Scuola Superiore Meridionale
University of Naples Federico II

Doctoral Thesis
in Modeling and Engineering Risk and Complexity

Collective decision making in multi-agent systems with applications to human-AI interaction

Author:
Ayman bin Kamruddin

Supervisors:
Prof. Michael J. Richardson
Prof. Mirco Musolesi
Prof. Mario di Bernardo

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Modeling and Engineering Risk and Complexity. Coordinator: Prof. Mario di Bernardo.

December 16, 2024

ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ

Praise belongs to the God, Lord of the Worlds
(the Holy Qur'an, Chapter 1, Verse 1/2)

Abstract

The research presented in this thesis aimed to model human behaviour in complex perceptual-motor tasks. It starts by investigating whether dynamical perceptual-motor primitives (DPMPs) could also be used to capture human navigation in a first-person herding task. To achieve this aim, human participants played a first-person herding game, in which they were required to corral virtual cows, called target agents, into a specified containment zone. In addition to recording and modelling participants' movement trajectories during gameplay, participants' target selection decisions (i.e., the order in which participants corralled targets) were recorded and modelled. The results revealed that a simple DPMP navigation model could effectively reproduce the movement trajectories of participants and that almost 80% of the participant's target selection decisions could be captured by a simple heuristic policy. Importantly, when this policy was coupled to the DPMP navigation model, the resulting system could successfully simulate and predict the behavioural dynamics (movement trajectories and target selection decisions) of participants in novel multi-target contexts.

Building on these findings, the study further explored multi-herder and multi-agent herding scenarios, which introduce additional layers of complexity. In these scenarios, human participants coordinated their movements with one another, dynamically adjusting their strategies based on the behaviour of both their fellow herder and the target agents. Modelling results revealed that human herding behaviour in these multi-agent contexts could be replicated using simple control rules and decision-making processes. Artificial agents were then developed - using these models and additionally, Deep Reinforcement Learning techniques - in order to cooperate with human agents to complete the multiagent herding ask. By integrating multi-herder dynamics into the model, the study provides insights into how artificial agents could be designed to work alongside humans in tasks requiring collaborative joint action, such as search and rescue operations, crowd management, or autonomous driving systems, where multiple entities must cooperate to achieve shared goals.

Acknowledgements

I would like to thank Professor Michael ("Mike") Richardson, for his incredible help, vast amounts of time dedicated, and valuable guidance in seeing this project through to fruition. Thank you, Mike, for being so accommodating and helpful. I have always appreciated your open-door policy, and even more your open-arms policy. Thank you for being so welcoming, supportive and patient. Whether it was through coding some of the harder bits in the earlier stages or advice and direction in general, you've been awesome throughout the entire process.

I would next like to thank Professor Mario di Bernardo for his valuable comments and feedback throughout the process. Thanks, Mario, for being so flexible with regards to my particular needs. I would also like to thank Professor Mirco Musolesi for enlightening discussions had throughout my PhD journey.

Thanks to all the honours students who have assisted along the way - Chris Lam, Sala Ghanem, and Hannah Sandison. Data collection would have been a nightmare for me without you, and I hope that you have gained some value from our interactions in return. Thanks also to Dr. Gaurav Patil for helping build solutions and resolve issues along the way.

Thanks to mum ("ammi") and pop ("abba") for supporting me through this journey with their love and kind words. Lastly, and most certainly not the least, I'd like to give thanks to my wife, partner and best friend, Najoua, for her unwavering support - I've written this doctoral thesis but I'd happily have her take the doctoral title.

Contents

Abbreviations

ODE: Ordinary Differential Equation(s)
PDE: Partial Differential Equation(s)
AI: Artificial Intelligence
AA: Artificial Agent
DPMP: Dynamic Perceptual-Motor Primitive
HA: Herder Agent
TA: Target Agent
TS: Target Selection
SCA: Successive Collinearity by Angle
SCD: Successive Collinearity by Distance
RL: Reinforcement Learning
DRL: Deep Reinforcement Learning
NN: Neural Network
ANN: Artificial Neural Network
DNN: Deep Neural Network
PPO: Proximal Policy Approximation
DDPG: Deep Deterministic Policy Gradient
MARL: Multi-Agent Reinforcement Learning
TSp: Target Selection policy
HS-DRL: Human-sensitive DRL (policy)
SP-DRL: Self-play DRL (policy)

1 Introduction

1.1 Context of the research topic

Animals may well be among the most complex entities in existence. Composed of elementary particles, they exhibit increasing layers of complexity at every organisational level—from particles to atoms, from atoms to molecules, from molecules to cells, from cells to organs, and finally, from organs to the entire being. Each level adds more intricacy, challenging the notion that reductionism can fully explain the behaviour of complex systems [1].

Yet, amidst this complexity, simple laws governing interactions—whether between agents, between agents and their environment, or both—can yield powerful explanations and predictions of behaviour. Importantly, an agent need not possess consciousness or intelligence akin to humans for complex behaviours to emerge. Ants build bridges, bees coordinate foraging, and termites construct elaborate mounds, all explainable via simple, self-organising dynamical rules. Indeed, these examples highlight how simple dynamical rules, once discovered, can explain behaviour that appears greater than the sum of its parts. Thus, despite the myriad components making up intentional agents, their behaviour can still be predicted and understood from mathematical laws.

With regards to understanding human action and perceptual-motor behaviour, a significant shift occurred in the 1980s with the work of Scott Kelso, Peter Kugler, and Michael Turvey and colleagues (e.g., [2, 3, 4]). Of particular relevance here, was the task dynamics framework outline by Saltzman and Kelso [2]. Here they provided the first detailed account of how the mathematics of dynamical systems could be used to understand human perceptual-motor behaviour, utilising simple coupled ordinary differential equations to model fundamental human movements, such as arm extensions, wheel cranking, or reaching for a cup. In the 2000s, these models were then extended to capture a wide range of human navigational movements [5, 6, 7, 8, 9], forming the foundation of the behavioural dynamics approach to modelling human perceptual-motor behaviours [7]. Not only were such task-dynamical models applicable to human navigation, but they were widely applied to interpersonal coordination and joint action [10, 11, 12] - providing understanding that has been central to the modelling of perceptual-motor behaviour in more complex tasks [13, 14, 15, 16] similar to the ones presented in this thesis. The research underpinning this thesis thus leverages both these navigation models, along with the task/behavioural dynamics approach and modern computational

(deep learning) techniques to model, understand, and predict human behaviour in first-person herding tasks.

Herding tasks—where one or more herder agents must capture, corral, and contain one or more target agents within a predefined containment zone [17]—may initially appear niche. However, they provide a rich testbed for studying human cooperation and coordination. These tasks incorporate elements of decision-making, apparent path planning, and the spontaneous emergence of roles. Furthermore, herding has practical applications in search-and-rescue scenarios faced by civil protection agents, who must guide groups to safety during emergencies such as earthquakes or fires. As such, herding is an ideal scenario for testing theories of human perceptual-motor behaviour.

Previous research on modelling human perception-action has typically focused on either very large-scale crowd dynamics [18, 19] or very small-scale individual or dyadic behaviours [20, 21]. Herding occupies a middle ground, involving an intermediate number of participants where individual actions significantly influence the collective behaviour of the group. Moreover, although past research on herding has often relied on simplifying assumptions such as flocking behaviour [22, 23, 24, 17, 25] and birds-eye view dynamics [26, 16], the work in this thesis distinguishes itself by removing these assumptions, focusing instead on the first-person perspective in herding a number of autonomous, non-interactive agents.

With regards to the practical applications of herding tasks to search-and-rescue scenarios, one may imagine scaling up the training of novice participants in simulation environments by replacing human coactors with artificial agents. In scenarios where access to expert human behaviour is limited, it becomes crucial to explore how these mathematical laws and computational tools can be harnessed to program such artificial agents. Such models can also find applications in developing robots capable of not only executing such actions [27, 28] but also physically cooperating with humans [29, 30], bridging the gap between expert knowledge and real-world applications.

1.2 Key research questions

Key research questions to be addressed in this thesis include:

- Can mathematical laws be used to capture human motion in shepherding tasks?
- Can simple heuristic rules or computational models be discovered to predict human decision-making in such tasks?
- What are the best combination of these laws/rules to model multi-agent shepherding behaviour?
- Can we generate human-autonomy teaming in this representative task context?

1.3 Contributions of this work

This work has advanced the modelling of human navigation in first-person shepherding tasks, proposing a new model to describe human motion in this context. Then, it combined this model with rules-based policies of human decision-making in single-herder contexts (i.e., the decision about which target to corral at a given point in time). It has furthermore explored computational tools, such as Deep Reinforcement Learning, to replace these simple rules-based approaches. Finally, it has integrated various combinations of these models to generate successful human-machine teaming in a real-time, multiagent human-machine hearing task.

1.4 Relevance to risk and complexity

Whether trying to design artificial agents that can cooperate with humans, or train them in simulation environments, in high-risk scenarios - such as military and defence training and operations, civil protection scenarios like fire and earthquake evacuation, the work presented in this thesis has direct applications to Risk and Complexity. Human behaviour is complex. Thus modelling human behaviour is modelling complex behaviour - human behaviour inherently is more than the sum of its parts, displaying many characteristics of complex, self-organising systems.

1.5 Thesis structure and outline

This thesis is divided into 8 chapters, with this being Chapter 1. Chapter 2 delves into the necessary mathematical frameworks and literature that provide the foundation for understanding and modelling the behaviour of herding agents (human or otherwise). It covers fundamental mathematical tools such as differential equations, control theory, and agent-based modelling. This chapter reviews past research, including classical herding strategies, agent-environment interactions, and the evolution of dynamical systems approaches to perceptual-motor tasks. Special attention is given to challenges such as multi-agent interactions and scalability, highlighting the gaps in existing models and outlining the goals of the present research in addressing these issues.

The third chapter details the methods used for data collection, analysis, and modelling. It begins by describing the experimental setup and the first-person herding game used to record human herding behaviour. Next, it covers the analysis techniques applied to understand the movement trajectories and decision-making patterns of human herders. The chapter also introduces the Deep Reinforcement Learning (DRL) algorithm used to develop the DRL artificial agents investigated in chapter 7. This methodological framework is crucial for understanding how both human and artificial agents were trained and tested in the herding task.

Chapter 4 marks the beginning of the modelling process, focusing on single-herder, single-target dynamics. Drawing on previous work in agent-based systems and dynamical models, this chapter modifies and fine-tunes the Fajen and Warren [5] navigational model

to capture how a single herder approaches and corrals a single target. It introduces an Ordinary Differential Equation (ODE) model tailored to the task and validates the model by comparing it with empirical human trajectory data. Through this validation, the chapter demonstrates the accuracy of the model in predicting human behaviour, setting the foundation for more complex herding scenarios.

Building on this single-target model, Chapter 5 incorporates multiple target agents to investigate the decision-making dynamics of human participants when tasked with corralling more than one agent. It describes how human participants navigate the complexities of multi-target scenarios, including how they prioritise targets and adapt their strategies dynamically. The chapter develops several heuristic models to explain these decision policies and evaluates their effectiveness in simulating human behaviour. One heuristic model emerged as the most fitting, providing insights into the decision rules that guide human herding behaviour in these more complex settings.

Chapter 6 increases the complexity of the task by introducing multiple human herder agents and additional target agents. It explores how two human herders coordinate their movements to efficiently corral targets while avoiding overlap and conflict in their strategies. The chapter models the cooperative dynamics between herders, drawing on both the movement models developed earlier and decision-making frameworks from multi-agent literature. The results demonstrate how the developed models are able replicate the coordination patterns that emerge among herders.

The final chapter (Chapter 7) shifts the focus to human-machine teaming. It describes the development and evaluation of autonomous herding agents that mimic human behaviour using the ODE motion model and decision-making strategies derived from either heuristic rules or DRL. The chapter compares the performance of human-machine teams to that of human-human teams, analysing the strengths and weaknesses of each computational architecture. The results reveal how closely autonomous agents can approximate human strategies in real-time human-machine teaming, and where their limitations lie. This chapter offers critical insights into the potential of artificial agents to collaborate with humans in complex task environments.

The concluding chapter summaries the key findings from the thesis, reiterating how the research contributes to understanding human herding behaviour and the development of artificial agents capable of similar tasks. It emphasises the success of DPMP-based models, heuristic decision-making, and DRL in replicating and predicting human performance in herding tasks. The chapter also discusses the broader implications for fields like robotics, AI, and human-machine interaction. Finally, it identifies several avenues for future work, including improving the scalability of the models for more complex multi-agent scenarios, enhancing the decision-making algorithms, and exploring real-world applications in dynamic environments.

1.6 Description of research periods

Year I, the coursework-based year, and Year II, were spent entirely in Naples, Italy, at the SSM. Year III and IV observed a division of research time across the SSM and Macquarie University, Australia, which housed the laboratories to conduct the experiments needed. Specifically, the periods

- 1 Nov 2022 - 1 June 2023 (7 months)
- 1 September 2023 - 25 May 2024 (9 months)
- 1 September 2024 - 31 October 2024 (2 months)

were spent at the Performance and Expertise Research Centre at Macquarie University, under the supervision of Prof. Michael Richardson, SSM board member.

2 Background

While modelling human behaviour in general presents a formidable challenge, significant progress has been made in using dynamical systems and computational tools to model, predict, and reproduce human behaviour in specific perceptual-motor tasks, such as simple limb movements [31] and navigation [5, 8]. Shepherding tasks [17], which involve the coordination and control of agents to achieve a collective goal (herding target agents into a predefined containment zone) offer a particularly valuable use case for further investigating the utility of these modelling approaches for capturing and understanding complex, human and multi-person (-agent) behaviour.

Previous research on shepherding, however, has often relied on several simplifying assumptions to make the modelling process more manageable, such as assuming flocking [22, 23, 24, 17, 25] and providing a birds-eye-view of the herding space [26, 16]. By removing these assumptions, we aim to extend this work to more realistic scenarios. The long term goal of my research program is to develop computational models of human expert herding and navigational behaviour that can be employed to train human participants in virtual or real search-and-rescue tasks, working alongside cooperative autonomous agent teammates. The thesis work presented here is a step towards this long term goal, by advancing our understanding of human and artificial herding behaviours with practical implications for enhancing human-machine collaboration in complex, dynamic environments.

2.1 Mathematical formulation

The herding problem can be formulated in mathematical terms as follows. Let $\mathbf{x}_i(t) \in \mathbb{R}^2$ and $\mathbf{y}_j(t) \in \mathbb{R}^2$ represent the positions of the i-th target agent and j-th herder agent relative to the center of the containment zone (see Fig. 2.1). The control action u governing the dynamics of the herder is defined by

$$\ddot{\mathbf{y}}_j = u(t, \mathbf{x}_1, ..., \mathbf{x}_{\mathrm{N_{TA}}}, \mathbf{y}_1, ..., \mathbf{y}_{\mathrm{N_{HA}}}) \tag{2.1}$$

where N_{TA} and N_{HA} denote the number of target agents and herder agents, respectively. The objective is to determine u such that

$$\|\mathrm{x}_i\| \leq R_C, \quad \forall i, \forall t > t', \tag{2.2}$$

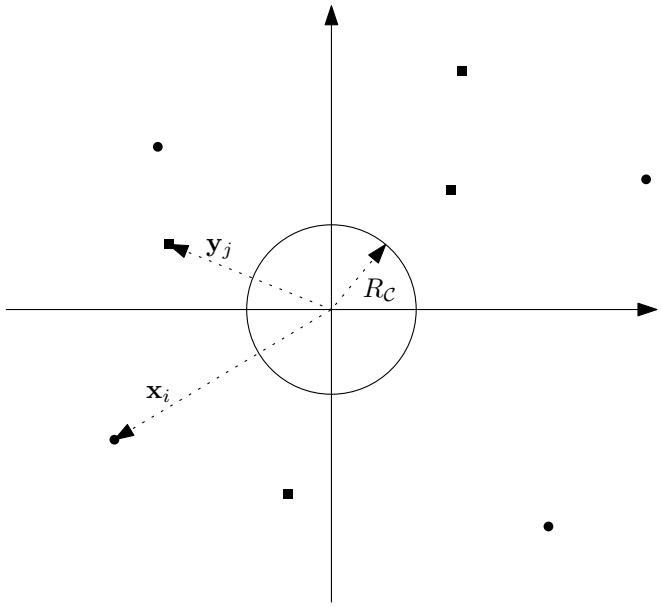


Figure 2.1: Basic formulation of the shepherding environment. The herder agents are depicted as squares and the target agents as circles. The large central circle is the containment zone.

where R_C is the radius of a specified containment zone (here taken to be circular), and for some arbitrary time t', ensuring that $\mathbf{y}_j(t)$ exhibits human-like behaviour for all j (the meaning of what "human-like" quantitatively means will be discussed later, starting from Chapter 4).

2.2 Dynamical Systems Approaches for Modeling Actions in Shepherding

Dynamical systems approaches have proven to be particularly effective in modelling the complex behaviours observed in shepherding tasks [32, 13, 14]. These approaches utilise low-dimensional, environmentally coupled ODEs or PDEs to capture the essential dynamics of both herders and targets. For example, research by [15] demonstrated that the navigational trajectories of human herders could be effectively modelled using simple mass-spring systems (see section 2.2.3 for more details), that incorporate attractive and repulsive forces to represent the influence of goals and obstacles. Similarly, studies by [33] employed similar models to control autonomous herding agents, validating these models through simulations and real-world experiments. The collective results of the previous work has demonstrated how these models, combining ordinary differential equations with rules-based decision policies, not only help in understanding the fundamental principles governing shepherding behaviours but also provide a robust framework for developing autonomous systems capable of human-like herding performance. The integration of these dynamical models within agent-based simulations [34, 16] has also allowed researchers to develop a more comprehensive understanding of human herding strategies, enabling the optimisation and refinement of control algorithms for both virtual and physical herding scenarios [33]. The thesis work presented here significantly extends this previous work, by developing a herding movement or navigational model that better generalises across a range of first-person, multiagent herding task scenarios. Before presenting the proposed model - that is able to navigate towards and corral moving targets - it is important to first understand (i) the dynamical functions or *dynamical perceptual-motor primitives* that form the foundation to dynamical models of human perceptual-motor behaviour and (ii) the DPMP navigational model proposed by Fajen and Warren [5, 6] that formed the foundation to the model investigate here.

2.2.1 Dynamical Perceptual-Motor Primitives

There is now a growing body of research demonstrating how nearly all human perceptual-motor behaviours can be modelled (approximated) using a small set of environmentally coupled non-linear dynamical functions or dynamical perceptual-motor primitives (DPMPs) [31, 35]; namely, environmentally coupled fixed-point (mass-spring) and limit-cycle (self-sustained oscillator) equations [36, 31, 2, 16, 37, 38, 39, 7]. In the case of modelling human perceptual-motor behaviour, these primitives relate to the two general classes of informationally coupled task directed human movements: discrete movement (e.g., moving towards goal, or reaching for or grasping an object, throwing a ball, tapping

a key), and rhythmic behaviour (e.g., walking, clapping, chewing, hammering a nail). Indeed, research has demonstrated how DPMP models can effectively capture everything from simple object reaching, collision avoidance [38], and postural control tasks [2], to more complex interpersonal object pick-place and pass tasks [40, 41, 42], competitive games such as air-hockey [43], and crowd behaviours [18]. DPMP models have also been used for the self-organised control of robots executing similar actions or tasks [27, 28], as well as for human-robot interaction [29].

