Scaffolding Deep Reinforcement Learning Agents using Dynamical Perceptual-Motor Primitives

Gaurav Patil (gaurav.patil@mq.edu.au)
School of Psychological Sciences & Centre for Elite Performance, Expertise and Training, Macquarie University, Sydney, Australia 2109

Patrick Nalepka (patrick.nalepka@mq.edu.au)
School of Psychological Sciences & Centre for Elite Performance, Expertise and Training, Macquarie University, Sydney, Australia 2109

Hamish F. Stenning (hamish.stenning@hdr.mq.edu.au)
School of Psychological Sciences, Macquarie University, Sydney, Australia 2109

Rachel W. Kallen (rachel.kallen@mq.edu.au)
School of Psychological Sciences & Centre for Elite Performance, Expertise and Training, Macquarie University, Sydney, Australia 2109

Michael J. Richardson (michael.j.richardson@mq.edu.au)
School of Psychological Sciences & Centre for Elite Performance, Expertise and Training, Macquarie University, Sydney, Australia 2109

Abstract

Agents trained using deep reinforcement learning (DRL) are capable of meeting or exceeding human-levels of performance in multi-agent tasks. However, the behaviors exhibited by these agents are not guaranteed to be human-like or human-compatible. This poses a problem if the goal is to design agents capable of collaborating with humans in cooperative or team-based tasks. Previous approaches to encourage the development of human-compatible agents have relied on pre-recorded human data during training. However, such data is not available for the majority of everyday tasks. Importantly, research on human perceptual-motor behavior has found that task-directed behavior is often low-dimensional and can be decomposed into a defined set of dynamical perceptual-motor primitives (DPMPs). Accordingly, we propose a hierarchical approach to simplify DRL training by defining the action dynamics of agents using DPMPs at the lower level, while using DRL to train the decision-making dynamics of agents at the higher level. We evaluate our approach using a multi-agent shepherding task used to study human and multi-agent coordination. Our hierarchical, DRL-DPMP approach resulted in agents which trained faster than vanilla, black-box DRL agents. Further, the hierarchical agents reached higher levels of performance not only when interacting with each other during self-play, but also when completing the task alongside agents embodying models of novice and expert human behavior. Finally, the hierarchical DRL-DPMP agents developed decision-making policies that outperformed heuristic-based agents used in previous research in human-agent coordination.

Keywords: hierarchical deep reinforcement learning; dynamical perceptual-motor primitives (DPMPs); multi-agent coordination; emergent coordination; shepherding

Introduction

Recent advances in model-free Artificial Intelligence (AI) and Deep Reinforcement Learning (DRL) techniques (Berner et al., 2019; Mnih et al., 2015; Pohlen et al., 2018; Vinyals et al., 2019) have resulted in artificial agents capable of meeting or exceeding human levels of performance. Most notable has been the success of DRL agents learning to perform complex individual or multi-agent video games (e.g., Atari 2600 games (Bellemare, Naddaf, Veness, & Bowling, 2013), DOTA (Berner et al., 2019), Starcraft II (Vinyals et al., 2019)). In many cases, however, the success of these DRL agents requires a complex, highly tuned, and task-specific structure of DRL methodologies and neural-network architectures, along with long and computationally intensive self-play training schemes (Berner et al., 2019; Vinyals et al., 2019). Moreover, even after constraining the action space of DRL agents to match human response limitations (Berner et al., 2019), the behavior of the DRL agents is often qualitatively different from humans (Shek, 2019; Carroll et al., 2019; Rigoli, Patil, Stening, Kallen, & Richardson, 2021), resulting in less fluid patterns of interaction (Nalepka, Gregory-Dunsmore, Simpson, Patil, & Richardson, 2021). Although this is not a problem if the goal is to achieve optimal task performance, it poses a major challenge when the aim is to develop DRL agents capable of effective human-AI agent interaction. Indeed, effective human performance in multi-agent task contexts requires that co-actors behave reciprocally, can anticipate each other's behaviors, and can readily perceive each other's behavioral intentions (Carroll et al., 2019). Thus, developing methods that produce DRL agents that are capable of human-like behavior and robust human-centered coordination and collaboration is often essential.

One way to improve the "human-like" quality of DRL agents is to employ pre-recorded human data or real-time human gameplay during the training process; e.g., behavior cloning (Bain & Sammut, 1999), generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016), or oracle learning (Maclin, Shavlik, Torrey, Walker, & Wild, 2005). The use of human data to pre-train AI agents helps to scaffold the essential "dynamics of gameplay" (e.g., basic action and

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

coordination patterns that lead to preliminary levels of task success), both ensuring effective task learning and decreasing training time (Amodei et al., 2016). Alternatively, human data can be used indirectly by exposing DRL agents to agents embodying models of human gameplay data during training (Carroll et al., 2019). In this way, DRL agents are encouraged to develop policies that complement human-like constraints. Unfortunately, these methods and approaches rely on the availability of large datasets of human gameplay, which are not readily available for most tasks (both real and computer based) and can suffer sharp performance declines when expert data is sparse or imperfect (Osa et al., 2018), or if agents interact with humans whose behaviors are not consistent with the distribution of the training data (Carroll et al., 2019).

