Curriculum–Based Reinforcement Learning for Pedestrian Simulation: Towards an Explainable Training Process

Giuseppe Vizzari[1], Daniela Briola[1] and Thomas Cecconello[2]

[1]*Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Milan, Italy*
[2]*Department of Electrical Electronic and Computer Engineering, University of Catania, Catania, Italy*

Abstract

Deep Reinforcement Learning (DRL) is a promising approach in the development of autonomous agents adopted in different contexts, from robotic control to virtual avatars in video games. The present contribution presents an application of DRL to the context of pedestrian simulation: building on previous results, we focus on wayfinding decisions, i.e. the decisions among different alternative trajectories within an annotated (planar) environment comprising rooms and passages, in which the agent might need to reach specific intermediate goals before moving towards a final exit. By employing a curriculum based approach, the learning process guides agents to develop a policy leading to the exploration of the environment to reach a set of intermediate waypoints and the final movement target, irrespectively of the specific map of the environment. We discuss the adopted approach, the achieved results, and we discuss potential steps towards improving the explainability of the training process by means of formalization of scenarios included in the curriculum, and their intended training goals.

Keywords

agent-based simulation, pedestrian simulation, reinforcement learning, curriculum learning

1. Introduction

Research in the area of pedestrian modelling has started investigating the application of Machine Learning (ML) approaches leveraging the growing amount of data describing pedestrian and crowd behavior[1]. Even assuming the validity of the experiments and observations that generated those data (a topic that however is raising important questions [1]), the data driven nature of these supervised approaches, however, makes it difficult to achieve models characterized by the level of *generality* (i.e. applicability to a relatively wide range of situations in which maybe no experiment was carried out, and therefore there is no available data) achieved by manually defined approaches.

Recent results [2] have investigated the adoption Reinforcement Learning (RL) [3], a particular type of ML approach in which agents situated in an environment explore the potential space of the policies (i.e., agent behavioral specifications) and converge to a specific behavior assuring

WOA 2023: 24th Workshop From Objects to Agents, November 6–8, Rome, Italy

giuseppe.vizzari@unimib.it (G. Vizzari); daniela.briola@unimib.it (D. Briola); thomas.cecconello@phd.unict.it (T. Cecconello)

0000-0002-7916-6438 (G. Vizzari)

[1] see in particular: https://ped.fz-juelich.de/da/doku.php

the maximization of an expected cumulative reward (a feedback signal evaluating the adequacy of agent's behavior in a given situation). By their own nature, *per se*, these forms of agent training processes do not need a huge amount of data, but they rather require a careful definition of agent's perceptions, actions, and the way they are evaluated within a given situation (i.e. a reward function, more on this point later in the paper). It is not an uncommon RL workflow to perform training on the specific problem / situation at hand, to achieve the best possible results, and to obtain costly models which fit the situation but are not necessarily applicable to others. Within a previous work [4] we proposed a *curriculum* [5] based approach, in which agents were proposed training scenarios of growing complexity, granting both a reasonably fast training and an interesting level of generality and direct applicability of the learned policy without retraining on a new scenario to be investigated. The model focused on *operational level decisions*, although agents were provided with basic information supporting the navigation of simple environments comprising interconnected rooms, but they were not able to choose among alternative paths towards a final goal.

The present contribution builds on the framework discussed in the above cited paper to investigate the possibility to train agents that are able to perceive and exploit environmental information supporting wayfinding [6]. Passages are associated to information indicating if they represent a reasonable way towards a final exit from a scenario, and also if they are to be followed to reach *intermediate goals*. The learning process guides agents to develop a policy leading to the exploration of the environment to reach a set of intermediate waypoints and the final movement target, irrespectively of the specific map of the environment. The contribution will discuss the overall framework, the experimented training process, and the achieved results.

The paper will describe the fundamental elements of the approach, its implementation within a software framework employing Unity[2] and ML-Agents[3], describing the promising achieved simulation results: in particular, we will show that the proposed approach is able to produce plausible results in environments that were not used for sake of training, so the approach seems promising at least in terms of generality. While the curriculum learning structuring implies a goal driven approach in the design of the training process (a step of the curriculum definitely has an intended outcome in teaching the agent how to deal with a certain situation, in pushing it to learn to do something) this passage is mostly intuitive, vague, and definitely not formalized as of this moment. It would be instead important, both for sake of explainability of the training process, as well as for supporting more systematic analyses of the curriculum structure and composition, and potentially to support its maintenance and extension, to take these aspects into account: the paper will provide an initial discussion of how we intend to approach this very interesting future development.

2. Related Works

Pedestrian and crowd dynamics, as suggested in the introduction, represent an area in which scientific research has produced valuable results, that are now being practically employed by

[2] https://unity.com
[3] https://github.com/Unity-Technologies/ml-agents

off-the–shelf tools: PTV Viswalk[4] officially states that it employs mechanisms based on the *social force model* introduced by [7]. An interesting and compact discussion of the field, from a research oriented standpoint, is presented by [8], although it is really difficult to provide a compact and yet substantial and comprehensive introduction to the field. Wayfinding and path planning activities, for instance, are object of recent intense research, and they also to try to consider factors like partial or imprecise knowledge of an environment (as discussed by [9]), its dynamic level of congestion, and human factors like imitation (as proposed by [6]) can influence overall observed system dynamics.