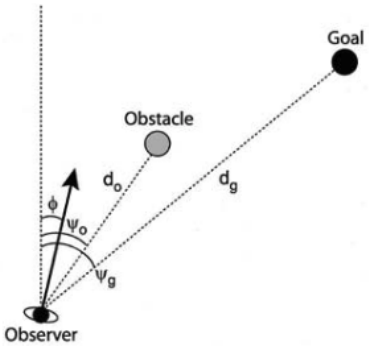
As noted above and of most relevance here, is that Fajen and Warren (e.g., [5, 6, 7, 8]) have remonstrated how a simple, fixed point, DPMP functions can be employed to capture and predict the route selection and navigational behaviour of humans across a range of different task environments.

2.2.2 Fajen and Warren's Navigational Model

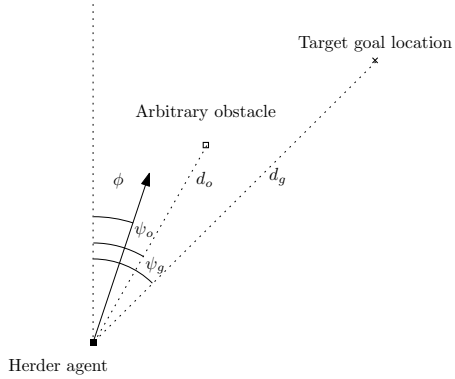
Consistent with the dynamical systems approach to understanding human behaviour, Fajen and Warren's navigational model [5] offers a comprehensive framework for understanding and predicting human path trajectories in complex environments. Their model emphasises the emergent nature of navigational behaviours, proposing that these behaviours arise from the dynamic interactions between an agent's heading direction and the attractive and repulsive influences of goals and obstacles. This approach aligns closely with the principles observed in herding tasks, where the movement of herders and targets can be understood as a result of similar dynamic interactions. By incorporating elements such as steering dynamics and obstacle avoidance, Fajen and Warren's model provides a robust theoretical basis for developing advanced herding strategies that can be applied to both human and artificial agents. Their work has significantly influenced subsequent research in the field, including studies that integrate these navigational principles with reinforcement learning ([37, 44, 45]) and other adaptive control techniques ([33]) to enhance the decision-making capabilities of herding agents.

To fully appreciate the contributions of Fajen and Warren's navigational model, it is essential to delve into its mathematical underpinnings. Their model is characterised by an ordinary differential equation (ODE) that describes how an agent's heading direction evolves over time in response to the spatial configuration of goals and obstacles. This equation incorporates terms for both attractive and repulsive forces, reflecting the agent's tendency to move towards goals while avoiding collisions. The mathematical formalism of this model provides a clear framework for simulating and analysing navigational behaviour, offering insights into how simple rules can give rise to complex and adaptive movement patterns. By reviewing the key terms and parameters of the Fajen and Warren model, we can better understand its applicability to herding scenarios and its potential for informing the design of autonomous herding agents.

Stated succinctly, the Fajen and Warren navigation model [5] captures the dynamical evolution in time of an agent's heading direction, ϕ, measured in angles with respect to an external frame of reference (see Fig. 2.2a), which is the direction in which the agent moves, and can be modelled by:



(a) Representation of the model variables in the Fajen and Warren model, reproduced from [5] with permission.



(b) Carry-over to the herding problem. The arbitrary obstacle could be any avoided game element.

Figure 2.2: Basic navigatory setup of the navigation task on the left and the shepherding task on the right. The dotted vertical line represents the external frame of reference and the arrow, the heading direction of the observer or the agent.

$$\ddot{\phi} = -b\dot{\phi} \underbrace{-k_g(\phi - \psi_g)(e^{-c_1 d_g} + c_2)}_{\text{attraction to goal location}} \underbrace{+k_o(\phi - \psi_o)(e^{-c_3|\phi-\psi_o|})(e^{-c_4 d_o})}_{\text{repulsion from obstacles}} \tag{2.3}$$

In this equation, the acceleration, $\ddot{\phi}$, is influenced by three components:

1. **Damping Term**: The term $-b\dot{\phi}$ represents a damping effect, which acts to slow down the agent's rotational acceleration over time, thereby stabilizing its heading direction. b is the damping coefficient.

2. **Attraction to Goal Location**: The term $-k_g(\phi - \psi_g)(e^{-c_1 d_g} + c_2)$ models the attractive torque towards a goal location. Here, ψ_g is the angular position of the goal relative to the external frame of reference, d_g is the distance to the goal, and k_g, c_1 and c_2 are positive constants that determine the strength and decay of this attraction.

3. **Repulsion from Obstacles**: The term $+k_o(\phi - \psi_o)(e^{-c_3|\phi-\psi_o|})(e^{-c_4 d_o})$ accounts for the repulsive torque from obstacles. In this term, ψ_o represents the angular position of an obstacle, d_o is the distance to the obstacle, and k_o, c_3 and c_4 are positive constants that define the intensity and decay of the repulsion effect.

Across a range of studies [5, 6, 7, 8, 9], Fajen and Warren and colleagues have demonstrate this model can effectively and successfully model human navigation in cluttered environments [5], interception of moving targets [6], and model collective

human motion in crowds [9]. The wide applicability of this model testifies to the unifying nature of the underlying dynamics of human motion, and implies that this same models could be adapted to more complex human perceptual-motor tasks. For instance, [41, 42] have also demonstrated how this same model can be used to capture the arm movement dynamics of pairs on individuals completing collaborative pick-and-place tasks.

Of particular interest here, was whether this model (2.3) could be adapted to model the navigational behaviour a human herders in a cooperative multi-herder, multi- target herding task, as well as its potential use in designing autonomous herding agents. As detailed in Chapter 4, the model can be adapted to effectively simulate the target approach and corralling movements of human herders, by simply treating target agents as goal locations and other environment objects as obstacles - other herders and non-targeted targets (Fig. 2.2b). Before detailing this model, the rest of this provides a short review of previously-used approaches to modelling human shepherding behaviour and the action decision or goal state activation functions required for the successful enactment of DPMP models in complex task scenarios.

2.2.3 Dynamical Systems applied to the tabletop shepherding task

The majority of previous research exploring and modelling human shepherding behaviour has employed tabletop versions of the task. These tabletop shepherding tasks, as described in [15, 16, 46], have served as an excellent laboratory task for studying human cooperation and coordination. The experimental setup has also provided an excellent platform for developing and testing artificial agents capable of successfully collaborating with human participants [44, 45]. In the typical experiment, pairs of participants stand on opposite sides of a large touchscreen displaying a virtual herding environment. the environment includes a containment zone, virtual sheep (targets), and two squares representing the herder agents (HA), which the human participants control controlled using motion tracking sensors or by touching and moving their HA with their index finger or a touch screen pen (see Fig. 2.3 for more details).

Research employing this task has demonstrated that, given enough practice, all pairs converge to the same two modes of perceptual-motor movements and coordination: (1) search-and-recover behaviour, where players/herders consistently move to and corral the sheep furthest from the containment zone on their side of the table, progressing from one distant sheep to the next until all are within the containment zone; and (2) coupled oscillatory containment, where the players oscillate around the containment zone, building a 'circular' wall to contain the herd. More importantly, both of these behaviours can be effectively modelled using a simple, environmentally coupled DPMP model [33, 15, 47, 48]. This model [15], takes the form:

$$\ddot{r}_i + b_r \dot{r}_i + \epsilon_r (r_i - r_{sf(t),i}) = 0 \tag{2.4a}$$

$$\ddot{\theta}_i + b_{\theta,i} \dot{\theta}_i + \epsilon_\theta (\theta_i - \theta_{sf(t),i}) = (\dot{\theta}_j - \dot{\theta}_i) \tag{2.4b}$$

where $\ddot{r}_i$ and $\ddot{\theta}_i$ represent the acceleration of the i-th herder's radial position and angle on the field, respectively, with regards to the center of the containment zone (see Fig.

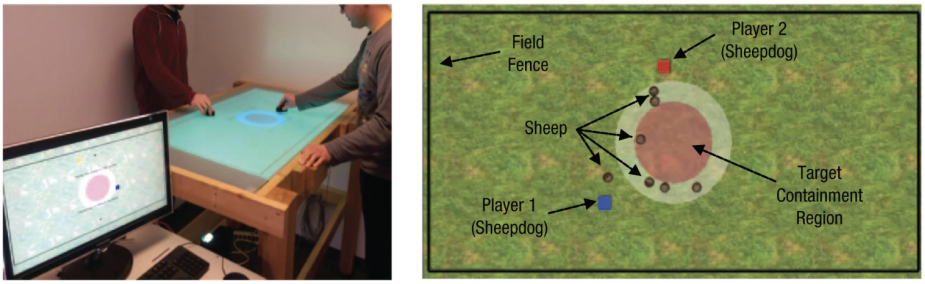


Figure 2.3: Experimental setup of the tabletop shepherding task. The left panel shows the physical layout of the experimental room, while the right panel displays a labeled diagram of the game environment, including all relevant game objects in the virtual experiment setup, such as the participant-controlled sheepdogs (red and blue cubes) and the sheep (brown dots), which must be herded into the red containment area. Reproduced from [15] with permission.

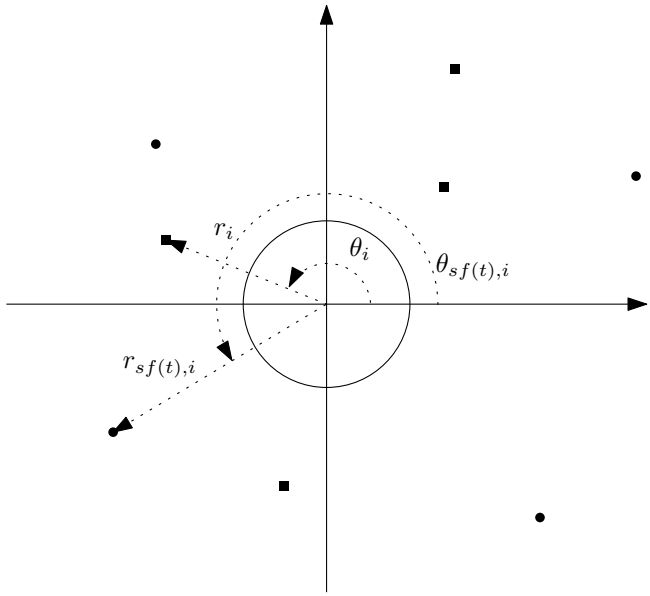


Figure 2.4: Setup of the variables in Eq. 2.4. r and θ are polar coordinates and the subscripts i and $sf(t), i$ refer to the i-th herder agent and its target position respectively. Refer to Fig. 2.1 for general setup.

2.4). The term $b_r \dot{r}_i$ denotes a damping coefficient for the radial velocity, while $\epsilon_r > 0$ is a stiffness coefficient, ensuring that the herder's position r_i converges towards a target position $r_{sf(t),i}$. Similarly, $b_{\theta,i}\dot{\theta}_i$ represents the damping for the angular velocity, and ϵ_θ is a stiffness coefficient for the orientation, with θ_i converging towards the target orientation $\theta_{sf(t),i}$. The right-hand side of equation (2.4b) accounts for the difference in angular velocities between the i-th and j-th herders, promoting coordination in their movements.

Not only were these equations able to reproduce human data effectively, but artificial HAs controlled by the resulting model were found to exhibit equivalent, human-like behavioural performance [16] and could impart the same level of skill training as human experts [46]. Indeed, when integrated into the control architecture of artificial herder agents, human participants completing the tabletop herding task with these artificial herder agents were unable to determine whether they were cooperating with an artificial agent or with another human herder [16].

Despite the success of Eq. (2.4b) in modelling human tabletop herding, the ability to generalise this model to first-person herding and navigation contexts remains limited. For one, individuals do not exhibit oscillatory containment in first-person herding, nor are the full-body movements produced at the same speed (velocity) as the hand movements used to control the position of HAs in tabletop herding tasks. Thus, while this research indicates the robust utility of the DPMP model for capturing human herding behaviour, a different DPMP model was required for the current work.

2.3 Modeling Decisions in Shepherding

Although much of the above research has provided clear evidence that DPMP models can effectively capture a wide range of human behaviours, including tabletop and potentially first-person shepherding behaviour, the effective realisation of these models in complex task environments depends on the goal or task activation functions that modulate how they unfold over time [16, 46]. That is, the utility of DPMP models relies not only on modelling movement patterns but also on modelling the action selection or decision-making behaviour of human actors completing a given task context. For herding, this entails modelling the target selection process individuals employ to choose which target to corral at any given point in time [34, 49].

Thus, while the current thesis demonstrates how (2.3) can be adapted to model the behavioural trajectories of human-controlled herders in a first-person task, this validation required uncovering an appropriate target selection (dynamic decision-making) process or policy. In line with previous research [34, 50, 48], the focus here was on developing a robust rule-based or heuristic policy that could not only effectively predict and mimic human target selection decisions but could also be easily integrated into the control architecture of DPMP-controlled herding agents [33]. However, given the recent success of reinforcement learning methods in developing optimal, task-specific action decision models [51, 52, 53, 54], this approach was also utilised in the final study of the thesis[37, 45, 44].

2.3.1 Heuristic policies

For the table-top herding task, numerous studies (e.g., [34, 48, 55]) have demonstrated how human players select targets using the following heuristic policy: at any point in time, pick the target that is (i) furthermost from and (ii) moving away from the containment zone and (iii) closer to me that my co-herder/player. Moreover, and as detailed above, that this imply heuristic policy can be integrated together with 2.4 for the development of artificial herders capable of human level [16, 48, 46].

Motivated by this latter shepherding work, [33] explored a range of different rule-based, heuristic target selection policies for herding, focusing on a dynamic, position-based field division to enhance decision-making and herding efficiency. This approach, termed Dynamic Peer-to-Peer target selection, enabled herder agents to adapt their control strategies in real-time, effectively corralling both small and large groups of non-flocking agents with Brownian motion dynamics. By dynamically partitioning the field, the model streamlined the herding process and allowed agents to efficiently gather and contain targets. This Dynamic Peer-to-Peer policy was compared against three alternative strategies—Global Search, Static Area Partitioning, and Dynamic Leader-Follower—with all on the examined policies highlighting the effectiveness of simple, yet responsive, position-based heuristic species for multiagent herding tasks (see [33] for strategy details).

A key distinction among the strategies explored in [33], was whether they were dependent on local vs. global information about TA states. For instance, the Global Search and Static Area Partitioning policies both required HAs to know the location of all TAs in the game field or in a pre-defined (static), pre-designated herding area, respectively. The dynamic target selection strategies (Dynamic Leader-Follower and Dynamic Peer-to-Peer), however, only required the HAs (players) to know about the targets within a certain radius of their current (immediate) location. Interestingly, [33]'s findings indicated that target selection strategies based on local, dynamically updated information yielded more efficient herding performance than global and static search area strategies. That is, strategies based on dynamic, local information resulting in significantly shorter HA path lengths, more expeditious target gathering, and a higher success rate. As will be seen in this thesis (Chapter 5), dynamically updated rules unsurprisingly performed better in our first-person herding task as well.

While rule-based approaches like these offer clear and intuitive models for decision-making in herding tasks, they can be limited by their predefined rules and lack of adaptability to new or unforeseen scenarios. Reinforcement Learning (RL), on the other hand, provides a flexible and powerful alternative for modelling decisions. RL algorithms can learn optimal strategies through trial and error, adapting to dynamic environments and complex tasks. This adaptability makes RL particularly suited for tasks where the optimal behaviour is not easily encoded by simple rules, allowing agents to autonomously discover and refine strategies that maximise their performance in diverse and evolving contexts.

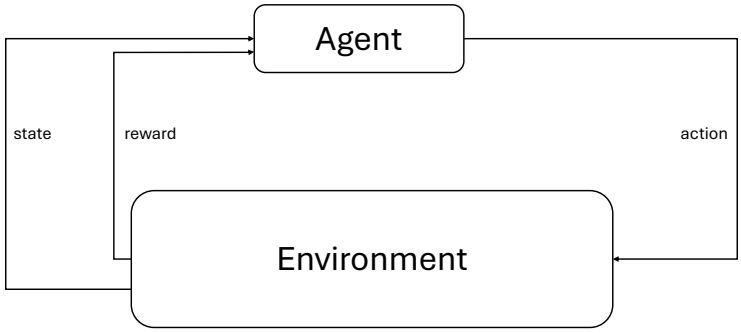


Figure 2.5: An illustration of the core structure of Reinforcement Learning, where an agent interacts with its environment through a continuous cycle of actions and feedback. The agent selects an action based on the current state of the environment, which responds by updating its state and providing a reward. This reward guides the agent's learning, encouraging actions that maximise long-term rewards, ultimately shaping its strategy for optimal decision-making in the environment.

2.3.2 Reinforcement Learning policies

Reinforcement learning (RL, [56, 57]) emerged as a powerful tool for modelling decision-making processes in shepherding tasks [37, 45, 44]. This approach enables agents to learn optimal strategies through trial and error, leveraging feedback from the environment to improve performance over time.

RL is a branch of machine learning where an agent learns to make decisions by performing actions in an environment to achieve a goal. This learning process is characterised by three main components: states, actions, and rewards (see Fig. 2.5). The state represents the current situation or context the agent is in, which can include information about the environment, the agent's position, and other relevant factors. The agent takes actions based on the current state, in order to transition to new states. After each action, the agent receives a reward, a numerical value that indicates the immediate benefit of the action taken. The ultimate objective of the agent is to maximise the cumulative reward over time, often referred to as the return.

In RL, the agent's learning process involves exploring the environment (trying new actions to see their effects) and exploiting its current knowledge (choosing actions that are known to yield high rewards). To effectively learn the optimal strategy, or policy, the agent must balance exploration and exploitation. Traditional RL methods, such as Q-learning [58] and SARSA, use tabular approaches where a value function or action-value

function is updated based on the rewards received. However, these methods struggle with large or continuous state-action spaces due to the curse of dimensionality.

Deep reinforcement learning (DRL) addresses these limitations by employing neural networks to approximate the value function or policy. In DRL, the agent uses deep neural networks to map states to actions or to predict the value of state-action pairs, enabling it to handle high-dimensional inputs like images or complex environmental states. One of the most notable algorithms in this domain is Proximal Policy Optimisation (PPO) [59]. PPO improves learning stability and efficiency by optimising a surrogate objective function that ensures the new policy does not deviate excessively from the old policy. This is achieved through a clipping mechanism that restricts the policy update, making PPO robust and widely applicable in various complex RL tasks.

Multi-agent reinforcement learning (MARL, [60, 54]) extends the principles of reinforcement learning to environments where multiple agents interact, each with their own goals and strategies. In MARL, agents must not only learn to optimise their individual rewards but also to coordinate and sometimes compete with other agents in the environment. This adds layers of complexity, as agents must account for the actions and potential strategies of others, leading to a rich set of dynamics and emergent behaviours. MARL has enabled applications in various domains, including robotics [61], autonomous driving [62], and complex strategy games [52].

There are numerous compelling studies on the use of RL for modelling decision-making in joint-action tasks, such as those found in [63, 64] (the game "Overcooked") and [37, 44, 45] (the tabletop herding task presented above in 2.2.3). In [64], the authors examined a "human-aware" artificial agent - an AA that was trained with a copy of an agent resembling a human agent. In RL, an agent must be trained over several (usually millions) iterations of the simulation environment. In MARL, a few options arise: train the AA with a copy of itself in a method referred to as Self-Play, or with a copy of a human-model-based AA (referred to as human-aware or human-sensitive). In [64], it was found that human agents teamed with the human-aware agent performed better in the task (the game "Overcooked") than humans teamed with another human.