Dynamical Perceptual-Motor Primitives

Research on human behavioral dynamics, however, has revealed that human movements typically reflect the context-specific realization of low-dimensional principles. Specifically, a growing body of research (Kelso, 1995; Nalepka et al., 2019; Nalepka, Silva, et al., 2021; Patil, Nalepka, Kallen, & Richardson, 2020; M. J. Richardson et al., 2016; Warren, 2006) has revealed that the spatiotemporal patterning of the actions that define human performance and decision-making can be modelled using a small, fundamental set of dynamical primitives (i.e., nonlinear dynamical functions). For instance, research has shown that these dynamical primitives can be employed to model human reaching, object passing, rhythmic wiping, cranking tasks (Kay, Kelso, Saltzman, & Schöner, 1987), goal-directed human navigation within an obstacle-ridden environment (Fajen, Warren, Temizer, & Kaelbing, 2003; Rigoli et al., 2021), and drumming (Ijspeert, Nakanishi, Hoffmann, Pastor, & Schaal, 2013) and racket ball tasks (Sternad, Duarte, Katsumata, & Schaal, 2001). The dynamical primitives used specifically to model human perceptual-motor behaviors are hereby termed as dynamical perceptual-motor primitives (DPMPs).

Relatedly, previous research has also shown how reinforcement learning can be used to model and parameterize dynamical primitives that can generate simple human motor behaviors (Ijspeert et al., 2013; Peters & Schaal, 2008). In multi-agent task contexts which require individuals to coordinate their actions physically and temporally to collectively influence the environment (Repp & Keller, 2004; R. C. Schmidt & Richardson, 2008), previous research has shown how stable patterns of such coordination naturally emerge as a result of the changing physical and informational couplings between the agents and the environmental constraints (Lagarde, 2013; M. Richardson, Marsh, & Schmidt, 2010; R. Schmidt & O'Brien, 1997; Nalepka, Silva, et al., 2021). Research has further shown that the same DPMPs used to model human perceptual-motor behaviors can also be employed to model the dynamics of numerous complex multi-agent tasks, including cooperative object pick-and-place tasks (Lamb et al., 2019) and goal-directed multi-agent navigation and collision avoidance behaviors (Warren, 2018), as well as multi-agent shepherding behavior (Nalepka, Kallen, Chemero, Saltzman, & Richardson, 2017; Nalepka et al., 2019; Rigoli et al., 2022).

Current Study

If the overarching aim is to replicate human-like behaviors for human-AI coordination, the implication of the research in DPMPs is that the processes used to train artificial agents by DRL should entail the same low-dimensional principles that characterizes human goal-directed behavior. Indeed, previous research has demonstrated, in a collaborative continuous control problem, that human participants exhibited higher levels of performance when working alongside an agent whose movements were constrained by DPMPs as opposed to one trained using DRL (Rigoli et al., 2022). Additionally, participants preferred to interact with these DPMP agents compared DRL agents.

Given these results, we present a hierarchical DRL-DPMP approach to training agents which can harness the power for DRL to generalize over a wide set of state-action-reward scenarios, while constraining its behavior to emulate the characteristics of human behavior. Our approach stems from the hierarchical nature of human actions such that goal-directed behavior can be split into a problem of *action selection* (which action to perform) and *action dynamics* (how to perform the action). By constraining the action dynamics of agents to conform to the dynamics of DPMPs, we can simplify training by only requiring agents to develop their own policies to identify the goals of their actions.

To illustrate this approach, we utilize the cooperative continuous control problem used in Rigoli et al. (2022), which is a modified version of the 'human shepherding task' employed in previous research investigating human multi-agent coordination (Nalepka et al., 2017, 2019). In this task paradigm, players controlling herding agents (HAs) are tasked to corral and contain a set of evasive target agents (TAs) within a red containment area located in the center of the field (see Fig. 1). In this task, the *action selection* component can be summarized by which TA each HA should pursue at each timestep, while the *action dynamics* component can be modelled using nonlinear mass-spring functions (see below for more details) parameterized to exhibit fixed point dynamics (where the selected TA is the system's terminal position). Previous research has demonstrated that agents whose behaviors were modelled using this DPMP model can exhibit human-like behaviors and can collaborate with humans in corralling TAs in a range of shepherding contexts (Nalepka et al., 2019; Rigoli et al., 2020).

In this study, we train and evaluate the proposed hierarchical DRL-DPMP agent methodology with agents which are required to develop separate policies for *action selection* and *action dynamics*. We compare the training

time required to train either agents, as well as evaluate their performance in completing the shepherding task when working alongside agents embodying human models of novice and expert behavior.

Method

Task and Environment

We employed a shepherding task previously used to study human and human-agent multi-agent coordination (Nalepka et al., 2017, 2019; Rigoli et al., 2020; Nalepka, Silva, et al., 2021). The task consists of two 'herding agents' (HAs), which are tasked to corral and contain a set of 'target agents' (TAs) within a red containment region located on the game field (see (Patil, Nalepka, Rigoli, Kallen, & Richardson, 2021) and Fig 1). The TAs exhibited Brownian motion, and fled from the HAs when within a critical distance. The task was deemed successful if the HAs could contain the TAs within the region for a specified period. The shepherding environment was developed using the Unity game engine (Unity Technologies, San Francisco, USA) and the DRL agents were implemented using the Unity ML-Agents package (Juliani et al., 2018). The environment size was set to 1m $\times$ 1.8m with two HAs corralling four TAs which spawned randomly in a $\pm$0.3m $\times$ $\pm$0.6m rectangle at the center of the field. The task goal was for the HAs to contain the TAs continuously for 5 seconds while each trial lasted 120 seconds. The velocity of the HAs was limited to 1 m/s in each direction and the rest of the parameters, e.g., TA speed, TA repulsion from the HAs, HA repulsion threshold, were set according to previous research (Nalepka et al., 2019).

