyond the specific results in the proposed application context, we think that the overall approach can be of wider interest and applicability, at least within a number of situations in which agent-based and multiagent models and technologies have already been applied. Whenever the analyzed system is intrinsically distributed in nature, when it comprises a number of autonomously deciding but interacting entities, the presented approach can represent at least a starting point to the definition of an effective simulation model. As a general comment to the approach, it must be clear that it represents a useful instrument in the management of what we intuitively called the "knowledge bottleneck" about how to define the actions of agents acting and interacting within the simulated environment, Nonetheless, this approach still requires substantial knowledge about the studied domain. In our example, knowledge and experience in the pedestrian and crowd behavior, and in how to measure, evaluate, and analyze it, were instrumental in the definition of a perception and action model, in the definition of the reward function, and in the configuration of the curriculum-based training process.

Author Contributions: Conceptualization, G.V.; methodology, G.V.; software, T.C.; validation, G.V.; formal analysis, G.V. and T.C.; investigation, G.V. and T.C.; resources, G.V. and T.C.; data curation, G.V. and T.C.; writing—original draft preparation, G.V.; writing—review and editing, G.V. and T.C.; visualization, G.V. and T.C.; supervision, G.V. and T.C.; project administration, G.V. and T.C.; funding acquisition, G.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Authors are in the process of preparation of a software repository in which the framework will be made available for download and for sake of reproducibility.

Acknowledgments: The authors would like to thank Thomas Albericci and Alberto Gibertini for their support in the development of the experimental framework and execution of the simulation campaign. The authors also want to express gratitude to the anonymous reviewers for their support in improving the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bazzan, A.L.C.; Klügl, F. A review on agent-based technology for traffic and transportation. *Knowl. Eng. Rev.* **2014**, *29*, 375–403. [CrossRef]
2. Savaglio, C.; Ganzha, M.; Paprzycki, M.; Bădică, C.; Ivanović, M.; Fortino, G. Agent-based Internet of Things: State-of-the-art and research challenges. *Future Gener. Comput. Syst.* **2020**, *102*, 1038–1053. [CrossRef]
3. Croatti, A.; Gabellini, M.; Montagna, S.; Ricci, A. On the Integration of Agents and Digital Twins in Healthcare. *J. Med. Syst.* **2020**, *44*, 161. [CrossRef]
4. Mualla, Y.; Najjar, A.; Daoud, A.; Galland, S.; Nicolle, C.; Yasar, A.U.H.; Shakshuki, E. Agent-based simulation of unmanned aerial vehicles in civilian applications: A systematic literature review and research directions. *Future Gener. Comput. Syst.* **2019**, *100*, 344–364. [CrossRef]
5. Sutton, R.S.; Barto, A.G. *Reinforcement Learning, an Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
6. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed.; Pearson: London, UK, 2020.
7. Bandini, S.; Manzoni, S.; Vizzari, G. Agent Based Modeling and Simulation: An Informatics Perspective. *J. Artif. Soc. Soc. Simul.* **2009**, *12*, 4.

8. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum Learning. In Proceedings of the Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 41–48. [CrossRef]
9. Silva, F.L.D.; Costa, A.H.R. A survey on transfer learning for multiagent reinforcement learning systems. *J. Artif. Intell. Res.* **2019**, *64*, 645–703. [CrossRef]
10. Helbing, D.; Molnár, P. Social force model for pedestrian dynamics. *Phys. Rev. E* **1995**, *51*, 4282–4286. [CrossRef]
11. Schadschneider, A.; Klingsch, W.; Klüpfel, H.; Kretz, T.; Rogsch, C.; Seyfried, A. Evacuation Dynamics: Empirical Results, Modeling and Applications. In *Encyclopedia of Complexity and Systems Science*; Meyers, R.A., Ed.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 3142–3176.
12. Andresen, E.; Chraibi, M.; Seyfried, A. A representation of partial spatial knowledge: A cognitive map approach for evacuation simulations. *Transp. A Transp. Sci.* **2018**, *14*, 433–467. [CrossRef]
13. Vizzari, G.; Crociani, L.; Bandini, S. An agent-based model for plausible wayfinding in pedestrian simulation. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103241. [CrossRef]
14. Junges, R.; Klügl, F. Programming Agent Behavior by Learning is Simulation Models. *Appl. Artif. Intell.* **2012**, *26*, 349–375. [CrossRef]
15. Tordeux, A.; Chraibi, M.; Seyfried, A.; Schadschneider, A. Prediction of pedestrian dynamics in complex architectures with artificial neural networks. *J. Intell. Transp. Syst.* **2020**, *24*, 556–568. [CrossRef]
16. Zhao, X.; Xia, L.; Zhang, J.; Song, W. Artificial neural network based modeling on unidirectional and bidirectional pedestrian flow at straight corridors. *Phys. A Stat. Mech. Its Appl.* **2020**, *547*, 123825. [CrossRef]
17. Kothari, P.; Kreiss, S.; Alahi, A. Human Trajectory Forecasting in Crowds: A Deep Learning Perspective. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 7386–7400. [CrossRef]
18. Martinez-Gil, F.; Lozano, M.; Fernández, F. Emergent behaviors and scalability for multi-agent reinforcement learning-based pedestrian models. *Simul. Model. Pract. Theory* **2017**, *74*, 117–133. [CrossRef]
19. Cheng, C.; Kolobov, A.; Swaminathan, A. Heuristic-Guided Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, Virtual, 6–14 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W., Eds.; NeurIPS Foundation: San Diego, CA, USA, 2021; pp. 13550–13563.
20. Crociani, L.; Vizzari, G.; Bandini, S. Modeling Environmental Operative Elements in Agent-Based Pedestrian Simulation. *Collect. Dyn.* **2020**, *5*, 508–511. [CrossRef]
21. Crociani, L.; Shimura, K.; Vizzari, G.; Bandini, S. Simulating Pedestrian Dynamics in Corners and Bends: A Floor Field Approach. In *Proceedings of the Cellular Automata*; Mauri, G., El Yacoubi, S., Dennunzio, A., Nishinari, K., Manzoni, L., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 460–469.
22. Dias, C.; Lovreglio, R. Calibrating cellular automaton models for pedestrians walking through corners. *Phys. Lett. A* **2018**, *382*, 1255–1261. [CrossRef]
23. Paris, S.; Donikian, S. Activity-Driven Populace: A Cognitive Approach to Crowd Simulation. *IEEE Comput. Graph. Appl.* **2009**, *29*, 34–43. [CrossRef]
24. Haghani, M.; Sarvi, M. Imitative (herd) behaviour in direction decision-making hinders efficiency of crowd evacuation processes. *Saf. Sci.* **2019**, *114*, 49–60. [CrossRef]
25. Hall, E.T. *The Hidden Dimension*; Doubleday: New York, NY, USA, 1966.
26. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
27. Baker, B.; Kanitscheider, I.; Markov, T.M.; Wu, Y.; Powell, G.; McGrew, B.; Mordatch, I. Emergent Tool Use From Multi-Agent Autocurricula. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
28. Hochreiter, S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **1998**, *6*, 107–116. [CrossRef]
29. Zhang, J.; Seyfried, A. Comparison of intersecting pedestrian flows based on experiments. *Phys. A Stat. Mech. Its Appl.* **2014**, *405*, 316–325. [CrossRef]
30. Zhang, J.; Klingsch, W.; Schadschneider, A.; Seyfried, A. Transitions in pedestrian fundamental diagrams of straight corridors and T-junctions. *J. Stat. Mech. Theory Exp.* **2011**, *2011*, P06004. [CrossRef]
31. Chraibi, M.; Steffen, B. The Automatic Generation of an Efficient Floor Field for CA Simulations in Crowd Management. In *Cellular Automata—Proceedings of the 13th International Conference on Cellular Automata for Research and Industry, ACRI 2018, Como, Italy, 17–21 September 2018*; Lecture Notes in Computer Science; Mauri, G., Yacoubi, S.E., Dennunzio, A., Nishinari, K., Manzoni, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11115, pp. 185–195. [CrossRef]
32. Haghani, M.; Sarvi, M.; Shahhoseini, Z. When 'push' does not come to 'shove': Revisiting 'faster is slower' in collective egress of human crowds. *Transp. Res. Part Policy Pract.* **2019**, *122*, 51–69. [CrossRef]
33. Pathak, D.; Agrawal, P.; Efros, A.A.; Darrell, T. Curiosity-driven Exploration by Self-supervised Prediction. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 2778–2787.

34. Kretz, T.; Wölki, M.; Schreckenberg, M. Characterizing correlations of flow oscillations at bottlenecks. *J. Stat. Mech. Theory Exp.* **2006**, *2006*, P02005. [CrossRef]
35. Zhang, J.; Klingsch, W.; Schadschneider, A.; Seyfried, A. Ordering in bidirectional pedestrian flows and its influence on the fundamental diagram. *J. Stat. Mech. Theory Exp.* **2012**, *2012*, P02002. [CrossRef]
36. Steffen, B.; Seyfried, A. Methods for measuring pedestrian density, flow, speed and direction with minimal scatter. *Phys. A Stat. Mech. Its Appl.* **2010**, *389*, 1902–1910. [CrossRef]
37. Zhang, K.; Yang, Z.; Başar, T., Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In *Handbook of Reinforcement Learning and Control*; Vamvoudakis, K.G., Wan, Y., Lewis, F.L., Cansever, D., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 321–384. [CrossRef]