The feasibility of human-(RL)AA teaming was observed in [37, 44, 45] (the tabletop herding task), with human-RL(AA) teams exhibiting comparable performance to human-human teams (trial time, TA travel, and other metrics). However, [46] demonstrated that human participants still preferred playing with human-sensitive, heuristic, and human-award DRL agents, compared to self-play MARL agents. Furthermore, these human-sensitive agents were able to impart game expertise to novice human participants at a level comparable to that of expert human co-players.

This important distinction between human-sensitive and self-play-trained MARL agents will reappear later in this thesis, as the aim will be to develop AAs that mimic human behaviour. As will also be detailed later in this these, however, the integration of RL with DPMP control strategies offers a promising avenue for enhancing the flexibility and robustness of herding algorithms, by making them more capable of handling real-world variability and uncertainty.

2.4 Discussion

This chapter has highlighted the progress made in modelling human perceptual-motor behaviours, particularly within shepherding contexts, through the use of dynamical systems and computational approaches. By examining the unique characteristics of shepherding tasks, such as coordinated movement and the control of agents towards a collective goal, we have seen the effectiveness of applying DPMP models. These models, combined with rules-based decision policies, have proven adept at capturing key human behaviours within structured environments, such as tabletop tasks, thereby advancing our understanding of human and multi-agent dynamics. However, as detailed, certain limitations remain when applying these models to more immersive, first-person herding tasks, where full-body movements and dynamic decision-making processes differ significantly from simpler, tabletop interactions.

This review has also underscored the need to develop adaptive, context-specific decision-making policies that can emulate human target selection behaviours. While traditional heuristic policies can model certain decision patterns effectively, the introduction of reinforcement learning (RL) has allowed for more refined, task-specific action models that can better account for the complexity of human decision-making in varied task environments. Incorporating RL-based policies offers a promising avenue for capturing human behaviours more accurately in dynamic, first-person settings, while also enhancing the robustness of artificial agents in tasks requiring close human-machine collaboration.

In the subsequent chapters, I will present my work on developing a novel experimental environment that addresses these limitations. I will begin by detailing the methods used in designing the experimental platform, data collection, and analytical techniques. This setup will serve as a foundation for the systematic exploration and validation of both traditional and RL-based decision-making models within complex, first-person herding tasks.

3 Methods

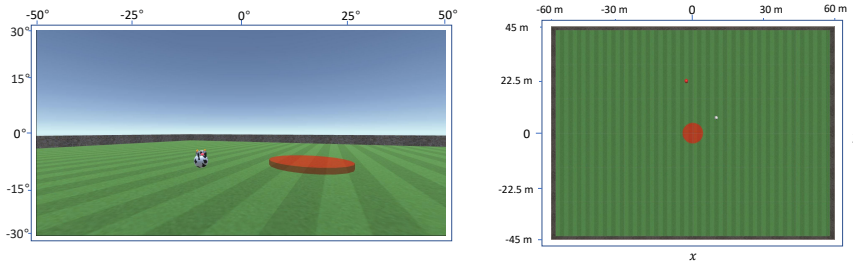
The thesis includes 4 experiments (chapters 4, 5, 6 and 7). Each experiment involved an assessment of human, human-human team or human-AA team herding behaviour playing a first-person herding game. In each experiment data was collection on herder and target agent positions, velocities, and movement trajectories, and analysed to understand the sequential phases of human herding behaviour, including the action (target) selection decision making of human herders. The experimental setup provided insights into how humans adapt their navigation strategies in response to dynamic agent interactions, contributing to the development of models for human-machine teaming in herding tasks. Given that the herding task and behavioural assessment measures were similar across all experiments, I outline these methods and techniques in this this chapter.

3.1 Experimental Paradigm

3.1.1 Apparatus and Task Environment

The herding task required human participants to control an HA, from a first-person point of view, in order to find and corral one or multiple TAs into a specified containment zone (circular red area) positioned in the centre of a large game field. Example views of the herding task environment are displayed in Figure 3.1, including a screenshot of the first person point of view of the participant (Figure 3.1a), with single target agent (virtual cow) and red containment zone in view, as well as a bird's eye view of the entire task environment (Figure 3.1b). The game was designed using the Unity game engine (Version 3.3.0, Unity Technologies, San Francisco, USA) and was presented to participants on a 27-inch computer monitor (1920x1080p).

The game area corresponded to a 120m $\times$ 90 m field, fenced off on each side with walls that prevented the HA and the TAs from leaving the game field. TAs were represented as spherical cows, each with a horned cubical head and a black and white textured body (see Figure 3.1a). The walls exerted a repulsive force of 4N on the TAs (of unit mass) upon collision, thus preventing the TAs from constantly moving along the walls of the game field. The HA could influence the TA by coming within a radius d_i = 10m of the TA, with the TA's movement dynamics defined as:



(a) First-person point of view.

(b) Corresponding bird's eye view.

Figure 3.1: Task environment. (a) Example of a participant/herder agent's first-person view of the game field. The TA (spherical cow with head) and sections of the containment zone (red) and the exterior walls of the field (grey) are also visible. (b) A birds-eye view of the game field (not seen by the participant), where the HA is the small red dot and the TA is the small white dot.

$$\ddot{\mathbf{x}}(t) = \alpha_r \frac{\mathbf{x}(t) - \mathbf{y}(t)}{\|\mathbf{x}(t) - \mathbf{y}(t)\|^2} - \beta \dot{\mathbf{x}}(t). \tag{3.1}$$

Here, $\mathbf{x}(t) \in \mathbb{R}^2$ and $\mathbf{y}(t) \in \mathbb{R}^2$ are the positions of the TA and HA with respect to the centre of the containment zone, respectively, and the dots above the variables refer to the standard time derivatives. β is the drag coefficient set to 0.2s^{-1} and α_r is a constant set to $20\text{m}^2\text{s}^{-2}$ when the HA is within the influence area of the TA and 0 otherwise (values set heuristically to maintain similarity across congruent environments, see [65, 66]). This results in the TA moving in the direction directly away from the HA approach vector when the HA enters the influence area. Note that the maximum repulsion the TA could experience was clamped at 30N. While it had been assumed and implemented in this study that the TAs would move deterministically in response to the HA's movements as in Eq. (3.1), other works [65, 66] consider noisy TA dynamics - and evasive TAs are the subject of current studies by the authors.

As noted above, the containment zone was specified by a red translucent, non-rigid-body circular area with a radius of 4m and was always located at the centre of the game field. Both the TA and the HA could freely pass through (in and out) of the containment zone without being impeded by the containment zone.

Participants controlled the direction of motion and angular orientation of the HA using keyboard and mouse controls, respectively similar to a classical first-person shooter game (e.g., Wolfenstein 3D). Using the W, A, S, and D keys, the HA could move forwards, left, backwards, and right, respectively. Upon pressing two keys - WD, WA, SD, and SA - simultaneously, the HA would move diagonally with respect to the forward direction. The HA's speed was fixed at 5 m/s. The HA's orientation could be rotated to the right or to the left by moving the mouse in a corresponding transverse (left-right) direction. The HA's field of view spanned 60° in the vertical direction and 97.6° in the horizontal direction. Note that the HA's head movement (camera orientation) was decoupled from

its body movement (translation) allowing participants to visually explore the environment without moving around.

3.1.2 Task Description

Participants were required to complete N trials, with each experiment containing a different number of trials (ranging from 20 to 24). For each experiment, the first quarter ($N/4$) trials were considered practice and were not included in any analysis. Each initial condition included different (x, y) game field positions of the HAs and TAs (never located within the containment zone), as well as a different initial HA heading angle. For each trial the HA and TA game field positions, heading angles (defined within regards to the centre of the game field) and movement velocities were recorded at 50 Hz. Trial order was randomised across participants in order to avoid extraneous order effects. Importantly, the first quarter of trials (i.e., practice trials) were randomised separately from the analysis trials. Note that including the same set of randomised experimental trials (and initial conditions) enabled an analysis of different participant and AA behaviour across identical initial conditions.

That is, to address concerns related to third-variable confounding and ensure the validity of the findings, a randomised trial order was employed to minimise any potential extraneous variable effects on participants' behaviour. The experimental design also included practice trials, which provided participants with an opportunity to familiarise themselves with the setup and minimise learning effects on the main data collection phase. Additionally, recruitment was conducted from a relatively homogeneous pool (students from the same age group), This context limits generalisability to other populations (e.g., older adults), which should be considered when interpreting the results.

A game trial started with the HA and TA positioned according to the predetermined initial conditions and came to an end when the TA was corralled within the containment zone and its speed was less than .1 m/s. This required the HA to stop influencing the TA prior to the TA reaching the containment zone, such that the TA slowed down and came to rest within the containment zone. That is, if the HA continued to influence the TA up until it entered the containment zone or after it entered the containment zone, the TA would simply pass through the containment zone without the trial ending.

3.1.3 General procedure

After arriving at the laboratory, participants were seated at a desk with the computer monitor, keyboard, and mouse used for the study positioned directly in front of them (refer to Figure 3.2). Participants then read and signed a consent form and completed a demographic survey. Participants were then informed that the study was investigating and modelling human navigation and herding behaviour and that they would be completing a simple herding task that required them to control a virtual HA to locate and corral one or more TAs into the red containment zone. Participants were not instructed about how to best corral and herd the TA(s), but were instructed to complete the trials in the shortest time possible. After participants indicated they understood the task, they completed all trials in a single session. Given that standard power analysis was not

Figure 3.2: The experiment setup. The person in the photo is the author of this Thesis.

appropriate for this type of modelling research—where standard hypothesis testing was not the primary focus—the participant sample size was chosen based on prior studies examining the dynamics of human herding behaviour (e.g., [34, 16, 46, 67]). The approach was designed to ensure the reliability of the herding behaviour patterns while allowing generalisable insights from a moderate number of participants.

3.2 Single herder - single target experiment

3.2.1 Participants

Twenty-four participants from Macquarie University, Sydney, Australia, were recruited for the study. The participants were between 18 and 36 years of age (M = 20.75, SD = 3.64). Twenty participants self-identified as female and four as male. Twenty-two participants were right-handed, one left-handed, and one identified as ambidextrous. The participants completed all 20 trials in a single session lasting 20 to 25 minutes.

3.2.2 Data Analysis

Pre-processing

In the single herder - single target experiment, participants were required to complete a total of $N = 20$ trials. As noted above, the first five trials were considered practice trials and were excluded from further analyses. Six highly erroneous human trajectories were also manually discarded from analysis, in which the human participant either (i) continuously overshot the TA or containment zone and had to circle back around (and around) the field to re-herd the TA, or (ii) was the only participant to circle around the field in a direction opposite to the rest of participants. These six trials (refer to Fig. 4.3 for an illustration of the data and the outliers) represented only 1.67% of the total number of experimental trials recorded.

Recall that the aim of this study was to model the navigational trajectories of the participants as they moved toward, approached, and then incorporated the TA into the containment area. It is important to note that at the start of a trial, if the TA(s) were not within their the HA's initial field of view, participants typically rotated their HA's head or body around to perform an initial visual scan for the location of the TA(s). However, given the above focus of the current work and the trivial nature of this scanning behaviour, an analysis or model of this behaviour was not considered here.

It is also important to note that at the end of the trial, some participants retreated away from the containment zone after successfully incorporating the TA(s) into the containment zone. To exclude this nonherding behaviour from model parametrisation, the end of the HA's TA-directed navigational and corral movements was defined as the last moment in time the HA influenced the last TA herded (or the TA in the single target experiment).

Train-test split

Out of the 15 experimental trials, the first twelve were chosen as the training set on which to parameterise our movement model and the last three were chosen as test trials on which to evaluate the simulated HA against. Given that trials were presented to the participants in a randomised order, the training and test split was also randomised. Furthermore, out of the three test trials, it was observed that in one trial approximately half (N=10) participants took a route tending towards the top half of the game field, and the other group (N=14) took a route tending towards the bottom. This was because the HA's initial position, centre of the containment zone and TA's initial position were almost perfectly aligned (angle = 179.5°). This further corroborates the previous observation that the initial heading angle of the human HA does not play a role in the route selection, as all 24 human HAs had the same initial heading direction. This particular trial was thus split into two sub-trials corresponding to the top and bottom groups respectively, refer to the bottom panel of Figure 4.4 for illustrations.

3.2.3 Mean human trajectories and confidence bounds

Once the post-navigational transients were removed from the participant HA trajectories, time normalised mean human HA trajectories were generated for each of the 20 initial conditions. These were obtained by first resampling the participants' HA (x, y) trajectories to 1000 points for each trial and then calculating the mean (x, y) value at each time index for the corresponding initial condition.

To assess the variation in human/participant trajectories within the model test set, a 90% confidence interval around the average human HA trajectory was calculated. The interval, symmetrically positioned, was determined by multiplying 1.645 times the standard deviation of point-to-point distances between individual participant trajectories and the mean human trajectory for each initial condition.

Trajectory measures and comparison analysis

This model, complete with all degrees of freedom defined, was thus able to generate simulated trajectories. To compare simulations with human data, the errors for individual humans in the trials corresponding to the test dataset were calculated as the Dynamic Time Warped (DTW) [68, 69] distances from the mean human trajectory, per trial. A brief overview of the DTW metric, as well as its suitability for our use case is covered in Appendix A. These errors were ranked and the 90-th percentile error was compared with the error of the simulation with respect to the mean human trajectory. The DTW distance was employed as it allows one to assess the difference between two time-series trajectories independently of differences in movements speeds or sampling frequency. This was crucial here as human participants could actually start and stop their movement while the simulated trajectories were made a constant speed.

Statistical tests

For the single-target study, as well as throughout the thesis, standard statistical tests such as t-tests and Bayesian Factors analyses were used. These tests identify the evidence for the alternate or the null hypothesis: in our case, the dissimilarity or similarity of the simulations from the human data, respectively. Classical independent samples t-tests, as well as Bayesian Null Hypothesis Significance Testing had been extensively used for this purpose.

3.3 Single herder - multiple targets experiment

In this multi-target herding experiment, the game setup mirrored that of the single-target experiment, except that there were three targets instead of one. Participants were also required to complete five single-target practice trials before beginning the multi-target trials to familiarise themselves with the game controls and mechanics. These initial single-target trials were excluded from further analysis.

3.3.1 Participants and setup

Twenty-one participants from Macquarie University were recruited for the study. The participants were between 18 and 33 years of age (M = 21.05, SD = 3.54). Sixteen participants identified as female, four as male, and one as nonbinary or third gender. Twenty participants were right-handed and one left-handed. The participants completed all 24 trials in a single session lasting 25 to 30 minutes. None of the data was discarded, and no pre-processing was performed. The aim of this experiment was to infer the target selection strategy adopted by human actors in this task context.

Once the target selection strategy was inferred from this first dataset, a new set of ten participants recruited from the Scuola Superiore Meridionale, Italy, were tasked with playing the same video game, the only difference being that a different set of initial conditions was used, with the number of trials unchanged. This second data set was used to validate the results of the first analysis. Ten participants were recruited, aged between 25 and 32 years (M = 27.80, SD = 2.32). Three participants identified as female and seven as male. The 10 participants were right-handed.

3.3.2 Data Analysis

The same data pre-processing and trajectory analysis methods employed in 3.2 were employed for this second experiment.

Target Selection Policy Evaluation

As noted previously, in addition to modelling the movement trajectories of participants the thesis also explored the target section polices adopted by participants in multi-target settings (i.e., the order or sequence in which the human herders selected and corralled the targets). I predominately focused on what heuristic policies best capture human multi-target herding behaviour (although also DRL policies will be considered in Chapter 7; see below). This entailed assessing the predictive accuracy of different sets of computational rules about what target a player would engage with next based on different position and distance criteria. To test the predictive accuracy of given policy, the following steps were performed. First, the actual order in which the TAs were corralled by a participant was extracted from trial data. The target order or sequence exhibited by a participant for a given trial (set of initial conditions) was then compared to the ordering predicted by a given policies. If (and only if) the predicted target selection order perfectly matched the observed ordering, a policy was deemed to have successfully predicted the observed order. For a given participant, the predictive accuracy of a given policy was examined across all trials, such that the predictive accuracy of the policy was the percentage of correct TA sequence predictions over trials (i.e., 0 = zero sequence predictions and 1 = perfect sequence predictions for all trials).

Given that different polices might predict in the same target selection sequence for a certain initial condition, percent accuracy scores were calculated in two ways. First, a non-mutually exclusive score was calculated, which corresponded to the number of successful trial predictions, independently of whether other policies also predicted the

same sequence. This non-mutually exclusive score was then used to rank order the overall accuracy of policies. Given this rank ordering of polices (from highest or lowest), the second method involved determine the mutually exclusion accuracy score. This involved determining the first policy that predicted the observed order for a given trial and participant from the non-mutually-exclusive rank order (highest to lowest accuracy). This policy was then deemed predictive, received a score of 1, and the mutually exclusive accuracy of all other policies for that trial for that participant was assigned a score of 0. Averaged across trials, this resulted in a predictive accuracy score from 0 to 1, where higher scores corresponded to a policy that better predicted participants' behaviour compared to other policies, including those that might yield the same predicted sequence for a given set of initial conditions.

Validity of simulated HA behaviour

To further validate the inferred target selection policy, simulations were made by coupling this policy to the single-target movement model developed in the single-herder single-target study (see Chapter 4). The simulations were then evaluated against the human data by first constructing a non-linear (specifically, square root) heat map of the spatial occurrences of trajectory points in square 5-by-5 m bins. The resulting heat map was filtered through for bins with a value (obtained through trial-and-error) greater than 10, to eliminate bins with only single trajectories passing through a bin, to obtain a binary map.

The percentage of any given trajectory falling through this binary map was calculated as the number of bins the trajectory passed through in the binary map for a given set of initial conditions, normalised by the total number of bins assigned to that given trajectory. This value was called the "binary trace" of a trajectory. The "weighted trace" of a trajectory was defined as the similarly normalised sum of the values of the nonlinear heat-map bins through which that trajectory had passed. The traces of each human participant were calculated and the mean and SD was taken across all trials. A detailed representation of human trajectories, the simulation, and the weighted and binary heatmaps is present later in Chapter 5, Fig. 5.2.

3.4 Multiple human herders - multiple targets experiment

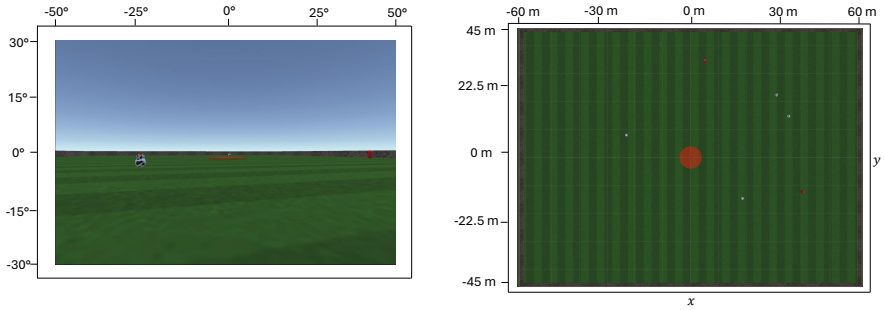
3.4.1 Participants

Forty-two participants (22 pairs) were recruited from the undergraduate student cohort at Macquarie University, Australia and were divided into pairs. Each pair was assigned to complete an online shepherding task, where they worked collaboratively to shepherd multiple (three to five) target agents into a central containment zone. The participants were between 18 and 30 years of age (M = 20.14, SD = 2.13), with 36 who self-identified as female and six as male. Twenty-seven participants reported that they played video

games, of which five played on a daily and seven on a weekly basis. All study procedures were approved by the Macquarie University Human Research Ethics Committee.