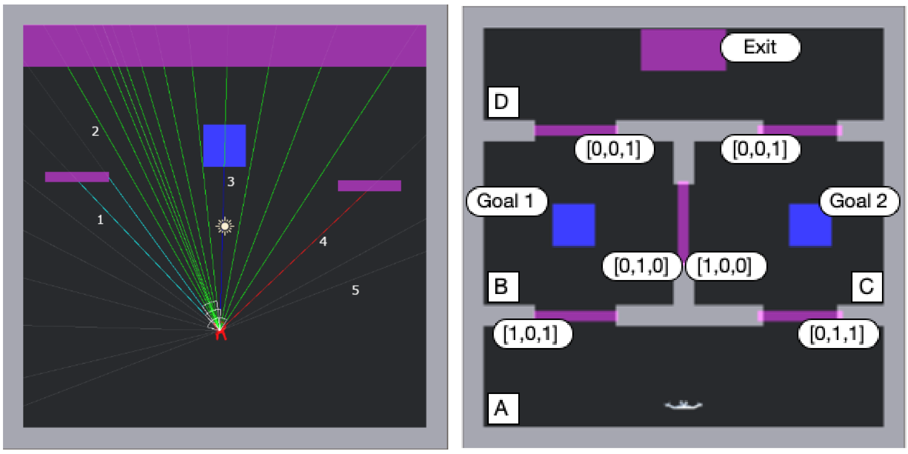
Machine learning approaches have not yet delivered results able to substitute the traditional hand crafted models adopted in commercial simulators, and they are still at the stage of active researches. One of the first approaches, by [10], has investigated both RL techniques (Q-learning) and a classification 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 in such a way that *regression* techniques could be employed, either to predict the scalar value of pedestrian's velocity vector (see, in particular, the work by [11]) or to predict the both the walking speed and the direction to be employed (as presented by [12]) considering the current perceived situation. The basic idea is that, thanks to the growing availability of raw data describing pedestrian experiments (see the above mentioned web site gathering and making available videos and tracking data about pedestrian and crowd experiments [5]), one 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 behaviour: even when the whole process should lead to a successful training of the network and to achieving even very good results in the specific situations documented in the dataset, there is actually no guarantee that the network would produce plausible movement predictions in different situations not covered by the experiments.

The RL approach has been recently applied again to the problem of pedestrian behavioural modeling and simulation producing very interesting results by the already mentioned [2]: authors clearly and very honestly discuss the limits of the achieved model. In particular, although trained agents achieve encouraging quantitative results, also from the perspective of 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 complex situation, but still this testifies that there is still need to perform further investigations to really evaluate the adequacy of the RL approach to the problem of pedestrians and crowd simulation.

A general consideration on RL compared to other ML approaches can however already be done: 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

[4] https://www.ptvgroup.com/en/solutions/products/ptv-viswalk/

[5] https://ped.fz-juelich.de/da/doku.php



(a) (b)

Figure 1: (a) Perception rays of the agent: cyan is associated to rays colliding with valid intermediate target (1), green with final target (2), blue with a goal to be reached (3), red with an invalid mid target (4), gray with a wall (5); (b) Double example environment with two sub goals to be reached before reaching the final exit in the northern room.

can embed (i) concepts about how the environment is actually conceived and interpreted by the agent in relation to its goal oriented behaviour, and (ii) an idea of what should be considered a desirable behaviour (and just as well also what should be considered bad choices), and this can represent a way of guiding the learning process in the large space of potential policies. From this perspective, the presented approach is in tune with recent works on heuristics-guided RL [13] (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. The Proposed Model

3.1. Representation of the Environment

For sake of simplicity in this experimental study environments are bound to be squares of 30 × 30 metres surrounded by walls. Concrete objects that prevent pedestrian passage, such as walls, obstacles are represented in gray, while violet rectangles are intermediate and final goals, whereas blue squares and rectangles are associated to intermediate movement goals not strictly associated to a passage (they can represent areas in which agents need to carry out specific actions). Blue and violet markers (in the vein of [14]), do not hinder the possibility of moving through them, and they are essentially a modeling tool to support agent's navigation in the environment. The models we want to achieve represent an alternative to Unity's path finding and (more generally) pedestrian agent control mechanisms.

The modeler must thus perform an annotation of the environment before using it in the proposed approach; an example of an environment annotated with this rationale is shown in Figure 1. On the left, the various objects are exemplified, while the right shows a vector of booleans associated to every intermediate target, and more precisely to every side of them. The semantics of this vector is that the presence of a 1 in the i^{th} position implies that crossing it from that side will lead an agent towards a specific goal: the small scale scenario depicted in the example comprises two sub goals (respectively associated to the first and second bits) and the final exit (associated to the third bit), so the vector is associated to three bits. The rationale of this representation is to mimic the presence of *signposting indications*, guiding visitors of a building in finding their way to reach. In the experiments in the remainder of the paper we considered vectors including nine positions for intermediate goals plus the final exit.