Heuristic DPMP Models

Previous research using the shepherding task with humans indicated two strategies adopted by participants. The first, referred to as search & recover (S&R), involved each participant selecting the TA farthest from the containment region, which was also closer to them than their partner, and repelling it towards the center. The second, referred to as oscillatory containment (OSC), involved both participants encircling the TAs by making oscillatory movements around the whole herd to keep them contained. Not all participants discover OSC behavior, but for those who do, its use led to superior levels of performance (Nalepka et al., 2017). More recent work validated a model that accounts for S&R, OSC, and the transitions between these behaviors which was embodied in an artificial agent tasked to complete the task with humans (Nalepka et al., 2019).

The behavior exhibited by human players can be modelled by using the DPMP based task dynamic model,

$$\ddot{r_i} + b_r \dot{r_i} + k_r(r_i - (r_{T,i} + r_{min})) = 0 \tag{1}$$

and

$$\ddot{\theta_i} + b_\theta \dot{\theta_i} + \beta \dot{\theta_i^3} + \gamma \theta_i \dot{\theta_i} + k_\theta(\theta_i - \theta_{T,i}) = 0, \tag{2}$$

which model the radial distance and angle of each HA, respectively. $\ddot{r_i}$ and $\dot{r_i}$ in Eq. (1) represent the velocity and acceleration of HA-i's radial distance, respectively, $r_{T,i}$ the radial distance of the TA being pursued and r_{min} the fixed radial offset to ensure the HA repels the TA towards the containment area.$\ddot{\theta_i}$ and $\dot{\theta_i}$ in Eq. (1) represent the velocity and acceleration of HA-i's radial angle, respectively and $\theta_{T,i}$ the radial angle of the TA being pursued. k and b terms in both Eqs. (1) and (2) represent the corresponding stiffness and damping parameters that determine both the acceleration of the HA towards the position $(r_{T,i} + r_{min}, \theta_{T,i})$, and the opposition to the resultant velocity, respectively. Additionally, $\beta \dot{\theta_i^3}$ and $\gamma \theta_i \dot{\theta_i}$ in Eq. (2) are nonlinear Rayleigh and van der Pol escapement terms which capture the amplitude-frequency and peak velocity-frequency relationship exhibited by human actors (Kay et al., 1987). The model defined by Eqs. (1) and (2) is hereby termed as *Limited-DPMP* model since it can only exhibit S&R behavior (when $b_\theta > 0$). A *Full-DPMP* model can be created from Eqs. (1) and (2) by allowing the damping parameter b_θ to take a negative value by using

$$\dot{b_\theta^{HA_i}} + \delta(b_\theta^{HA_i} - \alpha(r_{T,i} - r_\Delta) = 0 \tag{3}$$

such that when the radial distance of the TA being pursued, $r_{T,i}$, is within a set distance of the containment location, r_Δ, the δ and α terms repeatedly reduce the value of $b_\theta^{HA_i}$ to a negative value. On the contrary, when $r_{T,i}$ is outside r_Δ, the value of $b_\theta^{HA_i}$ slowly goes back to being positive. This flip in the value of the damping term of the model that approximates the angular dynamics of the HA results in a Hopf bifurcation (Patil et al., 2020) that changes the point attractor behavior to a limit cycle behavior resulting in the HA exhibiting OSC behavior. The model defined by Eqs. (1), (2), and (3) is referred to as the *Full-DPMP* model since it can exhibit both S&R and OSC behaviors. In previous implementations, a coupling term was defined to couple the *Full-DPMP* to its partner. The success of the model does not depend on this coupling function and for this paper the coupling term was not included. For both models, the target selection decision was defined by a heuristic policy presented in (Rigoli et al., 2020), where for the subset of TAs that are closer to the HA than to its partner, the HA will select the TA that is farthest from the containment region.

In summary, the *Limited-DPMP* model emulates novice-like behavior prior to the discovery of OSC behavior, while the *Full-DPMP* model emulates human expert-like behavior capable of exhibiting both SR and OSC behavior.

DRL Models

A classical approach to applying DRL to a multi-agent problem like shepherding would be to use a single artificial neural network (ANN) which takes the states of all the TAs and HAs as inputs and determines the HA's actions. This approach can be further decentralized by using separate networks for target selection and action dynamics which

can be trained independently. The hierarchical DRL-DPMP model, first proposed in (Patil, Nalepka, et al., 2021), uses an ANN for the the target selection policy, with the action dynamics defined by the above DPMP model. Here, the *Limited-DPMP* agent was used to control the action dynamics, and received input from the ANN as to which target to pursue. Agents employing this approach were referred to as DRL-target selection, *DRL-TS*. In addition to the *DRL-TS* hierarchical model, a black-box model was also evaluated *DRL-BB* which implemented separate ANNs to make TA selection decisions as well as control the action dynamics.