3.4.2 Apparatus and task

An illustration of the task is presented in Figure 3.3. The setup was the same as in the previous tasks, except that there were two herders in this game environment and 3, 4 or 5 target agents.



(a) First-person point of view.

(b) Corresponding bird's eye view.

Figure 3.3: Task environment. (a) Example of a participant/herder agent's first-person view of the game field. The TAs (spherical cow with head), the containment zone (red) and the exterior walls of the field (grey) are visible. The other HA is also visible in the distance as a red capsule. (b) A birds-eye view of the game field (not seen by the participant), where the HAs are the small red dots and the TA is the small white dot.

3.4.3 Trial Sequence and Initial Conditions

Each participant first completed five practice trials as a single herder corralling a single TA (referred to as the single-herder single-target task). Following these trials, participants then completed 24 trials of the multi-agent herding task in pairs, as two HAs with three, four of five TAs (i.e., 6 trails for each target number condition). Among the multi-agent multi-target trials, the initial six were also considered as practice trials (two trials per target number condition). As with the rest of the experiments in this Thesis, the practise and the analysis trials were randomised in order, separately, across participants.

3.4.4 Procedure

Participants in a pair completed the study in the same laboratory room, but were seated at separate computer desks, each desk with a computer monitor, keyboard, and a mouse positioned directly in front of the participant. The desks were positioned so that the participants in a pair were facing in opposite directions and could not see each other. Participants were also instructed not to talk during the experiment, with the experimenter in the room to enforce this instruction.

Upon arriving at the experimental laboratory room, participants first read and signed an informed consent form, and then completed a short a demographic survey, which captured their age, gender identification, and how often they played video games. Following this, the experimenter informed the participants they would be playing a herding (video-game) task, and that they would complete two versions of the task: (1) a single-herder-single-target and (2) a multi-player-multi-target version. Participants were instructed that they would be controlling a virtual HA to locate and corral TAs into a containment zone in the centre of the game field, and were instructed to complete the task in the shortest time possible. Once participants indicated that they understood the task and the keyboard and mouse controls, they then completed the 5 trials of the simple single-herder-single-target task independently (lasting approx. 5-10 minutes). Following a 5 minutes break, participants completed the 24 trials of the multi-player-multi-target task, with this lasting approximately 30 to 40 minutes.

It is worth noting that participants were not informed that the single-herder-single target task and the initial six trials of the multi-agent-multi-target task were practice trials. In addition, for the multi-player-multi-target task, participants were told that they would be playing as a team and to work together to corral the targets, but were not given any instructions about how best to do this. After participants pairs completed all 24 of the multi-player-multi-target trials, they were informed that the study aimed to model their herding behaviour and, following any questions they may have had about how and why this modelling was being done, they were thanked for their participation.

3.4.5 Data Analysis

The same data pre-processing and trajectory analysis (binary trace), methods employed in sections 3.2 and 3.3 were employed for this third experiment. Given the increased complexity of the task, (i.e., 2 herders and 3, 4 and 5 targets), the identification and evaluation of different possible target selection policies, including the predictive power of the resulting movement model + policy was conducted in a slightly different manner compared to 3.3. These evaluation and validation methods are detailed in Chapter 6, as well as the statistical methods of analysis employed.

3.5 Multi-agent herders - multiple targets experiment

The experimental setup for this experiment mirrored that of the multiple human herders study in section 3.4, with two significant differences. First, rather than involving two human participants, the experiment substituted one of the humans with an artificial agent (AA) whose behaviour was controlled in real-time by either the best navigation + policy model identified in the previous human-human, multi-herder, multi-target experiment, or by a DRL-based target selection policy. The latter was, therefore, the second significant difference. That is, the effectiveness of the navigation + heuristic policy was not only compared to human-human performance but also bench-marked against target selection policies developed using cutting-edge DRL methods. This included two DRL agents: (1) a self-play (SP) DRL agent, which learned a target selection policy by playing alongside copies of itself; and (2) a "human-aware" DRL AA (referred to as HS-DRL), which learned a target selection policy by playing (in simulation) alongside the navigation + heuristic policy. Details about how these DRL policies (agents) were trained and developed are provided in Chapter 7.

3.5.1 Participants

Seventy-one Macquarie University students participated in the study in exchange for course credits required for an undergraduate psychology course. Nine were excluded for not completing the study (six due to motion sickness, two due to difficulty with task mechanics, and one due to technological issues). A total of 62 participants completed the task, each randomly assigned to one of three AA conditions: Heuristic (n = 22), HS-DRL (n = 20), or SP-DRL (n = 20). Participants' ages ranged from 17–32 (M = 19.03, SD = 2.58), with 52 females and 10 males. Informed consent was obtained at the start of sessions, and all participants were naive to their assigned AA condition. Study procedures were approved by the Macquarie University Human Research Ethics Committee.

3.5.2 Procedure

The same procedure and trials employed in 3.4 were employed here, with the exception that participants were informed that they would be playing the game with an AA. No information about the specifics of the AA were provided to participants. In order to balance out the assignment of initial conditions to the two-player team, half of the experimental sessions had the human player start off with the set of initial conditions that were given to the AA in the other half fo trials (i.e., half of the participants played as player 1 and the other half played as player 2).

3.5.3 Data Analysis

Similar data pre-processing, trajectory analysis (binary trace), and policy evaluation methods employed in 3.4 were employed for this forth experiment. Details about how these methods were adapted, where necessary, are detailed in Chapter 7, as well as the statistical methods of analysis employed.

3.6 Discussion

In this section, I have reviewed the general methods, procedures, and data analysis techniques employed across the four experiments presented in Chapters 4, 5, 6, and 7. In the next chapter, Chapter 4, I will focus on the single-herder, single-target study, describing the prototypical movement dynamics exhibited by participants when engaging in the first-person herding task examined here. Importantly, the next chapter will also detail the DPMP herding navigational model that is foundational to the work presented in Chapters 5, 6, and 7.

4 Single herder, single target movement dynamics

This chapter focuses on how an individual herder agent can effectively corral a single target agent. The primary goals of this study included (i) identifying the movement or navigational trajectories that emerge during the herding process and (ii) to determine whether these behavioural patterns are consistent across individuals (participants). This involved analysing how the human-controlled herder agents first approached and then corralled a target toward the containment zone. By closely examining the herding trajectories exhibited by participants, the third specific aim of this study was to identify how (and whether) the Fajen and Warren [5] DPMP navigation model could be adapted to capture human first-person herding behaviour.

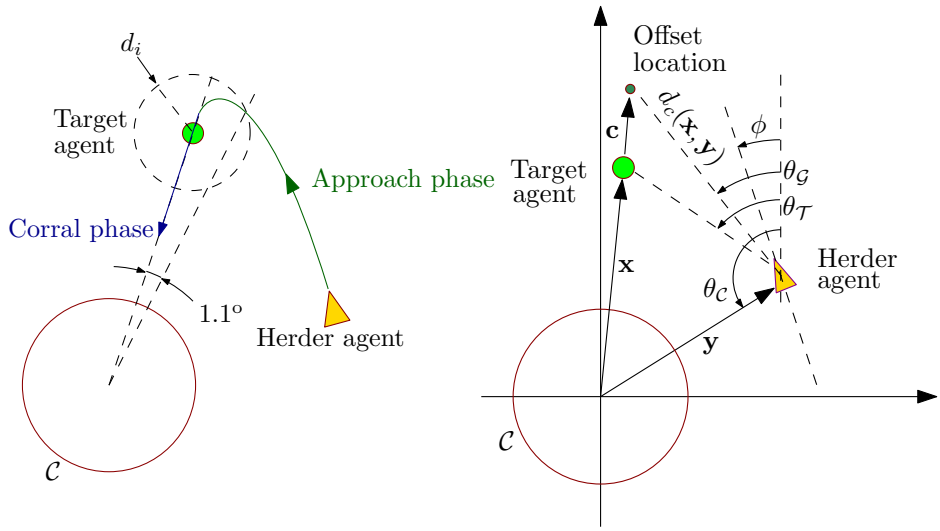
4.1 Modelling Movement Dynamics

4.1.1 Phase Identification

To determine whether the navigational model proposed by Fajen and Warren [5] could be adapted to capture the behavioural dynamics of human herders completing a first-person herding task, the movement and target approach strategies of 24 participants who were required to corral a single target into the containment zone were examined (see section 3.2 for data processing details). Participants completed 20 trials in total (5 practice and 15 experimental).

An analysis of participant data revealed that (human) HA trajectories consisted of two distinct phases: (i) an approach phase and (ii) a corral phase (see Figure 4.1a). During the approach phase, participants moved the HA to a position behind the TA with respect to the containment zone. That is, all participants navigated toward an offset location that was (on average) a distance c away and behind the TA with respect to the containment zone. For this first phase of behaviour, participants also typically maintained a distance greater than the HA-TA repulsion distance to ensure that the TA was not pushed further away from the containment zone.

For the second, corral phase, participants moved directly towards the TA entering the TA influence region in order to repel (drive) the TA in a (more or less) straight line



(a) (b)

Figure 4.1: (a) Shows the phases of the herding task. The HA first approached the TA and then switched to corralling it towards the containment zone C. The median angle at which human HAs first influenced TA was 1.1°. d_i is the influence radius of the TA, equal to 10 m. (b) Shows the model variables: C denotes the containment zone, $\mathbf{x}$ the position of the TA, $\mathbf{y}$ that of the HA. $\mathbf{c}$ is the offset from the TA, $d_c(\mathbf{x},\mathbf{y})$ the distance between the offset goal location and the HA. ϕ is the HA heading angle and $\theta_{\mathcal{G},\mathcal{T},\mathcal{C}}$ the angles of the offset goal location, TA and C, respectively (measured from the vertical axis). See the text for more details

towards the containment zone. Importantly, the angle at which the HA began influencing the TA could be used to differentiate the two phases. Furthermore, the median of these angles across all participants, measured as the angle between the positional vectors of the HA and the TA originating from the centre of the containment zone, corresponded to 1.1° (Figure 4.1a), indicating that the participants transitioned to corralling just before being perfectly aligned (on a straight line) with the TA and the centre of the containment zone. Finally, participants' motion never (except for 1 out of 360 trajectories) passed through the containment zone (even though they could and were never given instructions either way).

4.1.2 Herding navigational model - general formulation

To model the navigational trajectories of the participants during the task, the approach and corral phases were modelled using a modified version of the navigation model proposed in [5]. Specifically, assuming constant forward motion with $v(t) = v_0$ being the HA speed, and $\dot{\mathbf{y}}$ the velocity of the HA (refer also to Fig. 2.1),

$$\|\dot{\mathbf{y}}\| = v(t) = v(0) = v_0, \tag{4.1}$$

and an exo-centric reference frame, the heading direction ϕ of an HA was modelled (see Figure 4.1b for setup of variables) as

$$\ddot{\phi} = -b\dot{\phi} + \psi_{\mathcal{G}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{G}}) + \psi_{\mathcal{T}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{T}}) + \psi_C(\mathbf{x}_C, \mathbf{y}, \phi, \theta_C) \tag{4.2}$$

where $\theta_{\mathcal{G},\mathcal{T},C}$ are the angles of the offset goal location, TA and C, respectively (measured from the vertical axis). As usual, $\mathbf{x}$ denotes the position of the TA and $\mathbf{y}$, the position of the HA. $b \in \mathbb{R}^+$ is a positive damping constant and $\psi_{\mathcal{G}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{G}})$ represents the attractive coupling between the forward heading direction of the HA and the angular position of the target offset location. Again, this offset target location corresponded to the location behind the TA and is the goal location that the HA is attracted to. Moving towards this offset location, rather than the location of the TA itself, ensured that the HA does not influence the TA during the approach phase.

To further ensure that the HA does not influence the TA during the approach phase, the term $\psi_{\mathcal{T}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{T}})$ in Eq. (4.2) provides a repulsive coupling between the HA heading direction and the TA's angular position with respect to that heading. This repulsive torque disappears smoothly as the HA moves around and behind the TA (with respect to the containment zone) to allow the HA to approach and drive the TA into the containment zone.

Finally, the term $\psi_C(\mathbf{x}_C, \mathbf{y}, \phi, \theta_C)$ in Eq. (4.2) creates a repulsive coupling between the direction of the HA and the centre of the containment zone C, fixed at $\mathbf{x}_C = \mathbf{0}$. This was implemented to ensure that the HA avoids getting too close or entering that containment zone, similar to what was observed for the human participants. The full mathematical expressions of each of these terms is presented in the following section.

4.1.3 Herding navigational model - expression details

The term $\psi_{\mathcal{G}}$ represents the attractive coupling between the HA heading direction and the angular position of the offset location $\mathbf{x}_c = \mathbf{x} + \mathbf{c}$, where $\mathbf{c} := c\frac{\mathbf{x}}{\|\mathbf{x}\|}$. It was defined as

$$\psi_{\mathcal{G}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{G}}) := -k_g(\phi - \theta_{\mathcal{G}})(e^{-c_1 d_c(\mathbf{x},\mathbf{y})} + c_2), \tag{4.3}$$

where $\theta_{\mathcal{G}}$ is the angle of the offset location with respect to the vertical axis centred on the HA, k_g reflects the strength at which the HA's heading direction is attracted towards the off-set location, c_1 and c_2 are positive constants, and $d_c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} + \mathbf{c} - \mathbf{y}\| \in \mathbb{R}^+$ is the distance between the offset location and the HA. Note that $(e^{-c_1 d_c(\mathbf{x},\mathbf{y})} + c_2)$ ensures that the attractive coupling to the offset target location decays as a function of $d_c(\mathbf{x}, \mathbf{y})$ but does not completely vanish.

The term $\psi_{\mathcal{T}}$ which repels the HA from the TA's current location, was defined as

$$\psi_{\mathcal{T}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{T}}) := \zeta(\mathbf{x}, \mathbf{y}) k_o (\phi - \theta_{\mathcal{T}}) e^{-c_3 |\phi - \theta_{\mathcal{T}}|} e^{-c_4 \|\mathbf{x} - \mathbf{y}\|}, \tag{4.4}$$

where $\theta_{\mathcal{T}}$ is the angle between the TA and vertical axis centred on the HA, k_o represents the strength at which the HA's heading direction is repelled away from the location of the TA, and c_3 and c_4 are other positive constants. The exponential modulation terms ensure that the repulsion decays rapidly as the HA faces away from the TA and as the two agents get further away. The function $\zeta(\mathbf{x}, \mathbf{y})$ in Eq. (4.4) is a sigmoidal function chosen as

$$\zeta(\mathbf{x}, \mathbf{y}) = 1 - \sigma\left(\epsilon - \cos^{-1}\left(\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right)\right), \tag{4.5}$$

with $\sigma(z) = \frac{1}{1+e^{-z}}$, ensuring that the negative coupling term in Eq. (4.4) decays towards zero as the HA approaches the TA. The parameter ϵ was set to 5^o (as indicated from the human trajectory data, at 1.1^o the HA first influences the TA) measured from the line joining the centre of the containment zone and the TA (more simply the angle between $\mathbf{x}$ and $\mathbf{y}$ - see Figures 4.1a, 4.1b and 4.2). ζ thus defines when the transition between the approach phase and the corral phase occurs.

Finally, the term ψ_C, which repels the HA away from the containment zone C (the centre of which, $\mathbf{x}_C$ is stationary at the origin), was defined as

$$\psi_C(\mathbf{x}_C, \mathbf{y}, \phi, \theta_C) := k_o (\phi - \theta_C) e^{-c_5 |\phi - \theta_C|} e^{-c_6 \|\mathbf{x}_C - \mathbf{y}\|} \tag{4.6}$$

where θ_C is the angle between the centre of the containment zone and the vertical axis centred at the HA, and c_5 and c_6 are positive constants. The term ψ_C repels the HA away from the containment zone at all times.

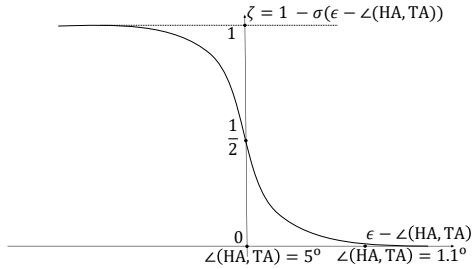


Figure 4.2: Justification of choice of $\epsilon = 5^o$. $\angle(\mathrm{HA}, \mathrm{TA})$ is the angle between the HA and the TA, subtended at the centre of the containment zone, equal to $\cos^{-1}\left(\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right)$. This brings the influence of the target-as-a-repeller term towards 0 as the simulated HA approaches $1.1°$. $\epsilon - \angle(\mathrm{HA}, \mathrm{TA})$ is plotted on the x-axis and the function ζ on the y-axis. $\zeta = 0$ when $\angle(\mathrm{HA}, \mathrm{TA}) = 5°$.

4.1.4 Model parametrisation and simulations

To parametrise the model on mean human data, a parameter search algorithm was employed that found the model parameters that minimised the Dynamic Time Warped (DTW, [68, 69]) distance between the simulated and mean human data. Sequential Least Squares Quadratic Programming (SLSQP) ([70, 71], see also Appendix A), was employed to find this minimum. This algorithm was used because it is one of the methods that consistently finds the optimum for similar problems in a reasonable number of steps [72]. In particular, the `scipy.optimize.minimize(method='SLSQP')` Python function was used. For each trial in the model training set, the heading angle input to the simulation was taken as the tangent of the mean human trajectory near the HA's initial position[1]. The initial positions of the HA and TA for the simulations corresponded to the respective inputs to the human HA trials.

The parameters c_1, c_2, k_g, k_o, c_5, and c_6 were heuristically chosen as the parameters to parametrise, with the other parameters held constant (with their values taken from [5]). Ten different uniformly randomly chosen initial values for c_1 and c_2 were chosen in the range (0.1, 0.9) each, with identically chosen exploration limits. The median of the parameters output by each of these ten runs was taken, and then the mean, median, and SD thereof across trials. Using these latter median values, the parametrisation algorithm was run for the parameters k_g and k_o, which were chosen with initial values in the range (35, 50) and (150, 220), respectively, with exploration bounds (25, 60) and (150, 250). This procedure was identically reproduced for c_5 and c_6 with ranges and bounds (0.1, 1) each.

Once the optimum parameters for each initial condition were chosen, the overall median values per parameter were then employed for model simulations. The following parameters values were employed: $c_1 = 0.1\text{m}^{-1}, c_2 = 0.4, k_g = 43.4, k_o = 176.4, c_5 = 0.64$, and $c_6 = 0.15\text{m}^{-1}$. $c_3 = 1.25$ and $c_4 = 0.05\text{m}^{-1}$ were, as mentioned, taken from [5] and were not needed to be re-parametrised, as after the current parametrisation the model already resulted in excellent fits. Note that the offset location's distance from the TA, c, was set to 8.5m (since the repulsion distance was 10m) and the damping constant b to 3.5 s^{-1} in the simulations, as these values resulted in good fits.

Using the latter parameters, equation (4.2) was then simulated for each trials (initial condition), at an integration step of 0.02s, with an HA speed of 5m/s (same as for participant controlled HAs). For each trial, the average initial position and heading direction of the human participant across trials and the initial TA position were used as the initial conditions for the simulation . The resulting simulation trajectories were pre-processed by discarding their last 20%, since the simulated HA remained in place while the human HAs withdrew from the TA at the end. The simulations were presented and compared against human data in the *Model Validation* section below. Independent samples and Bayesian Factors t-tests were performed on the trajectory measures across human and simulation data, for the model validation set.