3.2. Agent Perception

Agents perception is supported by a set of *projectors* generating rays extending up to a certain distance (14 m in these experiments) and supplying indications on what is potentially intersected and the associated distance from the generating agent. Projectors are distributed around the agent according to this rule: $\alpha_i = Min(\alpha_{i-1} + \delta * i,\ max\_vision)$ where δ has been 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 1(a) graphically depicts this distribution. There are thus 23 angles, and for each of them two rays are projected, being associated to different types of information, supporting navigation among different rooms (information about walls and targets), and within rooms (information about walls and goals). Wall information is therefore currently present twice in agent's perception (although in one case it is in the same input in which a target towards another room might be perceived, and in the other it is in the same input in which a goal within a room might be perceived) but, as we will see in the experimental results discussion, this redundancy does not lead to issues preventing training convergence.

The overall agent's observation is summarized in Table 1: in addition to rays, it includes information about agent's state such as the current velocity, a vector indicating if the goal associated to the i^{th} should be reached by the agent, and a boolean that is true when the agent has reached all intermediate goals and should move toward the final exit. To improve the performance of neural networks typically employed DRL algorithms all numerical observations have been normalized in the interval [0,1].

3.3. Action space

Each agent is provided with an individual desired velocity that is drawn from a normal distribution with average of 1.5 m/s and a standard deviation of 0.2 m/s. Each decision, and for these experiments we decided to grant agents three decisions per second (in line with [15], combining cognitive plausibility, quality of the achieved results, and computational costs), determines a potential change in its velocity and this is basically what agent's decision is all about for this model.

Agent's action space has been therefore modeled as the choice of two (conceptually) continuous values in the [-1,1] interval that are used to determine a change in velocity vector,

Type	Observation	Value
Intrinsic	Own velocity	Number
	GoalsVector	Vector of booleans
	AllGoalsAchieved	Boolean
Walls and targets	Distance	Number
	Type/tag	One Hot Encoding
	Direction	Vector of booleans
Walls and goals	Distance	Number
	Type/tag	Boolean

Table 1
Summary of agent's observations.

respectively for magnitude and direction. The first element causes a change in the walking speed defined by Equation 1:

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

Where $speed_{min}$ is set to 0 and $speed_{max}$ is 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).

The second element of the decision determines a change in agent's direction; in particular, $\alpha_t = \alpha_{t-1} + a_1 * 25$. The walking direction can therefore change 25° each 0.33s, that is plausible for normal pedestrian walking, but would be probably not reasonable for modeling running and/or sport related movements.

3.4. Reward Function

The reward function is a central component in a RL approach, representing the only feedback signal guiding the learning process. In this case, we are dealing with a particular form of decision making, with conflicting tendencies that are generally reconciled quickly, almost unconsciously, in a combination of individual and collective intelligence, that generally leads to sub-optimal overall performance [16, 14].

Given the above considerations, we hand-crafted a reward function, initially in terms of macro components, i.e. factors generally influencing pedestrian behavior. Later on we performed a sort of initial tuning of the related weights defining the relative importance of the different factors. Unlike in [4] we focused on individual motivation in exploring the environment, following indications offered by the environment, adjusting the current velocity so as to achieve a plausible overall trajectory. The overall reward function is defined in Equation 2:

$$
Reward = \begin{cases} +6 & \text{Final target reached, all subgoals reached} \\ -3 & \text{Final target reached, at least one subgoal not reached} \\ +0.5 & \text{Valid intermediate target reached} \\ -1.5 & \text{Invalid intermediate target reached} \\ -0.5 & \text{No target in sight} \\ +1 & \text{Subgoal reached} \\ -0.5 & \text{Wall in proximity} < \text{0.6 m} \\ -0.1 & \text{At least one subgoal has not been reached yet} \\ -0.01 & \text{All subgoals reached no target in sight} \\ -0.0001 & \text{Each step done} \\ -6 & \text{Reached the end of steps per episode} \end{cases} \quad (2)
$$

The only ways to increase the cumulative reward are the reaching of a final target (but only if all subgoals were reached), a valid intermediate target (meaning one that leads towards a pursued subgoal or the final target), or a subgoal. Most actions will instead yield a negative reward, associated to an implausible choice or simply to the fact that another decision turn has passed (this small penalty pushes agents to avoid wasting time). Reaching the end of an episode of training (we will provide more information about them later on) without having completed the scenario (i.e. reaching the subgoals and then the final exit) will lead to a substantial negative reward.