Training

The *DRL-TS* model for each HA used a neural network with 3 densely connected hidden layers with 64 neurons each and took the states (position and velocity) of all TAs and HAs as inputs (24 inputs) and outputted a one-hot vector of the TA to pursue. The states of the TAs were ordered in an egocentric way, such that, the states of the TA closest to the HA occupied the earlier spots in the input array. The neural networks were trained according to the Proximal Policy Optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017) algorithm with observations collected every 15th frame for the *DRL-TS* model while the environment updated at 50 Hz. The ANN obtained experiences from nine game environments that were run in parallel.

A 11-step curriculum learning was implemented to train all the agents with agents advancing through the curriculum steps when 80% of the last 10 trials in the 9 environments (72 out of 90 trials) were successful. The following curriculum was used: all training runs started with 2 free TAs and 2 TAs clamped at the center with the repulsion and speed of TAs set such that they were easy to repel by the HAs but moved slowly upon repulsion and the goal containment time set to 0.5 s. The 2 TA curriculum step was introduced to aid in successful training of the DRL-BB agents. The repulsion factors slowly changed until step 5 such that they matched the conditions used in Rigoli et al. (2020). At step 6, all 4 TAs were free to move while the repulsion factors were set back to the same level as step 1. At step 7 the goal containment time was increased to 5 s and by step 11 the repulsion factors were changed in steps to the normal environmental conditions used in Rigoli et al. (2020). For the DRL-BB agents, in steps 1 to 5, where only 2 TAs were active, TA selection decisions were made by the heuristic defined for the *Full-DPMP* and *Limited-DPMP* agents. This was done in order to avoid any training bias of the ANN that approximated the TA selection due to the presence of only 2 TAs.

During training, the reward for both HAs was calculated in each environment update such that the agents received a negative ($0.01 \times$ distance of TA from center of environment) reward for every TA outside the containment area and positive 0.01 reward for every TA in the containment area. In addition, the HAs received a small negative reward (0.002) if both HAs were within 5 cm of each other in order to avoid the HAs from performing the same actions. Twenty *DRL-TS* and *DRL-BB* agents (40 total) were trained, with the top five for each methodology used for evaluation by ranking them by the average episode length in the last 0.25 million training steps.

Evaluation and Performance Measures

Following training, the end state performance of the top five agents was assessed over twenty trials. The following measures were used to quantify and compare herding task performance: 1) *trial time*: The amount of time (s) needed to contain the TAs - up to a maximum 120 s allowed for each trial; 2) *TA travel*: The average cumulative distance travelled (cm) by the TAs, normalized by trial duration; 3) *propOSC*: The proportion of a trial where the HAs exhibited oscillatory behavior. For this measure, a windowed frequency analysis (window size = 3 s) was used to determine whether an HA was exhibiting OSC behavior (with a peak frequency > 0.5 Hz; see (Rigoli et al., 2020)).

The top five DRL-TS and DRL-BB agents completed twenty trials in the following dyad configurations: self-play (where both HAs were controlled by the respective DRL agent), Limited-DPMP partner and Full-DPMP partner. For Limited- and Full-DPMP partner, each DRL agent completed the task alongside an agent implementing the Limited-DPMP (i.e., human novice-like behavior) and Full-DPMP (i.e., human expert-like behavior) models.

All analyses were conducted using the statistics program Stata 17.0 MP (StataCorp LLC, College Station, Texas). Multi-level linear (linear mixed-effects) models were fitted for each dependent measure, where each trial (observation) was nested under their respective trained agent. The models were defined using the procedure recommended by (Meteyard & Davies, 2020; Barr, Levy, Scheepers, & Tily, 2013). Specifically, the random-effects portion of the model was defined first and defined using a maximal to minimal-that-converges approach. In the approach, the models were first defined with parameters for the random-intercepts variance, for the random-slopes variances (one for each fixed effect: *agent type*, *partner type*, and the $agenttype \times partnertype$ interaction), and for each of the covariances between the random-slopes variances and the random-intercepts variance (i.e., unstructured covariance-variance structure). If a model could not be fit using maximum likelihood, the covariances, followed by any parameters for which standard errors were not calculated (indicating an invalid model fit), were removed from the model one at a time until the model was fit correctly. Using the resulting random structures, the one-way fixed effects were added. The two-way interaction was included if its inclusion improved model fit, as assessed by likelihood-ratio (LR) test. The resulting mixed models were then re-fitted using restricted maximum likelihood (REML) so that Kenward-Roger degrees of freedom for the fixed effects could be estimated (Kenward & Roger, 1997).

All dependent measures were treated at the dyad level. For the dependent measure *propOSC*, its value was averaged