[1] tangent between the 200th and 50th point, as that is approximately when the human HA starts moving

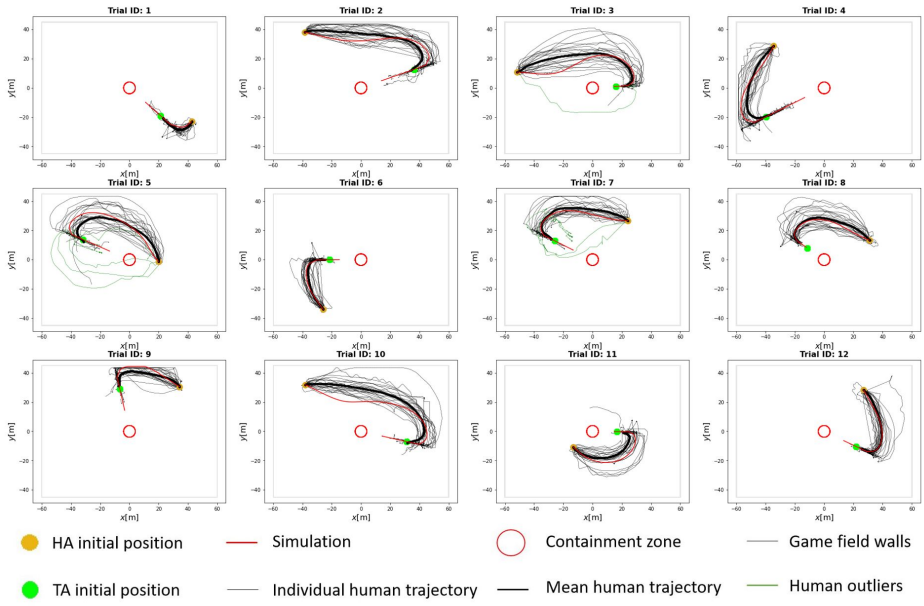


Figure 4.3: Model training set. Six dropped trajectory outliers are included in the figure, but were excluded from the model training. Legend has been used subsequently for all trajectory plots

4.1.5 Model Validation

In order to determine whether Eq. (4.2) could effectively capture observed human data (i.e. produce simulated HA movement trajectories equivalent to participant controlled HA movement trajectories), the model was first parameterised using 12 of the 15 experimental trials and then validated (tested) against the remaining three trials. As detailed previously, model parametrisation (optimisation) was conducted by minimising the normalised (dynamic time warped or DTW; [69, 68]) distance between simulated trajectories and the mean human trajectory for each of the 12 parametrisation trials (i.e., the corresponding 12 different initial conditions). From the trial optimised parameters, the median parameter settings were then calculated and used for model validation.

As can be seen in Figures 4.3 and 4.4, these median parameter values produced simulated trajectories that were equivalent to the average human trajectory for each of the 12 parametrisation trials and the three test trials, respectively, with regards to the metrics defined in section 3.2. The trajectories were also representative of the general distribution of the trajectories observed by the participants for each trial. To further validate the similarity of the simulated and mean participant trajectories for the three test trials, four different measures of trajectory similarity were employed (see Table 4.1). The first two were *navigation time*, calculated as the time of the approach and corral phase, and *path length*, calculated as the length of navigation during the approach and corral

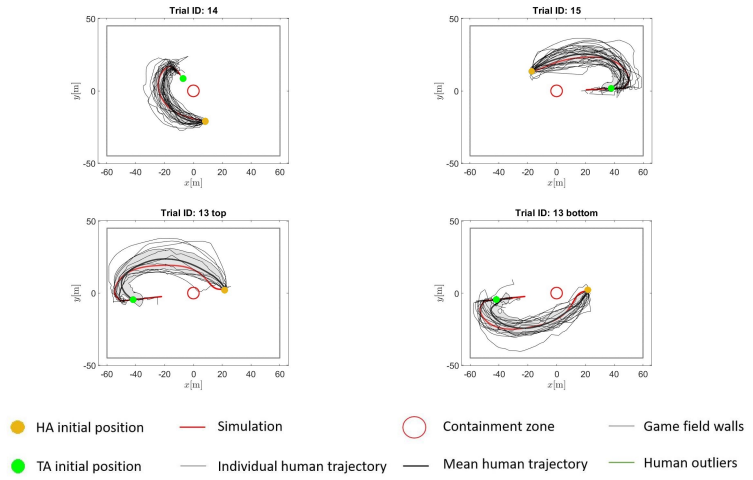


Figure 4.4: Model vs. data for the test set. Trial ID 13 was the one split into two sub-trials, as participants were divided according whether they took the top or bottom route. The shaded area in grey is the SD bound around the mean human trajectory.

phase, without observed differences between the simulated and mean human trajectories (both $t(6) > 0.25, p > 0.15$), with Bayesian factors with null effect Bayesian Factors of $BF_{10} = .534$ and $BF_{10} = 1.045$ for the path length and navigation times, respectively).

The third measure, *coverage percentage*, captured the percentage of a given trajectory that was within the 90% confidence interval (see Methods section 3.2.3 for more details) around the mean human trajectory. Consistent with the simulated trajectory being equivalent to a normative human trajectory, the simulated trajectories had a consistently higher coverage percentage than the 90th percentile human trajectory for all initial conditions in the test set (the 90th-percentile human trajectory corresponds to the 90th-percentile furthest human trajectory from the mean human trajectory as per the DTW distance). The fourth measure was the distance from DTW (or *error*) of the simulated trajectories from the mean human trajectories. Again, this error was less for the simulated trajectories compared to the 90 percentile human trajectory for each test trial, further indicating how the model was able to generate navigational trajectories similar and largely indistinguishable from those exhibited by the participants.

Table 4.1: Trajectory characteristics of human participants and simulations for the four set of initial conditions unseen by the model during parametrisation. Comp. human denotes comparative human, which is the 90th percentile furthest human trajectory from the mean human trajectory as per the DTW metric. Sim denotes simulation. The Mean and SD columns for the first two characteristics are for human data. The last row provides the mean for all trials.

	Path Length [m]			Navigation Time [s]			Coverage percentage [%]		Error from mean [m]	
Trial ID	Mean	SD	Sim	Mean	SD	Sim	Comp. human	Sim	Comp. human	Sim
14	72.1	16	70.2	17.8	6.2	18.8	29.8	97	5.829	1.866
15	101.1	17.7	108.9	24.8	4.8	37.7	23.2	83.3	6.184	5.255
13 top	119.2	22.7	126.6	28.4	4.9	42.7	26.4	69.9	8.257	7.138
13 bottom	121.4	20	126.1	29.5	5.3	42.5	47.3	70.7	5.571	5.112
Averages	103.5	19.1	108	25.1	5.3	35.4	31.7	80.2	6.46	4.843

4.2 Discussion

In this Chapter, the Single-Herder Single-Target Herding experiment was analysed. An analysis of participants' HA movement trajectories revealed that trajectories consisted of two distinct phases: (i) an approach phase, in which participants moved the HA to an offset position behind the TA, with respect to the containment zone; and (ii) a corral phase, where, after reaching the target-offset location, participants moved directly towards the TA, entering its influence region, to guide the TA in a straight line towards the containment zone. More importantly, computer simulations revealed that these behaviours could be captured using an adapted version of the Fajen and Warren [5, 6, 8] navigation model, in which the HA's heading direction was simply attracted towards a 8.5 metre offset location positioned behind a to-be-corralled target agent, while treating the to-be-corralled target agent as an obstacle. The transition between the approach and corral phases was simply achieved by incorporating a specific function that smoothly reduced the repulsive force between the target and the HA's heading direction as the HA moved around and behind the TA. In the following experiments, this model will be integrated with other models in progressively more complex herding tasks. In particular, the next chapter will present experiments that had been conducted in order to discern the decision policy(s) that individual herders adopt in containing multiple target agents.

5 Single herder, multiple target decision dynamics

In the previous chapter, a DPMP model was developed to capture the navigational movement dynamics of human controlled HAs tasked with corralling with a single TAs. In this chapter, we will explore the behavioural dynamics of herding multiple targets and identify the target selection policy (or policies) that human actors appeared to employ when selecting which TA to corral, choosing from multiple possible TAs.

5.1 Modelling decision dynamics

5.1.1 Definition of policies

Recall that for this experiment, single participants were required to corral three targets agents into the containment zone to complete a given trial (see 3.3. To identify the actual target selection strategy adopted by human participants when completing a trial, the observed TA selection order of a participant for a given trial was compared to the order that was predicted by various heuristic policies. The heuristic policies examined here were of four different general types, with each type including several sub-types based on a distance variable used to distinguish the relative location of targets with respect to the HA and/or the containment zone. As detailed below, these distances were calculated as (a) the linear, Euclidean distance, between a target and a HA or the target and the containment zone, or (b) the angular distance between the target and the HA. Note that the angular distance between a target and an HA is the angle between their position vectors, originating from the centre of the containment zone.

1) *Initial condition policies* - where the entire target selection sequence was defined by the initial state of the environment prior to herder movement or target engagement. The following six sub-policies were examined: (i) closest or (ii) furthest linear (Euclidean) distance from the HA; (iii) closest or (iv) furthest angular distance from HA; and (v) closest or (vi) furthest linear distance from containment zone.

2) *Successive policies* - where the first TA that was selected was defined by the initial state of the environment, but then subsequent target selections are made successively at the time of target engagement (but before target is corralled into the containment zone).

Again, six sub-policies were examined: (i) successive closest or (ii) successive furthest linear (Euclidean) distance from the HA; (iii) successive closest or (iv) successive furthest angular distance from HA; and (v) successive closest or (vi) successive furthest linear distance from containment zone.

3) *Dynamic policies* - similar to successive policies, but subsequent target selections are made dynamically; i.e., distance evaluations are made when a target agent has been corralled into the containment zone. The corresponding six sub-policies were examined. (i) dynamic closest or (ii) dynamic furthest linear (Euclidean) distance from the HA; (iii) dynamic closest or (iv) dynamic furthest angular distance from HA; and (v) dynamic closest or (vi) dynamic furthest linear distance from containment zone.

4) *Collinear policies* - the above policies that also take into account the vicinity (in terms of the angular distance) between TAs. Specifically, whether two or more TAs were essentially collinear with regards to containment zone and should therefore be consider a TA sub-group. As detailed below, based on analysis of the human data this grouping vicinity corresponded to an angular distance of less than 18.9°. In short, when the angular distance between TAs was less than this value, the TA that is further from the containment zone was chosen to be corralled first, otherwise the respective base (initial, successive, or dynamic) policy was followed according to the respective angular or linear distance variable employed. This resulted in an additional 18 policies, referred to as collinear policies; e.g., Initial Collinear Angle/Distance; Successive Collinear Angle/Distance; etc.

5.1.2 Policy testing

To test the accuracy of the predictions of the 36 defined policies, the actual order in which the TAs were corralled was extracted from the recorded data, and the non-mutually exclusive and mutually exclusive prediction scores were calculated for each trial (see section 3.3.2). Recall that the non-mutually exclusive scores capture how many times a policy predicts the observed data, ignoring the fact that multiple policies might predict the same sequence for a given initial condition. In contrast, the mutually exclusive scores capture how well a given policy predicts the observed behavior to the exclusion of other policies that perform worse overall. Thus, the mutually exclusive score serves as the best indicator of which policy best captures the data overall.

5.1.3 Generalisation of movement model to multiple targets

The navigational herding model, Eq. (4.2) from the Single-Target Herding Experiment was modified as follows to account for multiple TAs:

$$\ddot{\phi} = -b\dot{\phi} + \psi_{\mathcal{G}} + \psi_{\mathcal{T}} + \sum_{i=1}^{N} \psi_{O_i} \tag{5.1}$$

where $\psi_{\mathcal{G}}$ and $\psi_{\mathcal{T}}$ take the same form as in Eq. (4.2) but apply only to the currently targeted TA, and $\sum_{i=1}^{N} \psi_{O_i}$ is the sum of repulsive terms each taking the form of ψ_C from Eq. (4.2). More specifically, each term ψ_{O_i} shares the same form as ψ_C and $\sum_{i=1}^{N} \psi_{O_i}$

now encompasses not only the containment zone as a repulsive obstacle, but also the non-targeted TAs as repulsive obstacles. For example, for a non-targeted TA, O_j at position $\mathbf{x}_j$ and angle θ_{O_j} from the vertical,

$$\begin{aligned}\psi_{O_j} &= \psi_{O_j}(\mathbf{x}_j, \mathbf{y}, \phi, \theta_{O_j}) \\ &= k_o(\phi - \theta_{O_j})e^{-c_5|\phi-\theta_{O_j}|}e^{-c_6\|\mathbf{x}_j-\mathbf{y}\|},\end{aligned} \tag{5.2}$$

Note that the number of obstacles, N, in eq. (5.1) includes the containment zone and the non targeted TAs. As the containment zone and non-targeted TAs (who total $N_{TA} - 1$, with N_{TA} the total number of TAs) are each modelled by this repulsive form,

$$N = \underbrace{1}_{\text{for the containment zone}} + \underbrace{(N_{\text{TAs}} - 1)}_{\text{non-targeted TAs}} \tag{5.3}$$

The parameters that appear in each of the ψ_{O_i} terms were identical to the parameters for the ψ_C term in the Single-Target Herding Experiment in Chapter 4. Similarly, all other parameters in Eq. (5.1) were the same as in Eq. (4.2). The chosen target selection policy was evaluated once every .25s to determine the currently targeted TA.

5.1.4 Highest-ranked policies

As can be send from an inspection of Table 5.1, the TA selection decisions of participants were best capture by the collinear policy defined in terms of the closest angular distance of TAs form the herder; this policy was termed the *successive collinear angle* or SCA policy. This policy entailed three heuristic rules. First, participants preferred to select the TAs that were closest to them in angular rather than linear distance. Recall that the angular distance between a given TA and HA is measured with respect to the centre of the containment zone (which was at the origin), whereas the linear distance is the Euclidean distance on the plane. Second, participants selected successive TAs based on which TA was closest in angular distance to the previous TA selected. This approach was contrasted with the selection of TAs on the basis of their angular distance from the HA's initial (starting) position. Finally, if two TAs were collinear or almost collinear with respect to the containment zone (within a certain angular distance from each other, approximately 18.9°, with respect to the centre of the containment zone), the participants preferred to select the TA that was further away from the centre of the containment zone, even if that TA was further away from the HA in terms of angular distance.

The SCA policy was able to predict the TA selection order exhibited by the participants in 78.8% of the trials. Interestingly, the collinear aspect of this policy meant that participants minimised game effort, since when two (or potentially three) TAs were close to collinear with the containment zone, it was more efficient (given the constraints of the task) for participants to influence the TAs that were furthest from the containment zone first, and then influence the second furthest TA enroute. That is, participants could corral both near-collinear TAs simultaneously instead of corralling one all the way to the containment zone before repeating this step for the following TA(s).

To validate that SCA captured the majority of the participant's TA selection behaviour, other TS policies defined were also tested, with their prediction scores also reported in Table 5.1. Again, these policies tested a variety of possible rules and combinations of rules with regard to (1) distances from the HA or from the containment zone, (2) linear vs. angular distances from the HA, measured from the initial HA position, as well as successively and dynamically (in real time), and (3) identification of whether the TAs were nearly collinear with the containment zone or not. A two-phase process - non-mutually exclusive, then mutually exclusive process - was used to determine whether SCA better predicted the sequence in which participants corralled TAs compared to other possible policies. In the first phase, the number of times a policy correctly predicted a participant's target order sequence was recorded in a non-exclusive manner. This was done for the 21 participants for the 18 initial conditions, totalling 378 trials across the entire sample. The precision of each policy was then calculated as a proportion of this total. Given that the policies were not always mutually exclusive (i.e. several policies would correctly predict the same sequence depending on the specific herder-target environmental conditions), for the second phase the target order sequence a participant exhibited on a given trial was then reclassified in a mutually exclusive, stepwise manner, consistent with the policy order rank. To do this, the target order sequence for a given trial was checked to see if it was consistent with the most accurate policy (determined from the non-mutually exclusive policy analysis). If so, it was assigned to that policy, and no further classification checks were performed. If not, it was checked against the next policy in the rank until such a consistent policy was found. A target order sequence was classified as 'other' when no policy was consistent with that sequence.

This analysis revealed that after SCA, only two other policies, the furthest from the containment zone and Successive Closest Angle from Herder, resulted in mutually exclusive accuracy scores greater than 3.0%, specifically 6.34% and 3.96%, respectively. Regarding non-mutually exclusive accuracy, the next-best policy was Successive Collinear Distance, which entails the same three rules as SCA, but is defined in terms of linear (planar) distance. Another policy, Initial Collinear Angle, differs from SCA in that the TA order is based on the angular distance of TAs from the HA's initial position. Finally, the fourth-best policy, the Successive Closest Angle from Herder, entails the first two rules of SCA, but not the third collinear rule.

5.1.5 Policy validation

Given the similarity in the top four non-mutually exclusive policies, a subsequent policy validation experiment was conducted, in which participants completed a specific set of trials that were designed to better assess whether participants' TA selections were most aligned with SCA. Figure 5.1 presents examples from the set of validation initial conditions that were specifically designed to test whether the TA selections were better predicted by (i) linear versus angular distance, (ii) successive versus initial angular distance and (iii) collinear clustering or its absence. A full presentation of initial conditions used for the validation is available in Appendix B.

As expected, SCA was confirmed as the policy that predicted most of the human TA selection orders, with a prediction accuracy over the entire set of validation trials

Policy	Excl. N	Excl. P	non-Excl. N	non-Excl. P
Successive Collinear Angle	298	0.788360	298	0.788360
Furthest From Containment Zone	24	0.063492	71	0.187831
Successive Closest Angle From Herder	15	0.039683	242	0.640212
Other	10	0.026455	0	0.000000
Successive Furthest Distance From Herder	10	0.026455	33	0.087302
Dynamic Collinear Angle	8	0.021164	165	0.436508
Furthest Distance From Herder	6	0.015873	0	0.000000
Closest Distance From Herder	4	0.010582	128	0.338624
Closest From Containment Zone	2	0.005291	30	0.079365
Dynamic Collinear Distance	1	0.002646	156	0.412698
Successive Closest Distance From Herder	0	0.000000	182	0.481481
Successive Furthest Angle From Herder	0	0.000000	25	0.066138
Dynamic Furthest Distance From Herder	0	0.000000	0	0.000000
Dynamic Closest Angle From Herder	0	0.000000	162	0.428571
Dynamic Furthest Angle From Herder	0	0.000000	0	0.000000
Furthest Angle From Herder	0	0.000000	0	0
Closest Angle From Herder	0	0.000000	0	0
Initial Collinear Angle	0	0.000000	252	0.666667
Initial Collinear Distance	0	0.000000	224	0.592593
Successive Collinear Distance	0	0.000000	278	0.735450
Dynamic Closest Distance From Herder	0	0.000000	147	0.388889

Table 5.1: Total count and proportions of policy evaluations across mutually exclusive and nonmutually exclusive categories. The total number of TA selection policies for participants is the number of participants × number of trials = 21 × 18 = 378. Excl. N, is the total number of trials of mutually exclusive policies tested, Excl. P, is the respective proportion. Non-Excl. refers to the same but for the non-mutually-exclusive. The top three policies have been highlighted in bold.