3.5. Adopted RL algorithm

Within this work we adopted Proximal Policy Optimization (PPO) [17], a state–of–the–art RL policy–based algorithm, and in particular its implementation provided by ML-Agents [6]. PPO is a policy gradient algorithm learning directly the policy function π, selecting the action to be carried out in a given situation, without the need of a value function (an estimation of the expected return of an action carried out in a given state). These methods generally have better convergence properties compared to dynamic programming methods, but they need a more abundant set of training samples. Policy gradients work by learning the policy's parameters through a policy score function, $J(\Theta)$, through which it is possible to apply gradient ascent, maximizing 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:

$$
L^{PG}(\Theta) = E_t[log\pi_\Theta(a_t|s_t)]A_t \quad (3)
$$

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. When the advantage estimate is positive, the gradient will be positive as well; through gradient ascent the probability of taking the correct action will increase, while decreasing the probabilities of the actions associated to negative advantage. Constructing these estimates requires exploring

[6] https://github.com/Unity-Technologies/ml-agents

the effect of actions in different situations, but this approach is fundamentally different from supervised learning, since no annotated dataset is necessary.

For this specific work, we also employed a curiosity mechanism [18] pushing the agent to explore the state space more efficiently and generalize the acquired experience to unexplored scenarios.

The goal of the work was essentially to evaluate the adequacy of the approach to the problem of achieving a proper pedestrian simulation model and we did not analyze the performance of different RL algorithms yet.

4. The Curriculum Based Learning Process

The notion of Curriculum Learning [5] represents a general training strategy within machine learning, initially conceived to reduce the training times. The underlying rationale is to present examples (in particular labeled examples in supervised learning) in order of increasing difficulty during training, illustrating gradually more concepts and more complications to the decision. In the context of RL it has been employed as a *transfer learning* technique in RL, DRL and Multi–Agent RL [19]: the agent can exploit experiences acquired carrying out simpler scenarios when training to solve more complex ones, in an *intra–agent* transfer learning scheme. Besides improving convergence properties and performance, in some situations it was also reported to support a better generalization of the overall training process [20]. This aspect, coupled with the intuitiveness and cognitive effectiveness of the approach, led us to pursue this approach since achieving a good level of generalization of the acquired experience was also extremely important for our problem. We wanted to train a single model directly applicable to different, alternative designs on the same crowding condition, without having to perform training for every specific design (which would lead to achieve incomparable results, since they would be achieved by means of different pedestrian models).

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

For sake of clarity in the remainder of the section, we define more clearly some key concepts:

- *scenario or step of the curriculum*: the terms will be used interchangeably, and basically they imply a specific simple environment and specific conditions for considering this part of the training process completed;
- *episode*: each scenario generally needs to be experienced several times, each of them called an episode, to accumulate experience (often associated to errors and negative rewards); each episode has a maximum duration, but it can end with the achievement of the goal of the scenario;
- each episode therefore leads to the achievement of a *cumulative reward*, that is the summation of instant rewards achieved as a consequence of each decision and action step;
- the above mentioned *completion condition* for a given scenario is expressed in terms of a mathematical test for the evaluation of episodes cumulative rewards: for example, a

Environment	Episode (s)	Succ. thres.	Retraining
StartEz	50	5.5	No
CorridorObjective	150	0	No
StartObjectiveRandom	200	4.7	No
StartPassage	200	3.5	No
Path	300	5	No
DoubleChoice	200	3.5	Yes
TaskBeforeExit	300	2	Yes
2Rooms	400	-4.5	No
Rooms	400	1	No
LibrarySmall	300	0.5	Yes
Rows	300	-2	Yes
RowsMiddleSplit	300	1	No

Table 2
Thresholds for completion of different training environments. Some of the thresholds are negative since within those environments the occasions for achieving negative rewards throughout a plausible trajectory are numerous.

typical completion condition could be "the average cumulative rewards of the last 10 episodes is higher than threshold th_i".

4.1. 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 are sub–goals of the overall training. Table 2 reports the different environments that were included in the curriculum, including the duration of the associated episode and a threshold for evaluating the acquired cumulative reward. For sake of automation of the curriculum execution, we consider a step of the curriculum to be successfully completed whenever the average cumulative reward for trained agents, excluding the top and bottom 10% (for avoiding being excessively influenced by a small number outliers), exceeds a this empirically defined threshold, specifically configured for every step of the curriculum. The table also shows whose environment are included in the final retraining phase, that must be carried out before using the trained agents for simulation in environments not yet experienced.

The defined curriculum supports the acquisition of three main competences for an agent:

- the ability to steer and walk towards a target / mid-targets;
- the ability to choose among alternative paths, some of which might be preferable according to environmental annotations;
- the ability to avoid moving toward the final target unless all intermediate goals have been reached.