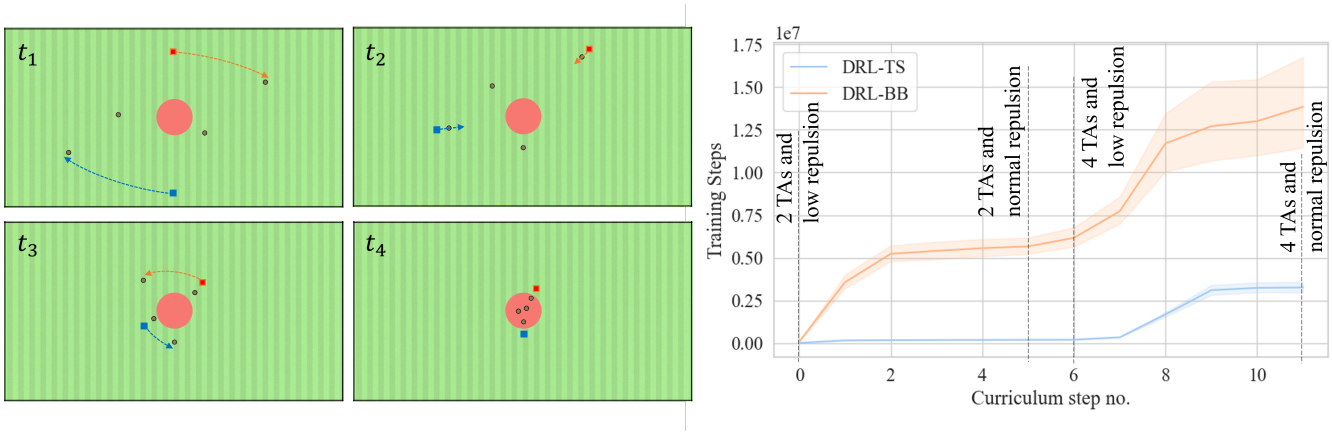

Figure 1: Task Environment and Training Time for the DRL Agents. (left) Example task progression for two herding agents (HAs) corralling four target agents (TAs). (right) Cumulative training steps needed to train the *DRL-TS* and *DRL-BB* agents through the curriculum (see text).

across both DRL agents in the self-play partner condition. For the Limited- and Full-DPMP conditions, only the value of the DRL agent was considered in the dyad. All pairwise comparisons following significant tests were Bonferroni corrected.

For both *trial time* and *TA travel*, the final models included random slopes for *partner type* but did not include random intercepts, random slopes for *agent type* nor the interaction, nor covariances. The *agent type*, *partner type*, and *agenttype* $\times$ *partnertype* fixed effects were all included in the model. For *propOSC*, the final model had the same random-effects structure but only included the main-effect fixed effects (i.e., *agent type* and *partner type*).

Results

Fig. 1 shows the comparison of training performance of DRL-TS and DRL-BB agents that successfully completed all curriculum steps. DRL-TS agents learned to successfully complete the shepherding task within 3.2 million timesteps (*SD* = 0.59 million), compared to an average of 13.84 million timesteps (*SD* = 5.8 million) for the DRL-BB agents. DRL-TS agents took less time to complete the training curriculum than DRL-BB agents, validating the expectation that using DPMP movement dynamics to control the action dynamics of agents would significantly simplify (and therefore speed-up) the agent training process. Additionally, the variability between the training steps required by all DRL-TS agents to complete the curriculum steps was much lower that the DRL-BB agents pointing to a higher level of stable learning patterns.

All performance data are presented in Fig. 2. For trial time, significant effects of *agent type*, $t(19.4) = -7.20, p < .0001$, and *partner type*, $F(2, 14.63) = 37.79, p < .0001$, were found, with dyads that included DRL-TS agents completing trials faster than dyads that included DRL-BB agents. Additionally, agents who completed the task with a similar partner (i.e., self-play) or with the Full-DPMP agent completed trials in a similar length of time ($t(11.6) = 1.29, p = .662$) and outperformed dyads who completed the task with the Limited-DPMP agent (both $t \leq -5.34, p < .0001$).

For *TA travel* (normalized by trial duration), significant effects of *agent type*, $t(19.8) = 6.84, p < .0001$ and *partner type*, $F(2, 14.44) = 162.00, p < .0001$, were observed as well as a significant *agenttype* $\times$ *partnertype* interaction, $F(2, 14.44) = 30.80, p < .0001$. In summary, TA travel was lowest for DRL agents during self-play compared to when the DRL agents worked alongside a Full-DPMP or Limited-DPMP partner, with the Limited-DPMP partner condition resulting in the largest magnitudes of TA travel (all comparisons $|t| > 5.12, p < .001$). Although TA travel was also larger for DRL-TS agents compared to DRL-BB agents during self-play ($t(578) = 16.41, p < .0001$), there were no differences between the DRL agents when completing the task with either the Limited- or Full-DPMP partners (both $t \leq 1.34, p \geq .655$).

Finally, the analysis of *propOSC* revealed a significant effect of *partner type*, $F(2, 21.04) = 21.04, p < .0001$, with DRL-TS agents exhibiting greater instances of oscillatory-like behavior compared to the DRL-BB agents, $t(40.9) = 4.15, p < .001$. Further, this behavior was reduced when the DRL agents completed the task with the Limited-DPMP agent, who was incapable of producing oscillatory behavior ($t(32.9) = -6.59, p < .001$). There was no difference between the self-play or Full-DPMP conditions ($|t| < 2.48, p \geq .079$).