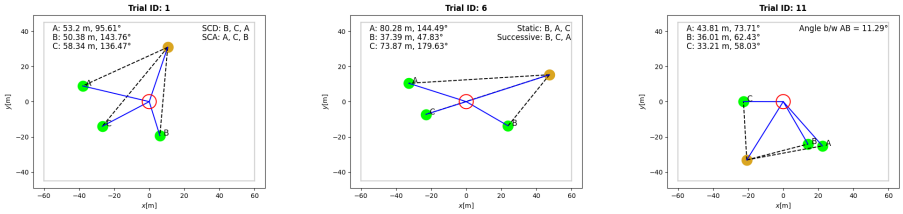
of 71.7%. With regard to subsets of validation trials, SCA more accurately predicted participant behaviour compared to all other alternative and competing policies tested. In the initial conditions testing the angular versus linear distance, the SCA had a score of 78%, while the policy based on successive collinear distance predicted 22% of the trials (only non-mutually exclusive scores are reported in this subsection). Where successive TA selection was tested against initial condition TA selection, SCA scored 78%, while initial condition TA selection scored 12%. Finally, in trials that tested "collinear cluster identification", SCA (using a collinear cutoff of 18.9°) obtained a score of 35%, while the successive closest angle predicted only 1% of the trials.

It is important to note that the considerably lower SCA score in this case compared to its overall score of more than 70% was due to the variation in the participant threshold for cluster/group identification. In the validation trials where the TAs clearly did not form a group (the minimum angle between them, subtended in the centre of the containment zone, was greater than 30°), the participants selected the TA closest in angle 84% of the time as their choice of the first TA and then preferred the SCA policy when choosing the second TA to corral 93% of the time. Furthermore, for the tests in which the TAs potentially formed a group (angle between two or more TAs less than 30° from the containment zone), the participants chose the TA farthest from the containment zone within that cluster 71% of the time. These percentages vary slightly after changing this 30° parameter (for example, after setting it to 25°, this last score increases to 74%), indicating that participants have varying thresholds to classify TAs as part of a group. To further illustrate this, the accuracy of the SCA was calculated for each participant in the validation dataset by varying the threshold angle for cluster identification. The thresholds for most participants to treat TA as a group ranged between 15° and 25°. These results revealed that while there was this (i.e., between 15° and 25°) variability in cluster/group identification threshold across participants, the empirically derived value of 18.9° on average fell within this bound.

5.1.6 Simulations

As a final validation of the SCA policy and Eq. (4.2), artificial HA simulations were performed, where the architecture of the artificial HA included Eq. (4.2) and the SCA target selection policy. Simulations were carried out using the same trial conditions used for the Multi-Target Herding Experiment. Note once again that Eq. (4.2) was slightly modified so that non-targeted TAs (the TAs that are not currently chased) were treated as standard obstacles, as per Eq. (5.1).

As expected, our Dynamic Perceptual-Motor Primitive (DPMP) model Eq. (4.2) combined with the SCA target selection policy generated simulated HA trajectories that were consistent with the observed human data. The similarity between the simulated and participant data was determined using weighted and binary trace maps, examples of which are provided in Figure 5.2. Recall that the weighted and binary trace maps (defined earlier in section 3.3.2) represent the areas on the movement plane most frequented by participants, with the white area in the binary trace maps approximating the confidence interval. As can be seen from an inspection of Fig. 5.2 and Table 5.2, the simulated HA behaviour consistently fell within the prototypical trajectory areas of the



(a) SCD vs SCA (b) Initial vs Successive (c) Cluster identification or not

Figure 5.1: Initial conditions were designed to specifically test leading policies. Each test would confirm or negate the validity of successive collinear angle (SCA) against competing subsuming policies. (a) comparing the successive collinear distance and SCA, (b) compares initial collinear angle and SCA, and (c) examining the relationship between successive angle and SCA. TAs are labelled A, B, and C, and their distances from the HA (with initial position in yellow), including the angle of the HA subtended at the centre of the containment zone, are indicated. The expected order of the TA engagement according to the competing tested policies is also included for the first two plots. In (c), it was found that individuals had different thresholds for cluster identification, and so the expected policy outcome is not included here but is detailed in sub-section 5.1.5 on Policy Validation.

participants, producing sequences of HA trajectories that overlapped with prototypical human behaviour 97% of the time. Figures of all model fits and binary and weighted trace maps are presented in Appendix B.

5.2 Discussion

This Multi-Target Herding experiment involved a more complex herding scenario where participants were required to corral three targets randomly placed around the game field. In addition to examining whether the newly proposed DPMP navigation model could be generalised to a multi-target task context, this experiment aimed to determine the target selection policy or policies that participants employed to complete the task. An analysis and subsequent validation experiment of the order in which participants corralled the targets revealed that participants employed the same target selection policy in almost 80% of the trials. This policy, termed the Successive Collinear Angle (SCA) policy, involved three heuristic rules. First, participants chose targets closer to them in angular distance measured with respect to the centre of the containment zone (rather than linear Euclidean distance). Second, they choose successive TAs nearest in angular distance to the previously chosen TA. Lastly, if two (or more) TAs were collinear or nearly collinear with respect to the containment zone (that is, perceived to be part of the same corral cluster relative to approaching the containment zone), then participants corralled the TA farther from the containment zone first, even if it was more distant from them in angular terms. While this experiment had identified and successfully validated the intricacies of

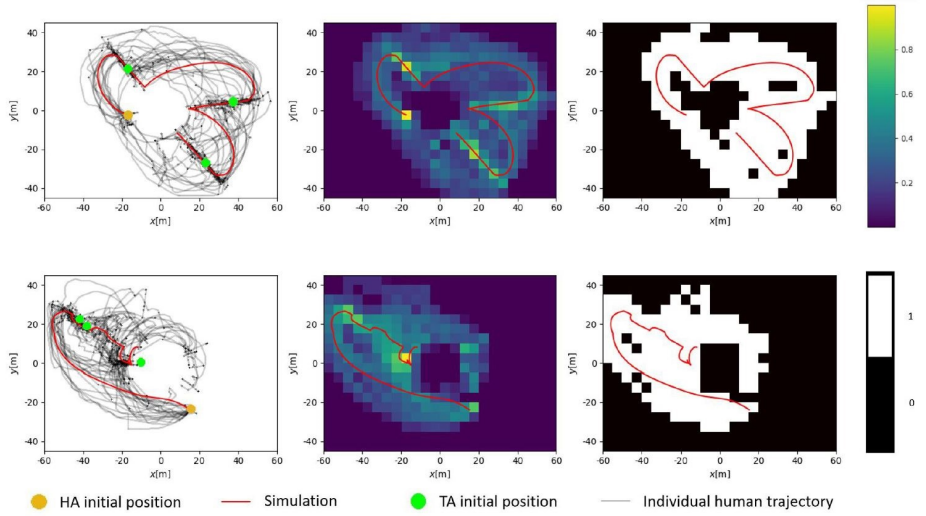


Figure 5.2: Plots of Simulations, Human Data, and Evaluation Methods. The first panel on the left displays both human and simulated trajectories (legend consistent with other 2D trajectory plots in this article). The upper plot demonstrates the selection of TAs based on the closest angular distance on a successive basis ("Successive Angle") and the lower plot illustrates the selection based on angular proximity, chosen successively ("SCA"). The second panel overlays the simulation onto a non-linear heatmap representing the spatial frequencies of human trajectories. The third panel features the simulation overlaid on the binary map derived from the non-linear heatmap. Colorbars indicating the relative spatial frequencies of the non-linear heatmap and the binary map are included in the final panel.

Trial	Mean	SD	Sim
1	0.93	0.06	0.98
2	0.94	0.05	0.98
3	0.90	0.09	0.98
4	0.95	0.05	1.00
5	0.92	0.07	1.00
6	0.93	0.08	0.97
7	0.91	0.08	0.96
8	0.92	0.10	0.98
9	0.93	0.07	1.00
10	0.94	0.05	0.92
11	0.92	0.08	0.93
12	0.94	0.06	0.94
13	0.95	0.07	1.00
14	0.92	0.07	1.00
15	0.93	0.08	1.00
16	0.94	0.08	0.87
17	0.92	0.06	0.88
18	0.94	0.04	0.98
Average	0.93	0.07	0.97

(a) Binary traces

Trial	Mean	SD	Sim
1	23.55	2.06	25.34
2	21.87	1.83	24.34
3	23.38	2.96	25.99
4	24.48	1.97	27.25
5	24.27	1.66	27.59
6	23.38	2.68	25.32
7	22.42	1.65	24.34
8	23.45	2.71	25.24
9	24.31	2.04	26.73
10	23.85	1.85	25.15
11	22.51	2.29	24.40
12	25.70	2.24	27.11
13	24.53	3.11	28.22
14	23.88	2.42	25.48
15	23.54	1.95	27.59
16	24.66	2.91	23.86
17	21.05	1.99	22.11
18	23.92	1.62	25.79
Average	23.60	2.22	25.66

(b) Weighted Traces

Table 5.2: Traces of the human vs. the simulated trajectories. The mean and SD columns (taken across participants) refer to the human data. Sim refers to simulated trajectories.

multi-target herding, the next chapter will examine its generalisability to multiple human herders herding additional (≥ 3) TAs.

6 Multi-player herding

The overall aim of the study presented in this chapter is to examine whether the navigational and decision model developed so far could be extended to a more complex, multi-player, multi-target, first-person herding task environment. It is important to recognise that achieving this aim required identifying the target selection policy that multiple (two) human players employed to collaboratively corral a group of TAs ($N_{\mathrm{TAs}} \geq 3$) within a containment area. Thus, the study also sought to identify the TS policy that could most accurately capture the action-decision behaviour that human herders employed in a collaborative first-person herding task. Of particular interest was whether different human herders converged on to the same target selection policy. To achieve these aims, the first-person herding task was adapted to a two-player/HA task, whereby each participant in a pair controlled a virtual HAs and were required to corral three to five TAs into a containment zone located within the centre of a game field. An analysis of the participant TS behaviour was then conducted to identify which TS policy most represented cooperative player behaviour. This policy was then combined with the navigational model developed in this Thesis, to evaluate the effectiveness of the combined movement and target selection model in simulating the human trajectories. It was expected that the navigational and decision models would extend to the new task environment, where the model would capture players' behaviours within the herding task. Moreover, in line with prior research modelling human behaviour [16, 5], the expectation for the present study was that simulated trajectories of the combined movement and target selection model would closely resemble the human (participant) data.

6.1 Identifying the decision policies

6.1.1 Tested Target Selection Policies

Drawing upon the study in Chapter 5 and other prior research examining human target selection decisions in herding tasks [33, 15, 50], the present paper examined whether SCA and four other TS policies (henceforth TS*p*'s) best captured the decision making behaviour of participants in the multi-agent-multi-target herding task. All of the policies examined were deterministic, with rules based on the either the relative Euclidean or angular distances of the co-herder, the containment zone, and all of the TAs in the herd. As in the previous study (chapter) , a "global" information approach was employed given

that a participant could see the positions of all agents simply by moving their HA's head around. Note also, that prior to selecting a TA, all of the examined TS*p*'s accounted for the TAs' distances from the participants co-herder (the other HA), such that each policy included a rule that a participant would select a target only if it were closer to oneself rather than being closer to the other HA. There was one exception to this rule, if all TAs were closer to the other HA (the other participants HA), then the policy would choose the TA furthest from the other HA.

As an illustration of how each policy was implemented, the logic for the SCA policy is detailed in Table 6.1. The four other policies examined were: Successive Collinear Distance (SCD) — which was the same as SCA, except that it used Euclidean distance to successively order TAs rather than angular distance; Successive Angle (SA) and Successive Distance (SD) — which excluded the collinear rule; and Distance From Containment Zone (DCZ) — where targets were chosen based on their Euclidean distance from the containment zone (furthest to closest). Again, all these policies included the TA-to-other-HA distance rule. All policies also included a 0.25 second delay between (re)decisions updates and a 1 second delay prior to making a new TA selection decision after corralling a TA into the containment zone.

These five policies were chosen because consistent with the results of Chapter 5, a preliminary analysis of the data revealed that participants selected targets successively (i.e., target-by-target), in contrast with selecting TAs based on their Euclidean or angular distance from the HA's initial (starting) position. Moreover, out of the 36 different policies examined in Chapter 5, these five represented the best alternatives to SCA, with SCD almost equal in performance to SCA in Chapter 5. That is, again, they all captured the successive (i.e., target-by-target) manner by which participants made their target selection decisions, they simply differed with regards to the use of Euclidean vs angular distance variables and the presence or absence of the Collinear rule.

Table 6.1: Rule logic of the successive collinear angle (SCA) target selection policy

Step	Condition	Action
1	HA current target = null (otherwise skip to Step 2)	Select the closest TA to the HA based on the angular distance of TAs from the HA. – If the selected TA is closest to other-HA, skip this TA and select the next closest TA. – If all TAs are closer to other-HA, select TA furthest from the other-HA. – Assign (return) the selected TA as HA's current target.
2	Check for collinear TAs	Identify if other TAs are within 20° of the currently selected TA. – If no collinear TAs, skip to Step 3, otherwise... – Order the collinear TAs based on their Euclidean distance to CZ. – Assign (return) furthest TA from CZ as HA's current target (resulting in target switching when corralling collinear TAs).
3	Check for TA containment	If HA's currently selected TA is in the containment zone and stationary, current target = null.
4	Pause and Return	Wait 0.25 seconds then return to Step 1.

Note: HA = Herding Agent; TA = Target Agent; CZ = Containment Zone.

6.2 Generalising the movement model

To generalise the navigational model, Eq. (5.1) required a minor modification to account for the presence of the additional HA and TAs in the game field. Specifically, the introduction of another HA brought about the need for avoidance behaviour towards the other HAs. Similarly, as in Chapter 5, as the TAs were typically corralled individually, the TAs that were not selected were treated as obstacles, referred to as "non-targeted targets".

To adapt the model to these conditions, an obstacle term for the other herding agent (not oneself) was added into the model. This term mirrored the existing obstacle term used for the containment zone and for the non-targeted targets. Thus, the resulting model was of the form,

$$\begin{aligned}\ddot{\phi} = &- b\dot{\phi} + \psi_{\mathcal{G}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{G}}) + \psi_{\mathcal{T}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{T}}) \\ &+ \sum_{i=1}^{N_O} \psi_{O_i}(\mathbf{x}_{O_i}, \mathbf{y}, \phi, \theta_{O_i})\end{aligned} \tag{6.1}$$

where the only difference between model in Eq. (5.1) and this updated model is represented by the ψ_{O_i} term on the right-hand side of Eq. (6.1), which now contains the additional obstacle coupling term corresponding to the other HA. Indeed, the number of obstacle terms N_O is given by

$$N_O = \underbrace{1}_{\text{for the containment zone}} + \underbrace{N_{\text{TAs}} - 1}_{\text{non-targeted TAs}} + \underbrace{N_{\text{HAs}} - 1}_{\text{other HAs}} \tag{6.2}$$

All parameters and functional forms for each of the terms in Eq. (6.1) carried over from the previous single-herder multi-target presented study in Chapter 5. To decide which TA was currently targeted, the respective TSp was employed. The TSp determined which terms in Eq. (6.1) were assigned to which TAs - i.e., which of the TAs were targeted and thus modelled by $\psi_{\mathcal{T}}$, and which were non-targeted and modelled by terms in ψ_{O_i}.

6.2.1 Target Selection Policy and Model Validation

To validate the model, Eq. (6.1) and explore the effectiveness of the model for replicating human-human herding behaviour when coupled with the different TSp's, simulations of the model with each of the five TSp's (i.e., SCA, SCD, SA, SD, and DCZ) were recorded for each of the initial conditions employed for the experimental trials (i.e., trials 7 to 24). Importantly, all model parameters use for the simulations were the same as those employed in Chapter 5, including using exactly the same repulsive coupling values for other-HA avoidance that were used for the avoidance of non-targeted TAs and the containment zone. These simulations were integrated at a time-step of 0.02 seconds, with a given TSp determining the current TA used to set the relevant parameters in

$\psi_{\mathcal{T}}(\mathbf{x}, \mathbf{y}, \phi, \theta_{\mathcal{T}})$ in Eq. (6.1) every 0.25 seconds. As noted above, there was also 1 second delay prior to making a new TA selection decision after corralling a TA into the containment zone.

Given these simulations and the recorded human-human (participant pair) data, we identified which TSp and model combinations best captured the human data using two methods.

Target-selection overlap

The first method, resulted in a normalised *target-selection overlap* score. This method involved calculating the Dynamic Time Warped (DTW) distance (or error) between time series of TA engagements of participant and model-controlled (simulated) HAs of over the course of a trial.

More specifically, a 50Hz time series (corresponding to the 50Hz of data collection frequency) was extracted for each participant and each model-controlled HA for each trial, indicating when they were corralling a TA (i.e., within the influence region of a TA). For every timestep, the presence of a given HA-TA pair engagement was tested and a N_{TA} - dimensional binary-encoded vector (that is, the vector elements could be either '0' or '1') was created, with '1' corresponding to the given HA-TA engagement being present at that particular timestep, and '0' corresponding to no engagement between the HA-TA pair at that timestep (refer also to Alg. 1). This resulted in a binary (values being either 0 or 1) time series of length being the length of the trajectory data, and width being N_{TA} corresponding to each HA-TA pair, thus accounting for all HA-TA pairs.

Next, the DTW distance was calculated between the given human-controlled HA target selection timeseries and the simulation target selection timeseries. Note that DTW distances can be calculated between two timeseries of different lengths, but of the same dimension. Given that the target number and herder number was fixed across the compared trials, this could be done here.

Then, the DTW distance, being non-normalised, was further normalised to a value between 0 and 1 (i.e., between 0 and 100%) by dividing the DTW distance by the product of the lengths of the two tiemseries (human-controlled and simulated). Finally, to obtain a value of confidence, rather than of error, the normalised DTW distance was subtracted from 1 to give a value where 1 indicated a perfect target selection match between the two target selection timeseries, and 0 no match.

This resulted in a target-selection overlap score between 0 and 1 for each ($HA_{\text{sim},j}$, $HA_{\text{part},i}$) combination, with 0 indicating 0% overlap in TA selection decisions and 1 indicating 100% or perfect overlap in time-normalized TA selection decisions. Thus, higher values reflect a better match between the simulated TSp and the actual TA selection policy employed by participant-i for a given initial condition.

The target-selection overlap for each ($HA_{sim,j}$, $HA_{part,i}$) pair was calculated for each initial condition of the experimental trials (i.e., 7 to 24) and the corresponding player position (i.e., 1 and 2). For each of the five TS policy and model combinations, this resulted in 21 target-selection overlap scores (21 pairs) for each player position (1 and 2) for each trial. The average across the 21 target-selection overlap scores was then calculated as the final target-selection overlap score for a given TS policy and model

Data: $HA_{sim,j}$ and $HA_{part,i}$
Result: value between 0 and 1 indicating no-to-perfect match, respectively
while $t < len(HA_{sim,j})$ **do**
  binary-encoding-sim(t) ← (0,0,. . . ,0) of length N_{TA}
  if $HA_{sim,j}(t)$ *interacted with* $TA_k(t)$ **then**
    binary-encoding-sim(t)$_k$ ← 1
  end
end
while $t < len(HA_{part,i})$ **do**
  binary-encoding-part(t) ← (0,0,. . . ,0) of length N_{TA}
  if $HA_{part,i}(t)$ *interacted with* $TA_k(t)$ **then**
    binary-encoding-part(t)$_k$ ← 1
  end
end
DTW ← DTW(binary-encoding-part, binary-encoding-sim)
normalised DTW ← $\frac{\text{DTW}}{\text{len}(HA_{sim,j})\times\text{len}(HA_{part,i})}$
output ← 1 − normalised DTW

Algorithm 1: Calculating DTW TSp overlap scores. Binary-encoding-sim and binary-encoding-part refer to simulation and participant binary encodings respectively.