We defined this sequence thanks to expertise in the context of pedestrian simulation, as well as to a preliminary experimental phase and to the experience acquired in a previous research

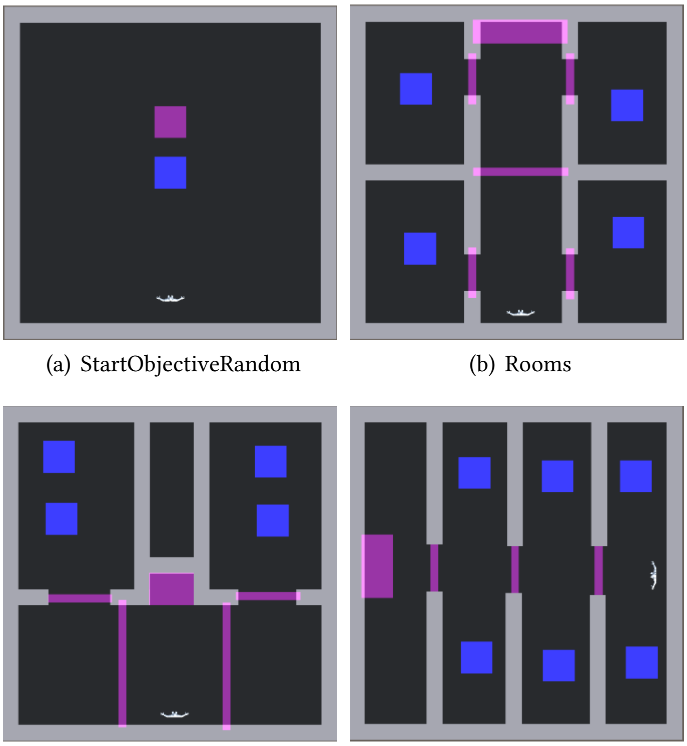


Figure 2: A selection of training Environments.

effort adopting a similar approach [4]. An ablation study, as well as analyses of the robustness of this training process are object of future works.

For sake of space, we cannot describe every environment and scenario included in the curriculum, but a selection of these training environments is shown in Figure 2.

Starting from basic situations in which the agent learns how to steer towards a randomly positioned goal, reach it, and only afterwards move towards the final exit (also randomly positioned) (Figure 2(a)), we have situations in which the agent can choose alternative paths in which the conflicting tendencies to save time and reach all subgoals make choices more complicated (Figures 2(b), 2(c) and 2(d)).

4.2. Reward Trend During Training

We executed the above described curriculum training process adopting the following choices for the configuration of the PPO algorithm implemented in ML-Agents:

- the employed neural network employed is a fully connected network with 2 hidden layers of 256 nodes each; a larger network leads to longer training times but it does not improve the quality of the achieved results, whereas a smaller network does not converge to a reasonable policies;

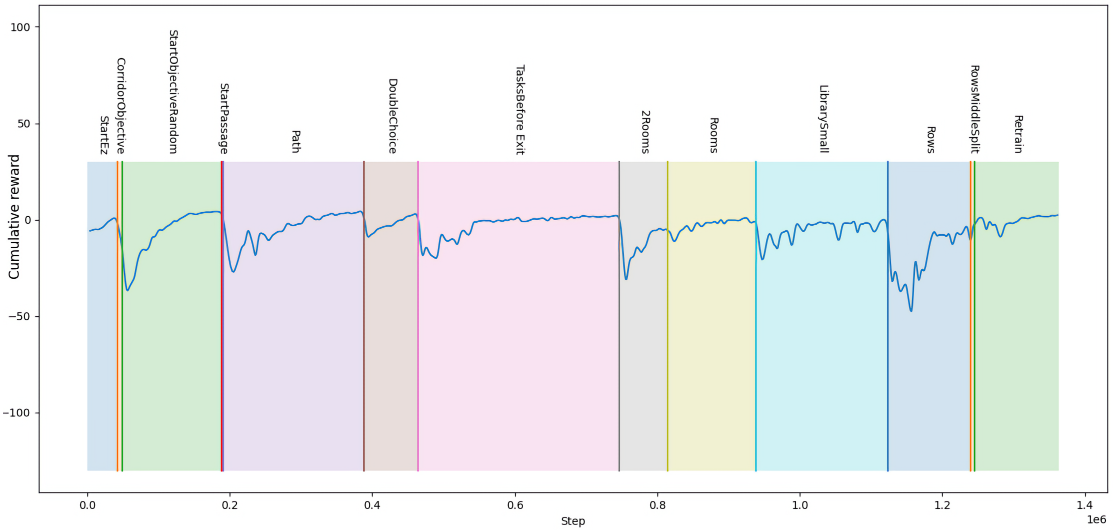


Figure 3: Cumulative reward throughout training: please notice that the line is achieved as a smoothing of actual data points, so some small artifact is present: in particular, within the CorridorObjective scenario the cumulative reward seems to be dropping, but this is due to the fact that the training phase is short and the beginning of the following scenario is instead leading to an extreme drop in the cumulative reward.

- we employed a basic PPO with curiosity mechanisms [18], combining extrinsic reward signal (strength 1) and a very moderate intrinsic signal (strength 0.004);
- we adopted a very high number for max_steps (maximum episode duration) to let the curriculum guide the actual training, rather than predefined parameters.

The overall training time with the defined curriculum varies according to different factors, but on a Windows based PC employing an AMD Ryzen 5 2600 (3.4 GHz) provided with 16 GB RAM, employing only the CPU[7] would require around 1 hour and half. Figure 3 shows the trend of the cumulative reward; in particular, the shown value is computed averaging out the cumulative reward achieved by agents in 12 episodes within the associated environment.