Discussion

The results revealed that the DRL agents trained using the hierarchical DRL-DPMP methodology (i.e., DRL-TS agents) not only outperformed DRL agents trained using the standard DRL methods (i.e., DRL-BB agents), but were also able

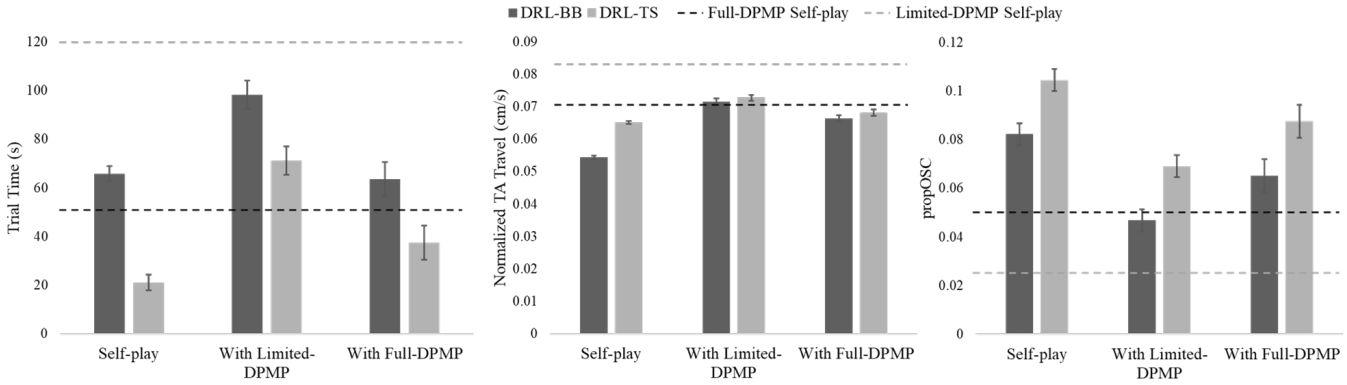

Figure 2: Summary of the results. (left) Trial time (center) Normalized TA travel (right) Proportion of a trial where the HAs exhibited oscillatory behavior. Plotted are the predictive margins from the fitted models. For comparisons, also included are the mean performance of two self-play DPMP models (Limited- and Full-DPMP).

to achieve effective task performance more than four times faster over the course of training. Recall that DRL-BB agents were required to develop separate policies for for both *action selection* and *action dynamics*. In contrast, the DRL-TS agents were trained to develop an *action selection* policy, while its action dynamics were constrained using DPMPs. Although it is not surprising that this configuration reduced training time, it is important to appreciate that agents trained using standard DRL methods often fail to learn to perform complex multi-agent tasks and that the use of DPMPs can provide a easy method for overcoming this issues. Further, in addition to (or instead of) the availability of human demonstration data, DPMP models could be used to generate synthetic data to augment existing (or to generate new) datasets for imitation learning approaches.

Because the DRL-TS agents' movements were constrained by a DPMP model that emulated the dynamics of human shepherding behavior, it is hypothesized that this would result in better human-agent collaboration as the movements of the DRL-TS agents would be aligned with human expectations. Future work will have to validate these agents with human participants, as a limitation of this work is that the DRL agents were evaluated alongside human models of novice and expert-level behavior. However, as demonstrated by Rigoli et al. (2022), participants prefer interacting with a DPMP-based shepherding model as compared to one trained using DRL. Therefore, it is expected that participants would equally prefer the DRL-TS agent as it exhibits movement dynamics consistent with the DPMP model, with a more flexible action selection policy, as opposed to the inflexible heuristic policy used in Rigoli et al. (2022).

Future work could also consider alternative methods to design the training environments of the DRL agents. In particular, the current study trained the DRL agents using self-play. However, an alternative approach, pursued by Carroll et al. (2019), is to insert models of human behavior (e.g., the Limited- or Full-DPMP agents) as agents within the training environment. In this way, DRL agents must develop policies which complement the behaviors of these DPMP models. It remains an open question to the extent to which DRL agents, in this setting, would adopt policies that resemble those of the DPMP agents, or if human participants would be accommodating to agents exhibiting heterogeneous strategies. Preliminary studies show that participants prefer heuristic models over the models trained by DRL (Patil, Bagala, Nalepka, Kallen, & Richardson, 2022; Patil, Bagala, Nalepka, Richardson, & Kallen, 2023). Additionally, it is also necessary to test the hierarchical approach in other task contexts where the movement dynamics can be modelled by DPMP models (Babajanyan, Patil, Lamb, Kallen, & Richardson, 2022; Ekdawi, Patil, Kallen, & Richardson, 2022; Patil, Rigoli, et al., 2021)

Interestingly, although the DRL-TS agents implemented the Limited-DPMP model to control its movements, the DRL-TS agents exhibited greater instances of oscillatory-like behavior compared to the DRL-BB agents. This is surprising because the Limited-DPMP model was parameterized to not exhibit oscillatory behavior (see Fig. 2). This suggests that the DRL-TS agents developed action selection policies that encouraged the emergence of this more effective behavioural strategy. Indeed, in model-based simulations, oscillatory behavior can be observed as an emergent property of systems that are only capable of point attractor dynamics (such as the case for the Limited-DPMP model) if certain action selection heuristics are used (Nalepka, Silva, et al., 2021). As the DRL-TS agents exhibited less control over the TAs compared to the DRL-BB agents, in terms of limiting their motion, it is possible that DRL-TS agents used the TAs' motion to its advantage to produce oscillatory-like dynamics. Indeed, the utilization of oscillatory, as well as circling strategies, is an effective mode of behavior observed not only in humans (Nalepka et al., 2017), but other animal systems in naturalistic shepherding and hunting contexts (Nalepka, Silva, et al., 2021). Similarly, during training, DRL-TS agents may be

learning policies which uncover these behaviors as latent solutions to the shepherding task context.