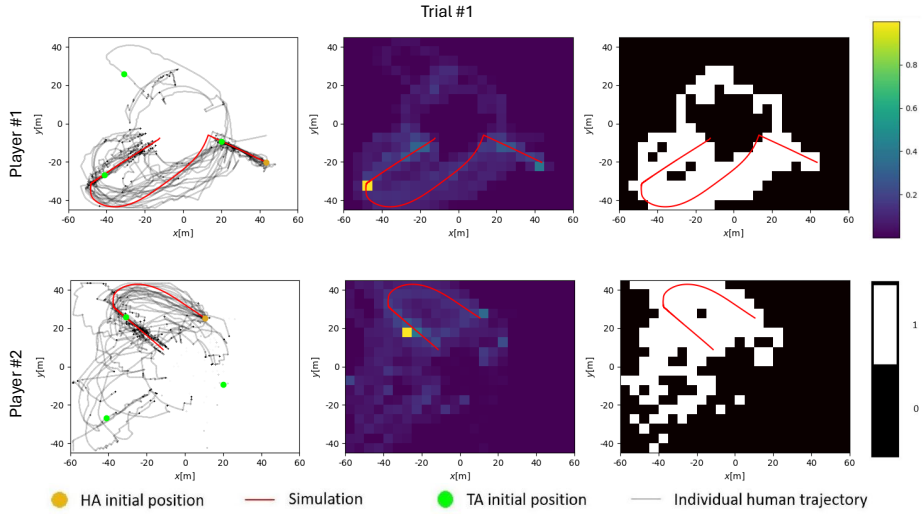
combination for a given trial and player position. These scores were then averaged across trials for each target number condition, resulting in a 3 (target condition: 3, 4 and 5) × 5 (TS policy: SCA, SCD, SA, SD, SCZ) × 2 (player positions) × 21 (participant pairs; i.e., N) target-selection overlap score data set.

Binary trace overlap

The second method employed to evaluate the degree to which each TS policy and model, Eq. (6.1), combination best replicated the human-human (participant data) was the same binary trace overlap method employed in Chapter 5. To calculate this score, weighted trajectory heat-maps were first generated for each trial from recorded participant data by applying a non-linear (specifically, square root) filter on the spatial (planar) frequencies of the participant's HA trajectories, binned into square 5-by-5 m bins. The bins of the resulting weighted heat-map or "weighted trace map" was then filtered using a cutoff value of 10, to eliminate bins where only one or two trajectories passed through that bin to produce a binary trace map (see Figure 6.1). For a given simulated HA trajectory, the percentage of the trajectory falling within this binary trace map was then calculated as the *binary trace score* for that HA for a given initial condition, with 0 correspond to no overlap and 1 corresponding to a HA trajectory path the fell completely within the trajectory area of approximately 90% of human trajectory data; i.e., the binary trace map represented an approximation of the 90% confidence interval for the human trajectory data used to generate the map.

To generate a data set appropriate for statistical testing, a kind of surrogate analysis

Figure 6.1: An example of the participant and simulated HA trajectories and TA engagement (data for trial 7, or the first analysis trial). The left panels display the participant trajectory data for each player using gray lines, with the red lines corresponding to the simulated trajectories of Eq. (6.1) using the SCA target selection policy. The green and orange dots correspond to the TA and player initial positions, respectively. The middle and right most panels illustrate the weighted and binary trace maps used to determine the the degree to which the simulated trajectory captured prototypical human behavior, with the corresponding color-bars on the far right indicating the relative spatial frequencies of the non-linear weighted and the binary trace maps. Refer to main text and Chapter 5 for more details on the generation of these maps.



was performed in which an $N = 21$ binary trace scores were calculated for each ($HA_{sim,j}$ for each initial condition of the experimental trials (i.e., 7 to 24) and corresponding player position (i.e., 1 and 2). This was done by creating 21 binary traces maps for each trial, by removing the data of one participant pair (1 to 21) for each binary trace map. That is, for each participant pair a binary trace maps was produced which excluded that pairs data, with the binary trace overlap score for ($HA_{sim,j}$ calculated for each map. For each of the five TS policies and model combinations, this resulted in 21 scores (21 simulated HA pairs) for each player position (1 and 2) for each trial. These scores were then averaged across the trials for a given target number condition, resulting in a 3 (target condition) × 5 (TS policies) × 2 (player positions) × 21 (simulated HA pairs) binary trace score data set.

6.3 Results

Figure 6.2 includes box plots of the target selection overlap scores (top graph) and binary trace overlaps scores (bottom graph), averaged across participant and trial for each target number condition. Recall that for both measures, high scores indicating better overlap, with 1 corresponding to 100% or perfect overlap and 0 corresponding to no overlap at all. With respect to target selection overlap, this measure can also be interpreted as the predictive power or normalised accuracy of the corresponding TSp.

In order to determine the degree to which the different TSp's better predicted the human target selection decisions and the degree to the Eq. 6.1 + TSp combination best capture the overall behavioural dynamics of the participant pairs, a mixed-effects linear model was used to analyse the target selection and binary trace overlap scores as a function of target number and policy type. As noted above, scores were averaged across the trials for the same target number condition, such that the model employed for analysis included fixed effects for target number and policy, as well as their interaction, and random intercepts of player (1 vs 2; with respect to initial condition) nested within participant pair (N = 21). The Kenward-Roger approximation was applied to calculate F-tests for the fixed effects (rather than χ^2 tests), ensuring more reliable p-values. Due to the balanced nature of the data, the degrees of freedom were integer values. All post-hoc comparisons were conducted with Tukey HSD correction to conserve family-wise error. The analysis was conducted using R (version 4.0.3) and the following packages: lme4 (for fitting mixed-effects models), lmerTest (for Type III ANOVA F-tests using the Kenward-Roger method), and emmeans (for post-hoc pairwise comparisons).

The analysis of target selection overlap revealed a significant main effect of the target number, $F(2, 574) = 216.798$, $p < 0.001$, a significant main effect of policy, $F(4, 574) = 9.504$, $p < 0.001$, as well as a significant interaction between target number and policy, $F(8, 574) = 2.762$, $p < 0.01$. Although the difference between the different TSp's was rather small (see Figure 6.2, top), consistent with the results of Chapter 5, SCA did overall better predict the participants' target selection decisions than the other policies, particularly for the 4 and 5 target conditions. Indeed, post-hoc analysis examining the differences between the policies for each target number condition revealed that although the only significant difference for the 3-target condition was between

Successive Containment Zone and Successive Distance ($p = 0.045$), SCA better predicted human target selection decisions than the Successive Angle ($p = 0.017$) and Successive Containment Zone ($p < 0.001$) polices for the 4-target condition, and both Successive Angle ($p < 0.001$) and Successive Distance ($p < 0.001$) for the 5-target condition.

Also consistent with the results of Chapter 5, Successive Collinear Distance or SCD performed comparably to SCA, with Tukey's HSD post-hoc revealing no significant differences between SCD and SCA for any of the target number conditions (all $p > 0.861$). Furthermore, SCD also significantly outperformed Successive Angle ($p = 0.044$) and Successive Containment Zone ($p = 0.003$) for the 4-target condition, and Successive Angle ($p < 0.001$) and Successive Distance ($p = 0.025$) for the 5-target condition.

The analysis of the binary trace overlap scores also resulted in a significant main effect of the target number, $F(2, 574) = 29.599$, $p < 0.001$, as well as a significant main effect of policy, $F(4, 574) = 224.836$, $p < 0.001$, and a significant interaction between target number and policy, $F(8, 574) = 51.039$, $p < 0.001$. As can be discerned from an inspection of Figure 6.2 (bottom), post hoc analysis conducted for each target number condition, revealed that Successive Containment Zone policy resulted in significantly worse overlap scores compared to SCA and SCD for all target number conditions (all $p < 0.001$). The Successful Angle policy was also significantly worse than the SCA and SCD policies for the 4 and 5 target number conditions (all $p < 0.001$). Although SCD and Successive Distance policies did result in slightly higher overlap scores compared to SCA for the 3 and 4 target conditions, this difference was not significant (all $p > .065$). Finally, SCA did result in higher binary trace overlap score for the 5 target conditions, with this difference being significant with respect to Successive Distance ($p < 0.001$), but not significantly higher than SCD ($p = .142$).

6.4 Discussion

The aim of this experiment was twofold: first, to evaluate whether the single-herder DPMP navigational model proposed in Chapter 5 could be extended to a more complex, multiplayer context; and second, to identify the target selection policy employed by participants in a cooperative multi-herder task. With regard to the latter aim, we sought to determine whether the Successive Collinear Angle (SCA) target selection policy identified in the single-herder study in Chapter 5 could also capture the majority of target selection behaviour in the multiplayer herding task explored here.

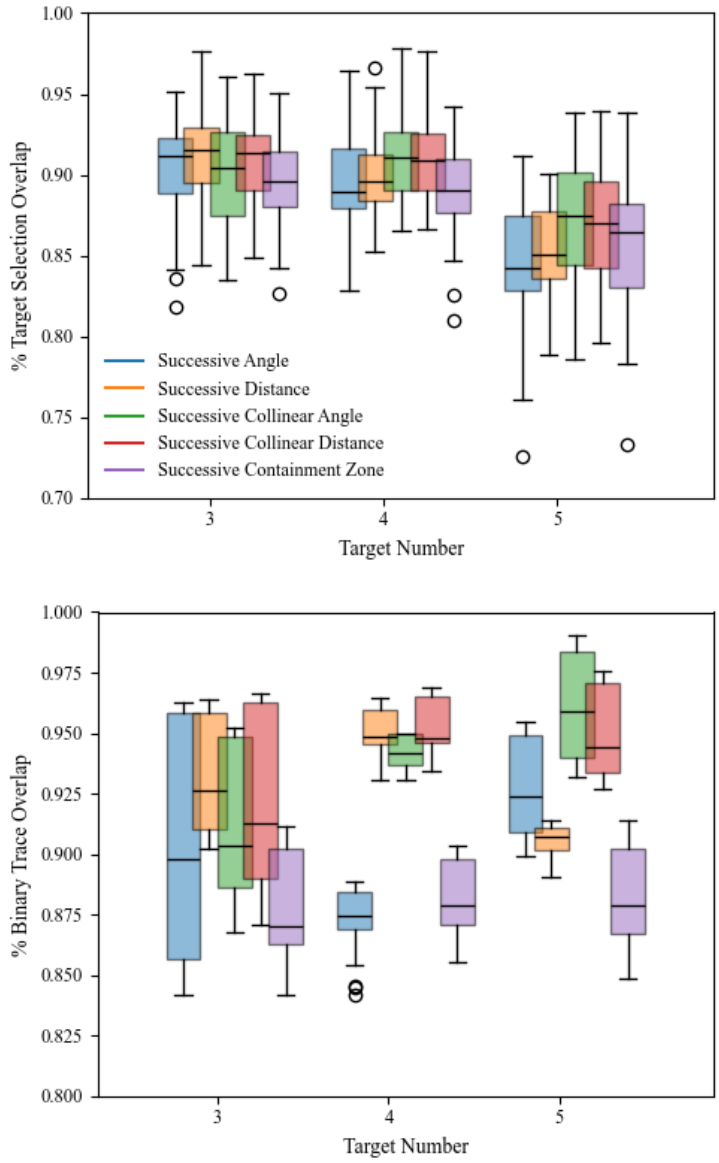
Overall, the findings provided clear evidence that the DPMP navigational model can be generalised to a multi-herder context. Indeed, as can be easily discerned from an inspection of Figure 6.1 and the figures in Appendix C, Eq. (6.1) could effectively capture the movement trajectories of participant pairs collaboratively herding 3, 4, or 5 targets. Interestingly, the model was relatively robust across the tested target selection policies, with the behavioural dynamics of simulated herder agent (HA) trajectories closely mirroring the topological form of participant trajectories, regardless of the target selection policy employed. Accordingly, these findings provide further evidence that goal-directed human movement can be effectively modelled using simple dynamical principles [5, 7, 9, 38, 16, 2]. Moreover, that the complexity of cooperative human

behaviour can emerge from real-time agent-environment interactions, underscoring the self-organising nature of human behaviour in cooperative task settings ([20, 21, 38]).

With regard to the target selection policy employed by participants, the results demonstrated that the SCA policy was marginally better than the other policies investigated, particularly as the number of targets to be herded increased. Consistent with the findings of Chapter 5, the performance of the Successive Collinear Distance (SCD) policy in predicting participants' target selection behaviour was comparable to that of the SCA policy. This similarity is expected given that the difference between these two policies is simply whether the distance of targets from the herder agent's (HA) current position is defined with respect to Euclidean or angular distance, which in most situations leads to identical predictions. Indeed, the fact that the collinear rule only applied in a subset of the 18 experimental trials also accounts for the small differences in predictive performance between the Successive Angle and Successive Distance policies compared to SCA and SCD, particularly in the three-target condition where the collinear rule never applied.

Nevertheless, consistent with previous work (Chapter 5, [50]) including research exploring cooperative herding in tabletop games [34, 16, 48, 46] and larger-scale contexts [49, 66], the results of this experiment do suggest the majority of participants employed a target selection policy at least similar to SCA (or SCD). Of course, one way this can be tested further is to examine the behavioural dynamics of human participants completing the multiplayer herding game employed here alongside an HA controlled via Eq. (6.1) and the SCA policy. Accordingly, the experiment in the following chapter was designed to assess whether the combined DPMP and SCA herding model simulated in this experiment could effectively control the behavioural dynamics of an artificial herder within a human-machine teaming context.

Figure 6.2: Box plots of target selection overlap (top) and binary trace overlap (bottom) for the different TSp as a function of target number.



7 Multi-agent herding: human-autonomy teaming

7.1 Introduction

In Chapter 6, we observed that a rule-based or heuristic approach effectively simulated human decision-making in the first-person shepherding task. The SCA policy for the multi-TA first-person task closely mirrored the behaviour of human HAs in tabletop (bird's-eye view) herding tasks [16, 48], and it aligned with the optimal strategy for herding both small and large groups [33]. The simplicity of the SCA and heuristic methods in general is beneficial when integrated with DPMPs into AA control systems, as these policies are fully deterministic and produce predictable, human-like decisions that support reciprocity in Human-Machine Interaction (HMI) [73]. However, heuristic policies represent simplified versions of human decision-making processes, often excluding the subtle dynamics, adaptability, and randomness characteristic of human behaviour [33]. Moreover, their task-specific nature limits their applicability to new or increasingly complex tasks [34], where even small shifts in task dynamics may require time-consuming rederivation of the heuristics [74].

7.1.1 Deep Reinforcement Learning for Target-Selection Action Decisions

To address the limitations of heuristic policies, an increasing body of research has been investigating whether machine learning methods, particularly Deep Reinforcement Learning (DRL), can more effectively modulate DPMP models for AA control. DRL is increasingly used to derive target-selection policies for guiding DPMPs [37, 45, 46], offering greater flexibility compared to the rigidity of heuristic-based AAs [75]. Rather than relying on human-inspired rules, DRL methods are designed to discover optimal action-decision policies through autonomous self-learning processes [76], often resulting in DRL AAs that achieve equal or better task performance compared to humans [77].

A deep neural network (DNN) is embedded in the DRL architecture, which learns iteratively via reinforcement learning, through trial and error, to form a behavioural policy (i.e., state-action mapping) that maximises expected task rewards [57]. As pow-

erful function approximators, DNNs allow DRL AAs to make decisions based on high-dimensional data [78, 79]. Crucially, DRL AAs can learn to perform tasks via self-play, without needing prior knowledge of task dynamics [53, 80]. In DRL, self-play (SP) refers to the process where an AA trains by playing alongside copies of itself, without direct human involvement.

Despite surpassing expert human performance in complex, competitive video games, DRL AAs often adopt decision strategies that differ significantly from those used by humans [52]. While excelling in SP tasks, this discrepancy in strategies compared to humans frequently leads to poor HMI and suboptimal performance [37, 64], which is a key aim of AI development [81]. Furthermore, human collaborators often perceive SP-DRL AAs as unpredictable and uncooperative [46]. Effective multi-agent human collaboration requires anticipation of intentions and reciprocity of behaviours between co-actors [82], yet SP-DRL AAs assume their human counterparts will make decisions that maximise positive outcomes in the same way, resulting in misunderstandings and ineffective HMI [64, 46].

To address HMI challenges with SP-DRL AAs, human-sensitive DRL strategies have been proposed [64, 63]. In these strategies, DRL AAs are trained alongside a human proxy AA controlled by either a heuristic or computational model. Given that DRL training typically requires millions of time steps, real-time human involvement is impractical. However, using the game Overcooked, [64] demonstrated that a "human-aware" DRL AA (referred to as HS-DRL), trained with a human behavioural clone, adapted better to human co-actors compared to SP-DRL trained AAs.

Although a similar HS-DRL approach has been suggested for determining action-decision policies for DPMP-controlled AAs [37], it remains unclear whether this leads to better HMI than SP-DRL or heuristic approaches [46]. Furthermore, no research has yet compared SP-DRL and HS-DRL target-selection policies for DPMP HAs in a first-person, multiplayer shepherding task. Addressing this gap was one of the objectives of this study.

7.1.2 Current Study

Building on previous chapters, the present study aimed to: (1) validate the SCA policy for DPMP herder control and assess its generality in a human-AA dyadic context, and (2) benchmark its effectiveness by comparing the performance of AAs with target-selection decisions determined by the SCA (Heuristic), SP-DRL, or HS-DRL policies. Although their action decisions varied, all AAs shared the same DPMP action dynamics (movement) model. Additionally, this study sought to determine whether the HS-DRL AA led to better human-AA performance than the SP-DRL AA, and whether it more closely resembled the performance of the Heuristic AA.

To achieve these objectives, human-AA dyads performed a multiplayer shepherding task, where they corralled three to five TAs into a containment zone. Using a between-subjects design, participants co-herded with either the Heuristic, SP-DRL, or HS-DRL AAs. To evaluate the effects of the three policies on HMI, the selection decisions and movement trajectories of both AAs and participants were analysed. Decision and trajectory measures were compared to typical human behaviour from human-human

dyads performing the same task (Chapter 6), with scores quantifying how closely an HA's behaviour aligned with typical human herding.

Since the SCA was human-derived, the Heuristic AA was hypothesised to best capture human herding dynamics, displaying the highest decision and trajectory overlap scores, thereby validating the SCA for DPMP-HA control in HMI. Regarding the DRL AAs, the HS-DRL AA was expected to show greater overlap scores (i.e., more human-like) than the SP-DRL AA. Similarly, participants co-herding with the Heuristic AA were predicted to exhibit behaviour more aligned with human-human baseline data compared to participants paired with HS-DRL and SP-DRL AAs, reflecting the Heuristic AA's ability to effectively embody a human co-herder (i.e., co-actor) for enhanced HMI. Finally, participants co-herding with the HS-DRL AA were also expected to display more human-like behaviour than those paired with the SP-DRL AA.

7.2 Artificial Agents for Target-Selection Policies

The three AA types formed dyads with participants, creating three independent conditions. It is important to note that although all AAs shared the same DPMP navigation model, they varied in their target-selection policies. Each policy output was assigned a TA identification (ID) number (i.e., 1, 2, 3, 4, 5, with 0 indicating no TA).