The different colors highlight the duration of the different scenarios of the curriculum: as expected the reward drops in a very significant way when agents moves on from a step of the curriculum to the next one, but through time the training converges. It is also apparent that different environments in the curriculum have a different difficulty, at least within this specific training process. The TaskBeforeExit environment seems particularly challenging, at least in terms of slow convergence to the necessary threshold for advancing to the next stage of the curriculum: this is due to the fact that the agent repeatedly passes nearby the final exit and therefore is tempted to move toward it despite it still has subgoals to reach. Once again, an ablation study of the curriculum would be important and it is object of future works.

[7] The adopted version of ML-Agents suggests doing so, since it would not properly exploit a GPU.

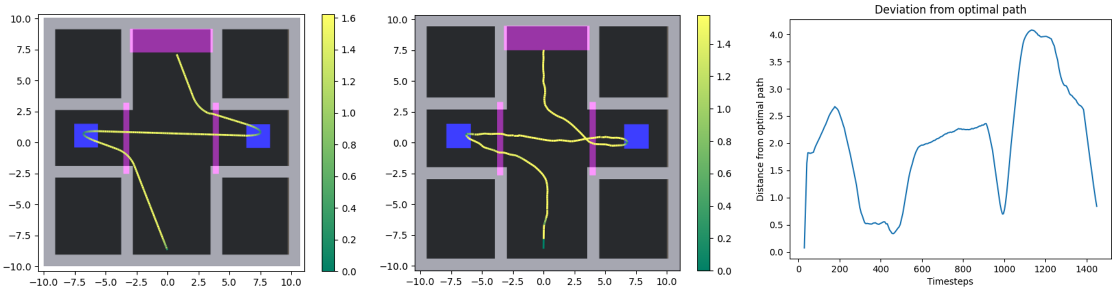


Figure 4: Cross environment execution: baseline optimal path on the left, trained agent in the middle, spatial deviation on the right.

5. Analysis of Achieved Results

In line with the goal of evaluating the possibility to achieve a policy and an agent model able to generalize the experience acquired in the training process, we tested the achieved pedestrian model in some specific environments not included in the curriculum.

In particular, Figure 4 shows the *"Cross"* environment, in which the agent must reach two goals situated in opposite rooms before moving toward the exit situated in the upper part of the environment. The left diagram shows the trajectory achieved by adopting the built-in Unity AI navigation system [8] in which specific intermediate points are indicated in terms of absolute coordinates and the A* algorithm is used to achieve the trajectory. Strictly speaking, this is not comparable to the one achieved through our approach, since our agent does not know the coordinates of the subgoals. Moreover, the goal of our simulations is not to minimize the travel time, but to achieve trajectories (and walking speeds) that are close to those that a human pedestrian would follow. From this perspective, the results achieved in this scenario are interesting since the pedestrian does not make sharp bends to minimize the lenght of the trajectory reaching points very close to the walls, but it rather makes smoother bends, unless a 180° turn is the most sensible behavior (i.e. after reaching the subgoals). The diagram on the right, showing the distance among the points in space among the two agents throughout the simulation, highlights that the RL agent mostly differs in the observable behavior in the bends before reaching the first subgoal and after reaching the second one.

Figure 5 shows the same kind of diagrams and analyses in a much more complicated scenario, denoted as Supermarket, resembling a mini market in which different regions can include subgoals of interest to the agent. Here the agent must follow signposts associated to intermediate targets (passages among sections of the market) and pursue the subgoals associated to its own "shopping list". Also in this case, the RL agent generates a longer trajectory, avoiding getting too close to walls and taking smoother bends (in this case no 180° turn is plausible, and the only case in which a 90° bend would be reasonable the adopted trajectory is rather smooth).

These results by no means represent a complete validation of the model, but rather an illustration of the achieved results, which are promising although the approach present significant limits (first of all, unlike in [4] social aspects and the fact that multiple pedestrians can be

[8] See https://docs.unity3d.com/Packages/com.unity.ai.navigation@1.1/manual/NavInnerWorkings.html

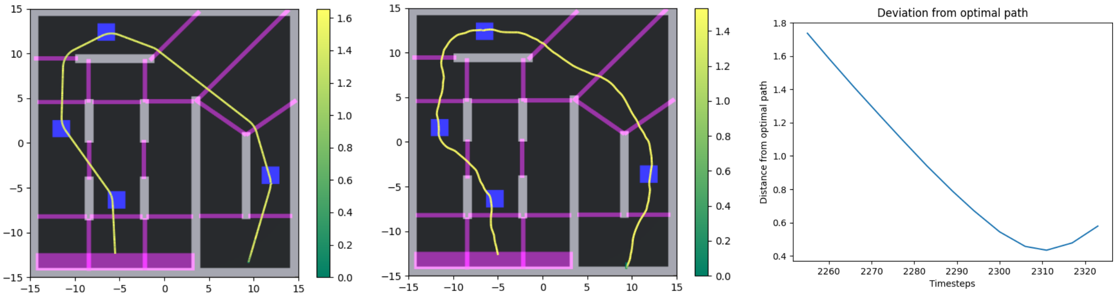


Figure 5: Supermarket environment execution: baseline optimal path on the left, trained agent in the middle, spatial deviation on the right.

present at the same time in the same environment were not considered in this experiment).