To conclude, agents trained using DRL are capable of learning policies that can meet or exceed human levels of performance. However, if the goal is to develop agents that can interact with humans seamlessly in collaborative, human-agent interaction, agents must exhibit behaviors that align with human expectations and behavioural strategies. Because agents trained using DRL require a vast amount of experiences, humans interacting with these agents directly during training is not practical (Carroll et al., 2019). Therefore, models of human behavior are needed to constrain the development of DRL agents. For perceptual-motor tasks specifically, the results of the present work demonstrate that this can be achieved using DPMPs, which represent the primitive features of human movement, to control the movements and actions of DRL agents. Such hybrid DRL-DPMP agents also significantly reduced the training time needed by reducing the number of dimensions the agent needs to control. For more complex and hierarchical tasks, agents exploiting models composed of DPMPs to constrain its actions is likely to outperform standard black-box DRL agents, while exhibiting dynamics that are human-compatible.

Acknowledgments

This work was supported by the Australian Department of Defence, Science and Technology (DST) group (partnership grant MyIP8655) and Human Performance Research Network (HPRNet, partnership grant ID9024). HFS was supported by the HPRNet grant. PN was supported by the Macquarie University Research Fellowship. MJR was supported by the Australian Research Council Future Fellowship (FT180100447). This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *CoRR*, *abs/1606.06565*.

Babajanyan, D., Patil, G., Lamb, M., Kallen, R. W., & Richardson, M. J. (2022). I know your next move: Action decisions in dyadic pick and place tasks..

Bain, M., & Sammut, C. (1999). A framework for behavioural cloning. In *Machine intelligence 15, intelligent agents [st. catherine's college, oxford, july 1995]* (p. 103–129). GBR: Oxford University.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278. doi: https://doi.org/10.1016/j.jml.2012.11.001

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013, may). The arcade learning environment: An evaluation platform for general agents. *J. Artif. Int. Res.*, *47*(1), 253–279.

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., ... Zhang, S. (2019). Dota 2 with large scale deep reinforcement learning. *CoRR*, *abs/1912.06680*.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T. L., Seshia, S. A., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. In *Proceedings of the 33rd international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Ekdawi, S., Patil, G., Kallen, R. W., & Richardson, M. J. (2022). Modelling competitive human action using dynamical motor primitives for the development of human-like artificial agents..

Fajen, B. R., Warren, W. H., Temizer, S., & Kaelbing, L. P. (2003). A dynamical model of visually-guided steering, obstacle avoidance, and route selection. *International Journal of Computer Vision*, *54*(1-3), 13–34. doi: 10.1023/a:1023701300169

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. , *29*.

Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. *Neural Computation*, *25*(2), 328–373. doi: 10.1162/NECO$_{a0}$0393

Juliani, A., Berges, V., Vckay, E., Gao, Y., Henry, H., Mattar, M., & Lange, D. (2018). Unity: A general platform for intelligent agents. *CoRR*, *abs/1809.02627*.

Kay, B. A., Kelso, J. A., Saltzman, E. L., & Schöner, G. (1987, may). Space–time behavior of single and bimanual rhythmical movements: Data and limit cycle model. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(2), 178–192. doi: 10.1037/0096-1523.13.2.178

Kelso, J. A. S. (1995). *Dynamic patterns : the self-organization of brain and behavior*. Cambridge, Mass: MIT Press.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997.

Lagarde, J. (2013). Challenges for the understanding of the dynamics of social coordination. *Frontiers in Neurorobotics*, *7*(OCT). doi: 10.3389/fnbot.2013.00018

Lamb, M., Nalepka, P., Kallen, R. W., Lorenz, T., Harrison, S. J., Minai, A. A., & Richardson, M. J. (2019, jun). A Hierarchical Behavioral Dynamic Approach for Naturally Adaptive Human-Agent Pick-and-Place Interactions. *Complexity*, *2019*, 1–16. doi: 10.1155/2019/5964632

Maclin, R., Shavlik, J., Torrey, L., Walker, T., & Wild, E. (2005). Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression. In *Proceedings of the 20th national conference on artificial intelligence - volume 2* (p. 819–824). AAAI Press.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. doi: https://doi.org/10.1016/j.jml.2020.104092

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015, feb). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi: 10.1038/nature14236

Nalepka, P., Gregory-Dunsmore, J. P., Simpson, J., Patil, G., & Richardson, M. J. (2021). Interaction Flexibility in Artificial Agents Teaming with Humans. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 112–118.

Nalepka, P., Kallen, R. W., Chemero, A., Saltzman, E., & Richardson, M. J. (2017, may). Herd Those Sheep: Emergent Multiagent Coordination and Behavioral-Mode Switching. *Psychological Science*, *28*(5), 630–650. doi: 10.1177/0956797617692107

Nalepka, P., Lamb, M., Kallen, R. W., Shockley, K., Chemero, A., Saltzman, E., & Richardson, M. J. (2019, jan). Human social motor solutions for human–machine interaction in dynamical task contexts. *Proceedings of the National Academy of Sciences*, *116*(4), 1437–1446. doi: 10.1073/pnas.1813164116

Nalepka, P., Silva, P. L., Kallen, R. W., Shockley, K., Chemero, A., Saltzman, E., & Richardson, M. J. (2021, nov). Task dynamics define the contextual emergence of human corralling behaviors. *PLOS ONE*, *16*(11), e0260046. doi: 10.1371/journal.pone.0260046

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., & Peters, J. (2018). An Algorithmic Perspective on Imitation Learning. *Foundations and Trends in Robotics*, *7*(1-2), 1–179. doi: 10.1561/2300000053

Patil, G., Bagala, P., Nalepka, P., Kallen, R. W., & Richardson, M. J. (2022). Evaluating human-artificial agent decision congruence in a coordinated action task. In *Proceedings of the 10th international conference on human-agent interaction* (p. 327–329). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3527188.3563923

Patil, G., Bagala, P., Nalepka, P., Richardson, M. J., & Kallen, R. W. (2023). Action decision congruence between human and deep reinforcement learning agents during a coordinated action task..