6. Towards an Explainable and Systematic Training Process?

The proposed approach based on the exploitation of a curriculum aims at creating a unique pedestrian model able to make the agent move in an unknown environment, exploiting its previous experience in simpler scenarios: the choice of these simple scenarios is of paramount importance, since they represent a set of "basic, archetypal situations" that ideally would be a good representation of (parts of) any unknown environment.

A problem in exploiting this approach is that we get a unique final model, where it is impossible to identify the part of the model related to a specific scenario: this means that if the model performs poorly in new unknown environments, it is plausible to consider evaluating if it is a situation that the present curriculum does not face sufficiently well or skips at all. This means that either a scenario already present in the curriculum must be changed (either in spatial structure or other relevant parameters, such as the threshold for completion), or a brand new scenario must be conceived and inserted in the curriculum. This also implies identifying the proper position in the sequence of scenarios. Of course, then, the whole training process must be executed again to achieve a new policy, a new behavioral model for the agent. This obviously represents a problem, both for sake of technological transfer, and especially should this approach be adopted in a distributed computing setting or (even worse) in a federated learning [21] context.

Our proposal to tackle these issues requires the definition of a semantic and explicit formulation of the notion of scenario, intended goals (supporting the acquisition of a given competence) and difficulties proposed to the agent in training. This basic element could first of all make visible the *design* process in the definition of the overall curriculum, making it more understandable for the modeler and for future sharing. First, this change would enable an explanation of the training process and a sort of justification of why the trained agent can perform some kind of behavioral pattern. Second, in case of implausible or problematic agent performance in some new environment, it could support the identification of scenario(s) to be changed to better face the new challenges strengthening some trained competence, or the creation of a new scenario to face newly identified training goals. Of course this implies remaining in the same deep structure of the model (agent's perceptions, actions, environmental building blocks).

In the long run, this could even lead to the conception of an automated process, integrated in the overall knowledge of the agent, supporting the self-revision of the training process or even a federated learning scenario in which multiple training agents could share experiences.

A longer term and even more challenging research direction, in the vein of recent proposals in cognitive sciences [22], would lead to the definition of a more articulated agent architecture comprising a knowledge-level component (potentially supported by sub-symbolic components) (i) inspecting the environment (or portion of environment) the agent is situated into and identifying the current situation, (ii) selecting / retrieving the most appropriate decision making model to be employed from a repertoire of pre-trained behavioral models (even achieved by means of different Machine Learning techniques or any inference mechanism for deciding how to move within a certain environmental condition). Also, potentially the agent could share experiences / models within a community of agents of this type by means of some collaboration mechanism, exploiting other's agents experience in different environments (so, improving the obtained behavioral models) and speeding-up the learning phase.

7. Conclusions

The paper has presented a research effort aimed at experimenting the adequacy of applying RL techniques to pedestrian simulation, especially considering the need to achieve general models applicable to a wide range of situations without the need of performing a training for each analyzed scenario. In addition to the longer term goals discussed in the previous section, a list of considerations of more immediate concern can be provided:

- we did not show a quantitative analysis of the achieved results, also for sake of space: this analysis, representing a first step in the direction of model validation, is object of current and future works;
- an extensive analysis of the effects of changes in RL algorithm, hyperparameters, configuration of the 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 work;
- additional quantitative experiments to improve the evaluation of the achieved results on the side of pedestrian simulation may be performed, possibly with a comparison with results of experimental observations, towards a validation of the model;
- overcoming some current limits in the expressiveness of the model: we focused here on wayfinding, while in a previous work we achieved a more general model for pedestrian operation decisions [4]. Group presence and social influence [23], for instance, were not considered.

Acknowledgements

Giuseppe Vizzari and Daniela Briola carried out this work within the context of MOST - Sustainable Mobility National Research Center and received funding from the European Union—Next-

GenerationEU (National Sustainable Mobility Center CN00000023, Italian Ministry of University and Research Decree n. 1033—17/06/2022, Spoke 8).

References

[1] M. Haghani, The notion of validity in experimental crowd dynamics, International Journal of Disaster Risk Reduction 93 (2023) 103750. URL: https://www.sciencedirect.com/science/article/pii/S2212420923002303. doi:https://doi.org/10.1016/j.ijdrr.2023.103750.

[2] F. Martinez-Gil, M. Lozano, F. Fernández, Emergent behaviors and scalability for multi-agent reinforcement learning-based pedestrian models, Simulation Modelling Practice and Theory 74 (2017) 117–133. URL: https://www.sciencedirect.com/science/article/pii/S1569190X17300503. doi:https://doi.org/10.1016/j.simpat.2017.03.003.

[3] R. S. Sutton, A. G. Barto, Reinforcement Learning, an Introduction (Second Edition), MIT Press, 2018.

[4] G. Vizzari, T. Cecconello, Pedestrian simulation with reinforcement learning: A curriculum-based approach, Future Internet 15 (2023). URL: https://www.mdpi.com/1999-5903/15/1/12. doi:10.3390/fi15010012.

[5] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 41–48. URL: https://doi.org/10.1145/1553374.1553380. doi:10.1145/1553374.1553380.

[6] G. Vizzari, L. Crociani, S. Bandini, An agent-based model for plausible wayfinding in pedestrian simulation, Engineering Applications of Artificial Intelligence 87 (2020) 103241. URL: https://www.sciencedirect.com/science/article/pii/S0952197619302246. doi:https://doi.org/10.1016/j.engappai.2019.103241.

[7] D. Helbing, P. Molnár, Social force model for pedestrian dynamics, Phys. Rev. E 51 (1995) 4282–4286. URL: https://link.aps.org/doi/10.1103/PhysRevE.51.4282. doi:10.1103/PhysRevE.51.4282.

[8] A. Schadschneider, W. Klingsch, H. Klüpfel, T. Kretz, C. Rogsch, A. Seyfried, Evacuation Dynamics: Empirical Results, Modeling and Applications, in: R. A. Meyers (Ed.), Encyclopedia of Complexity and Systems Science, Springer, 2009, pp. 3142–3176.

[9] E. Andresen, M. Chraibi, A. Seyfried, A representation of partial spatial knowledge: a cognitive map approach for evacuation simulations, Transportmetrica A: Transport Science 14 (2018) 433–467. URL: https://doi.org/10.1080/23249935.2018.1432717. doi:10.1080/23249935.2018.1432717. arXiv:https://doi.org/10.1080/23249935.2018.1432717.

[10] R. Junges, F. Klügl, Programming agent behavior by learning is simulation models, Applied Artificial Intelligence 26 (2012) 349–375. doi:10.1080/08839514.2012.652906.

[11] A. Tordeux, M. Chraibi, A. Seyfried, A. Schadschneider, Prediction of pedestrian dynamics in complex architectures with artificial neural networks, Journal of Intelligent Transportation Systems 24 (2020) 556–568. URL: https:

//doi.org/10.1080/15472450.2019.1621756. doi:10.1080/15472450.2019.1621756. arXiv:https://doi.org/10.1080/15472450.2019.1621756.

[12] X. Zhao, L. Xia, J. Zhang, W. Song, Artificial neural network based modeling on unidirectional and bidirectional pedestrian flow at straight corridors, Physica A: Statistical Mechanics and its Applications 547 (2020) 123825. URL: https://www.sciencedirect.com/science/article/pii/S0378437119321272. doi:https://doi.org/10.1016/j.physa.2019.123825.

[13] C. Cheng, A. Kolobov, A. Swaminathan, Heuristic-guided reinforcement learning, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 13550–13563. URL: https://proceedings.neurips.cc/paper/2021/hash/70d31b87bd021441e5e6bf23eb84a306-Abstract.html.

[14] L. Crociani, G. Vizzari, S. Bandini, Modeling environmental operative elements in agent-based pedestrian simulation, Collective Dynamics 5 (2020) 508–511. URL: https://collective-dynamics.eu/index.php/cod/article/view/A85. doi:10.17815/CD.2020.85.

[15] S. Paris, S. Donikian, Activity-driven populace: A cognitive approach to crowd simulation, IEEE Computer Graphics and Applications 29 (2009) 34–43. doi:10.1109/MCG.2009.58.

[16] M. Haghani, M. Sarvi, Imitative (herd) behaviour in direction decision-making hinders efficiency of crowd evacuation processes, Safety Science 114 (2019) 49–60. URL: https://www.sciencedirect.com/science/article/pii/S0925753518309275. doi:https://doi.org/10.1016/j.ssci.2018.12.026.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, CoRR abs/1707.06347 (2017). URL: http://arxiv.org/abs/1707.06347. arXiv:1707.06347.

[18] D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, Curiosity-driven exploration by self-supervised prediction, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 2778–2787. URL: http://proceedings.mlr.press/v70/pathak17a.html.

[19] F. L. D. Silva, A. H. R. Costa, A survey on transfer learning for multiagent reinforcement learning systems, Journal of Artificial Intelligence Research 64 (2019) 645–703. doi:10.1613/jair.1.11396.

[20] B. Baker, I. Kanitscheider, T. M. Markov, Y. Wu, G. Powell, B. McGrew, I. Mordatch, Emergent tool use from multi-agent autocurricula, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkxpxJBKwS.

[21] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, Knowledge-Based Systems 216 (2021) 106775. URL: https://www.sciencedirect.com/science/article/pii/S0950705121000381. doi:https://doi.org/10.1016/j.knosys.2021.106775.

[22] H. Mercier, D. Sperber, The Enigma of Reason, Harvard University Press, Cambridge, Massachussets, 2017.

[23] L. Crociani, Y. Zeng, G. Vizzari, S. Bandini, Shape matters: Modelling, calibrating and validating pedestrian movement considering groups, Simulation Modelling Prac-

tice and Theory 87 (2018) 73–91. URL: https://www.sciencedirect.com/science/article/pii/S1569190X18300844. doi:https://doi.org/10.1016/j.simpat.2018.06.001.