Patil, G., Nalepka, P., Kallen, R. W., & Richardson, M. J. (2020, aug). Hopf Bifurcations in Complex Multiagent Activity: The Signature of Discrete to Rhythmic Behavioral Transitions. *Brain Sciences*, *10*(8), 536. doi: 10.3390/brainsci10080536

Patil, G., Nalepka, P., Rigoli, L., Kallen, R. W., & Richardson, M. J. (2021). Dynamical Perceptual-Motor Primitives for Better Deep Reinforcement Learning Agents. In *Advances in practical applications of agents, multi-agent systems, and social good. the paams collection. paams 2021. lecture notes in computer science()* (Vol. 12946, pp. 176–187). Springer Science and Business Media Deutschland GmbH. doi: 10.1007/978-3-030-85739-4_15

Patil, G., Rigoli, L., Wahlin, C., Nalepka, P., Kallen, R. W., & Richardson, M. J. (2021). Perceptual sensitivity to an artificial co-actor in competitive 2d pong..

Peters, J., & Schaal, S. (2008, may). Reinforcement learning of motor skills with policy gradients. *Neural Networks*, *21*(4), 682–697. doi: 10.1016/J.NEUNET.2008.02.003

Pohlen, T., Piot, B., Hester, T., Azar, M. G., Horgan, D., Budden, D., ... Pietquin, O. (2018). Observe and look further: Achieving consistent performance on atari. *CoRR*, *abs/1805.11593*.

Repp, B. H., & Keller, P. E. (2004, apr). Adaptation to tempo changes in sensorimotor synchronization: Effects of intention, attention, and awareness. *The Quarterly Journal of Experimental Psychology Section A*, *57*(3), 499–521. doi: 10.1080/02724980343000369

Richardson, M., Marsh, K., & Schmidt, R. (2010). Challenging the egocentric view of coordinated perceiving, acting, and knowing. In B. Mesquita, L. Feldman Barrett, & E. Smith (Eds.), *The mind in context* (pp. 307–333). United States: Guilford Press.

Richardson, M. J., Kallen, R. W., Nalepka, P., Harrison, S. J., Lamb, M., Chemero, A., ... Schmidt, R. C. (2016). Modeling Embedded Interpersonal and Multiagent Coordination. In V. M. Muñoz, O. Gusikhin, & V. Chang (Eds.), *Proceedings of the 1st international conference on complex information systems* (pp. 155–164). Setúbal, Portugal: SCITEPRESS - Science and and Technology Publications. doi: 10.5220/0005878101550164

Rigoli, L., Nalepka, P., Douglas, H., Kallen, R. W., Hosking, S., Best, C., ... Richardson, M. J. (2020). Employing models of human social motor behavior for artificial agent trainers. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems* (p. 1134–1142). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Rigoli, L., Patil, G., Nalepka, P., Kallen, R. W., Hosking, S., Best, C., & Richardson, M. J. (2022, apr). A Comparison of Dynamical Perceptual-Motor Primitives and Deep Reinforcement Learning for Human-Artificial Agent Training Systems. *Journal of Cognitive Engineering and Decision Making*, 155534342210929. doi: 10.1177/15553434221092930

Rigoli, L., Patil, G., Stening, H. F., Kallen, R. W., & Richardson, M. J. (2021, sep). Navigational Behavior of Humans and Deep Reinforcement Learning Agents. *Frontiers in Psychology*, *12*, 4096. doi: 10.3389/fpsyg.2021.725932

Schmidt, R., & O'Brien, B. (1997, sep). Evaluating the Dynamics of Unintended Interpersonal Coordination. *Ecological Psychology*, *9*(3), 189–206. doi: 10.1207/s15326969eco0903$_2$

Schmidt, R. C., & Richardson, M. J. (2008). Dynamics

of Interpersonal Coordination. , *2008*, 281–308. doi: 10.1007/978-3-540-74479-5$_1$4

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR*, *abs/1707.06347*.

Shek, J. (2019). *Takeaways from openai five (2019) [ai/ml, dota summary]*.

Sternad, D., Duarte, M., Katsumata, H., & Schaal, S. (2001). Bouncing a ball: Tuning into dynamic stability. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1163–1184. doi: 10.1037/0096-1523.27.5.1163

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., … Silver, D. (2019, nov). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), 350–354. doi: 10.1038/s41586-019-1724-z

Warren, W. H. (2006, apr). The dynamics of perception and action. *Psychological Review*, *113*(2), 358–389. doi: 10.1037/0033-295X.113.2.358

Warren, W. H. (2018, aug). Collective Motion in Human Crowds. *Current Directions in Psychological Science*, *27*(4), 232–240. doi: 10.1177/0963721417746743