7.2.1 Heuristic (SCA) Artificial Agent

The Heuristic AA is a rule-based agent whose target-selection decisions were governed by the Successive Collinear Angle (SCA) policy described earlier in Chapter 6. This AA first selected TAs based on angular distance and then selected successive TAs based on their proximity in angular distance to the previous target. If two TAs were collinear with respect to the containment zone, the AA would select the furthest TA first, then influence the second one while en route. In this study, the SCA policy was adapted according to the other HA's location. Specifically, the Heuristic AA would choose a TA farther from the participants but closer to itself.

7.2.2 Self-Play-DRL Artificial Agent (SP-DRL)

Self-Play-DRL AAs were created using a fully connected multilayer perceptron (MLP) neural network (2 hidden layers, 128 nodes each), trained with the Proximal Policy Optimisation (PPO) algorithm, selected for its strong learning capability and high efficiency in state-action-reward processes. The inputs to the MLP included the positions (x, y) and velocities of all HAs and TAs. Rewards were updated at each environment step with the following structure to guide behaviour: a small negative reward per timestep (-0.0001), a positive reward for successfully containing a TA (0.6 divided by the number of TAs in a trial), and a negative reward for a TA escaping the containment zone (-1.2 divided by the number of TAs in a trial). This reward system encouraged efficient herding by discouraging delays, emphasising TA containment, and enforcing strict monitoring to prevent escapes. Ten SP-DRL AAs were trained using 20 parallel environment copies

on CPUs running at 40× regular speed, over 150 million environment time steps. During the simulations, the AAs trained by engaging in self-play with another DRL AA, without any human data. At the end of training, the two best-performing SP-DRL AAs were selected based on their task completion times across all trial sets.

7.2.3 Human-Sensitive-DRL Artificial Agent (HS-DRL)

Human-Sensitive-DRL AAs utilised the same MLP architecture as the SP-DRL AAs, with selection policies refined using PPO and the same reward structure. Ten HS-DRL AAs were trained over 150 million environment steps, with the two best-performing AAs chosen for this study. The key difference from the SP-DRL AAs is that HS-DRL AAs employed hybrid-DRL training to develop target-selection policies [37]. In this case, HS-DRL AAs trained with a Heuristic AA rather than using self-play with another DRL AA.

7.3 Results

Separate mixed-design ANOVAs were conducted to investigate differences in target-selection (decision) and binary trace (trajectory) overlap scores for both AAs and participants. Note that the overlap scores had been calculated in the same manner as detailed in previous chapters, in particular in sections 6.2.1 and 6.2.1. For the AA analyses, the between-subjects independent variable (IV) was AA Type, which included the Heuristic, HS-DRL, and SP-DRL AAs. For the participant analyses, the corresponding IV was AA-Coherder, referring to the participant conditions, where they co-herded with either the Heuristic, HS-DRL, or SP-DRL AAs. Additionally, the previous Chapter 6's study on shepherding with human-human dyads found that participants' trajectory scores varied based on the number of TAs in a trial, with trials involving an even number of TAs yielding higher trajectory scores. Consequently, TA-Num (the number of TAs in the trials) was included as a repeated factor with three levels: 3 TA, 4 TA, and 5 TA. Stata version 18 was employed for all statistical analyses. The nominal alpha level was set at $\alpha = .05$. Post-hoc pairwise comparisons were adjusted using the Bonferroni correction. Assumptions of normality and homoscedasticity were evaluated by inspecting histograms, Q-Q plots, and Residual vs. Fitted (RVF) plots, respectively. Mauchly's test was used to assess sphericity for the TA-Num variable, and any assumption violations were addressed through transformations and corrections as required.

7.3.1 Artificial Agents on Target-Selection Decision Overlap

A 3 (AA-Type) × 3 (TA-Num) mixed-design ANOVA was conducted to evaluate differences in target-selection decision overlap scores among the Heuristic, HS-DRL, and SP-DRL AAs. The assumptions of normality and homoscedasticity were satisfied. However, Mauchly's test indicated a violation of sphericity for the TA-Num effect, $\chi^2(2) = 12.98$, $p = .525$. As a result, the Greenhouse-Geisser correction was applied ($\epsilon = 0.976$).

After applying the correction, the ANOVA revealed significant main effects for AA-Type, $F(2, 177) = 33.46$, $p < .001$, $\eta_p^2 = .274$, and TA-Num, $F(2, 177) = 197.57$, $p < .001$, $\eta_p^2 = .691$. There was also a significant AA-Type × TA-Num interaction, $F(4, 177) = 8.34$, $p < .001$, $\eta_p^2 = .159$. To further investigate this interaction, simple effects were analysed at each TA-Num level.

For 3 TA trials, there were no significant differences in decision overlap scores between AAs, $F(2, 59) = 2.93$, $p = .061$. However, there was a significant effect of AA-Type for 4 TA trials, $F(2, 59) = 4.79$, $p = .012$, with post-hoc comparisons showing that the Heuristic AA had higher scores than the SP-DRL AA (contrast = .015, $p = .010$). The HS-DRL AA did not significantly differ from the Heuristic AA (contrast = -.009, $p = .203$), nor from the SP-DRL AA (contrast = .006, $p = .736$). The AA-Type effect was also significant for 5 TA trials, $F(2, 59) = 29.99$, $p < .001$. Bonferroni-adjusted comparisons indicated that the Heuristic AA showed higher decision overlap compared to both the HS-DRL AA (contrast = .029, $p < .001$) and the SP-DRL AA (contrast = .051, $p < .001$). Additionally, the HS-DRL AA exhibited higher decision overlap than the SP-DRL AA (contrast = .022, $p = .005$). As depicted in Figure 4 (Panel A), these results overall support the hypothesis that the Heuristic AA's decisions would more closely align with typical human TA selections compared to the DRL AAs. The expectation that the HS-DRL AA would display more human-like decisions than the SP-DRL AA was also confirmed for 5 TA trials.

7.3.2 Artificial Agents on Binary Trace Overlap

Differences in binary trace overlap scores among the AAs were analysed using a 3 (AA-Type) × 3 (TA-Num) mixed-design ANOVA. The assumptions of normality, homoscedasticity, and sphericity ($p = .151$) were satisfied. The ANOVA revealed significant main effects of AA-Type, $F(2, 177) = 64.18$, $p < .001$, $\eta_p^2 = .420$, and TA-Num, $F(2, 177) = 3.29$, $p = .040$, $\eta_p^2 = .036$. However, the interaction was not significant, $F(4, 177) = 1.59$, $p = .179$, $\eta_p^2 = .035$.

As expected, Bonferroni-adjusted comparisons indicated that the Heuristic AA exhibited significantly higher trajectory overlap scores compared to both the HS-DRL (contrast = .071, $p < .001$) and SP-DRL AAs (contrast = .121, $p < .001$). Additionally, the HS-DRL AA scored significantly higher than the SP-DRL AA (contrast = .049, $p < .001$). As illustrated in Figure 4 (Panel B), these findings supported the hypothesis that the Heuristic AA would demonstrate the most human-aligned trajectories, followed by the HS-DRL and SP-DRL AAs.

Post-hoc comparisons for the TA-Num effect revealed that trajectory overlap scores were significantly higher in 3 TA trials compared to 4 TA trials (contrast = .028, $p =$

.034). There were no significant differences between 5 TA trials and 3 TA (contrast = .013, p = .731) or 4 TA trials (contrast = -.015, p = .497).

7.3.3 Participants on Target-Selection Decision Overlap

A 3 (AA-Coherder) × 3 (TA-Num) mixed-design ANOVA was performed to evaluate differences in target-selection decision overlap scores among participants co-herding with the Heuristic, HS-DRL, and SP-DRL AAs. The assumption of normality was satisfied, but visual inspection of the Residual vs. Fitted (RVF) plot indicated moderate heteroscedasticity. To address this, a log transformation was applied, which improved the consistency of residuals. The transformation did not affect the omnibus test F-statistics, effect sizes, or p-values, but changes were observed in the pairwise comparisons. Given the heteroscedasticity present in the raw data, the log-transformed results are considered more reliable and are reported here. Sphericity was violated, $\chi^2(2) = 14.36$, $p = .001$, so the Greenhouse-Geisser correction ($\epsilon = 0.866$) was applied.

After correction, the ANOVA revealed significant main effects for AA-Coherder, $F(2, 177) = 4.17$, $p = .017$, $\eta_p^2 = .045$, and TA-Num, $F(2, 177) = 85.14$, $p < .001$, $\eta_p^2 = .490$. However, the AA-Coherder × TA-Num interaction was not significant, $F(4, 177) = 1.61$, $p = .183$, $\eta_p^2 = .035$.

As shown in Figure 7.1 (Panel C), pairwise comparisons revealed that participants co-herding with the HS-DRL AA had significantly lower decision overlap scores compared to those co-herding with the Heuristic AA (contrast = 0.216, $p = .044$) and SP-DRL AA (contrast = 0.023, $p = .034$). Moreover, there were no significant differences in decision scores between participants co-herding with the Heuristic AA and those with the SP-DRL AA (contrast = -.001, $p > 1.00$), contrary to the hypothesis that participant decisions would differ when co-herding with DRL AAs compared to the Heuristic AA.

Not surprisingly, post-hoc comparisons for TA-Num showed that 3 TA trials resulted in significantly higher decision overlap scores compared to 4 TA trials (contrast = 0.022, $p = .045$). Additionally, 5 TA trials produced significantly lower overlap scores compared to both 3 TA trials (contrast = -0.109, $p < .001$) and 4 TA trials (contrast = -0.087, $p < .001$). In other words, as the number of TAs increased, target-selection decisions became more variable across HAs.

7.3.4 Participants on Binary Trace Overlap

Finally, differences in participants' binary trace overlap scores were analyzed using a 3 (AA-Coherder) × 3 (TA-Num) mixed-design ANOVA. The assumptions of normality, homoscedasticity, and sphericity ($p = .454$) were satisfied. The ANOVA revealed no significant main effects for AA-Coherder, $F(2, 177) = 1.80$, $p = .168$, $\eta_p^2 = .020$, or for TA-Num, $F(2, 177) = 3.00$, $p = .052$, $\eta_p^2 = .033$, nor a significant interaction effect, $F(4, 177) = 1.31$, $p = .270$, $\eta_p^2 = .029$. To further assess these null effects, a Bayesian repeated-measures ANOVA was performed using JASP version 0.19, following [83]'s guidelines. Moderate evidence was found supporting the null AA-Coherder effect (BF_{10} = 0.255) and anecdotal evidence for the null interaction effect (BF_{10} = 0.392). There was

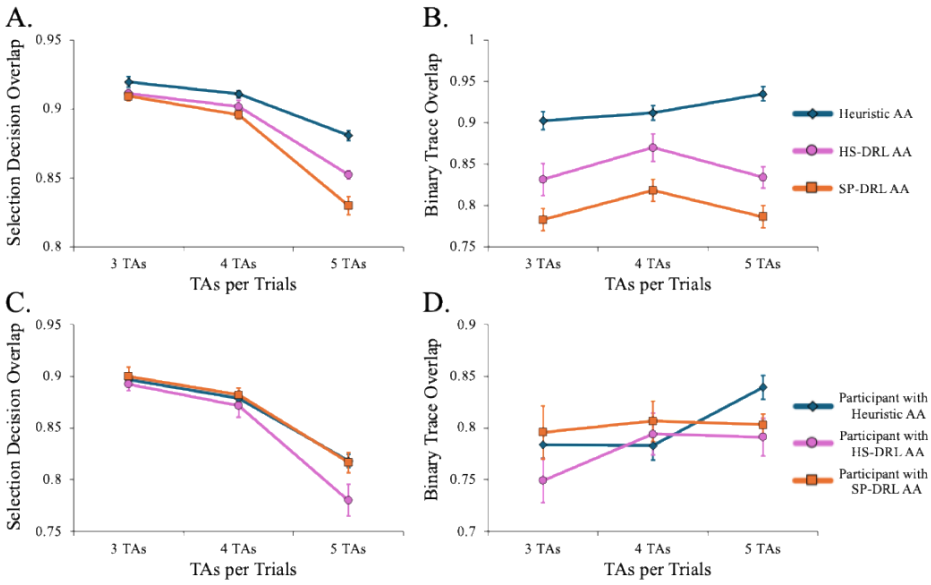


Figure 7.1: Line Graphs depicting average AA and Participant decision and trajectory overlap scores. Panel A and Panel B depict the decision and trajectory overlap scores for the different AAs, for 3 TA, 4 TA, and 5 TA trial conditions, respectively. Panel C and Panel D shows the decision and trajectory overlap scores for participants as a function of TA and coherder condition. The error bars represent ±1 standard error.

anecdotal evidence against the null TA-Num effect (BF_{10} = 2.408). No further analyses were conducted.

7.4 Conclusion

The present study investigated the effects of heuristic, HS-DRL, and SP-DRL target-selection policies in AAs collaborating with humans during a first-person shepherding task. The results demonstrated that the Heuristic AA, guided by the SCA policy, most accurately captured human behaviour (refer to figures in Appendix D for a 2D representation of human-human team and human-AA team trajectories with the AA being the heuristic agent), validating the effectiveness of human-derived heuristics for DPMP control. Additionally, participants displayed strong adaptability across the AA conditions, suggesting that DRL-based policies can achieve comparable HMI. These findings emphasise the potential of integrating heuristics within a dynamical systems framework to create AAs that exhibit human-like behaviours and strategies, with promising applications for synthetic coactors in real-world human-AA teams. Future research should investigate alternative machine learning models and consider increasing task complexity to further refine AA design. As human-AA interactions become more common in daily life, optimising AAs and their dynamic behaviour will be crucial for ensuring effective and seamless multi-agent collaboration in increasingly complex environments.

8 Conclusions and future work

This thesis explored the dynamics of human decision-making and movement in first-person herding tasks, combining human data collection experiments, computational modelling, and machine learning techniques to tackle the inherent complexities of multi-agent coordination. This research provided new insights into how humans control target agents and manage their movement within defined spaces. It also led to the development of artificial agents (AAs) capable of collaboratively completing the first-person herding task alongside human agents.

The initial chapters framed the herding problem within a broader context, highlighting its relevance in multi-agent systems and human-machine interaction. In particular, prior work [22, 23, 24, 25] had addressed importing herding models into the control architecture of robots for autonomous target agent control, often in complex (dynamic or cluttered) environments. This background on herding dynamics laid the groundwork for understanding how multi-agent herders (humans and machine) navigate and coordinate in environments where targets need to be contained efficiently.

The experimental work with human participants was a central focus. The studies revealed distinct patterns of behaviour in how individuals approached and managed target agents. Specifically, participants' navigational trajectories often involved two key phases: an approach phase, where they aligned themselves with the target, and a corral phase, where they influenced the target's movement toward a desired location. These findings were consistent across different task scenarios, demonstrating the robustness (or consistency) of human strategies in first-person herding tasks. A Dynamical Perceptual-Motor Primitive (DPMP) model was derived by modifying and fine-tuning an existing human navigational model [5]. By modelling how participants approached and corralled targets, the simulations mirrored human performance.

One of the key findings of the thesis was the effectiveness of simple strategies in modelling human decision making. As with [16, 33], it was demonstrated in this thesis that human-inspired models incorporating rules-based (heuristic) decision policies into DPMP models worked best in reproducing human behaviour. This was seen to easily generalise to multi-person, and then, multi-agent herding. The most appropriate heuristic decision model was itself derived from a detailed study of the single-herder, multiple-target scenario, with further validation through follow-up experiments in that same context.

This thesis also explored whether heuristic-based decision models, when integrated

into artificial agents, could replicate human-like teamwork in human-agent interactions. Similar to work conducted in simpler environments (such as [44, 45]), computational models like those based on reinforcement learning are generally more adaptable than heuristic-based models to new or unfamiliar environments. However, as had been demonstrated in [46, 44] and in this thesis, in multi-agent herding tasks where human and artificial agents collaborate, Deep Reinforcement Learning (DRL) agents were less effective at replicating human-like behaviours than heuristic models. This suggests that, at least in our representative task context, simpler strategies may better capture complex human behaviours.

8.1 Limitations

However, several challenges and potential limitations were identified. One significant shortcoming was the inability of the models to capture more nuanced aspects of human movement, such as the start-and-stop behaviours observed in some participants. Another limitation of our models is the lack of ability to model different humans' decisions in under the same experimental conditions (or trial). This distribution of human decisions could be addressed through a stochastic (or random) decision model and had not been detailed here. Another major challenge of this work is to implement the models into real-world (or physical) robots, an avenue that had not been tested but would have significant applications-based implications.

8.2 Future work

Future work in this domain can, more immediately, include studying the Desert Herding task [66, 65], which incorporate more complex TA dynamics and a much larger (500m x 500m) playing field. Other target selection policy modelling techniques can also be explored, such as using Supervised Machine Learning techniques, such as was performed in [55]. In this work [55], the authors had employed Long Short-Term Memory (LSTM) networks to predict the target selection decisions of human herders in the tabletop herding task (section 2.2.3), alongside using explainable AI to identify what features most influenced the decision model's predictions. Additional next steps could include training two Deep Reinforcement Learning Neural Networks - one for the navigational dynamics (hence replacing the navigational DPMP model) the other for decision dynamics (similar to what had been performed in this thesis) and combining the two to generate a model-free artificial agent, for comparison with the other AAs developed in this thesis.

However, on a more general scale, the importance of promoting the task to real-like-like scenarios cannot be understated. Work can be done in the first instance in implementing the models into controlled robotic environments (for e.g., [84]), and then in uncontrolled real-world environments. It will be very promising to implement these algorithms and models into the design architecture of real robots collaborating alongside humans, in search-and-rescue operations and in general, collaborative perceptual-motor tasks.

8.3 Conclusion

In conclusion, this thesis provided substantial contributions to the understanding of human decision-making and control strategies in first-person herding tasks. The experimental findings, coupled with the development of computational models, advance the field of multi-agent coordination and offer new possibilities for improving human-machine collaboration. While there are areas for further refinement, the insights gained from this research lay a strong foundation for future exploration in both theoretical modelling and practical applications of human-agent systems.

8.4 List of publications

Published:
bin Kamruddin, A., Sandison, H., Patil, G., Musolesi, M., di Bernardo, M., & Richardson, M. J. (2024). Modelling human navigation and decision dynamics in a first-person herding task. *Royal Society Open Science*, 11(10), 231919.

Under review:
bin Kamruddin, A., Lam, C., Ghanem, S., Patil, G., Musolesi, M., di Bernardo, M., & Richardson, M. J. Modeling and Reproducing Cooperative Human Behavior in a Complex First-Person Herding Task. *Journal of Experimental Psychology - Human Perception and Performance*

bin Kamruddin, A., Lam, C., Patil, G., Musolesi, M., di Bernardo, M., & Richardson, M. J. Human-Machine Teaming through Deep Reinforcement Learning in a Complex First-Person Herding Task. *IEEE Transactions on Human-Machine Systems*

Simpson, J., **bin Kamruddin, A.**, Nalepka, P., Richardson, M. J. Can an AI Agent Lead a Human Team? *Computers in Human Behavior: Artificial Humans*

Crone, C., **bin Kamruddin, A.**, Nalepka, P., Richardson, M. J. Assessing Team Performance Dynamics during a Competitive Team Multiplayer Video Game. *Scientific Reports - Nature*

Conference Proceedings:
bin Kamruddin, A., Patil, G., Musolesi, M., di Bernardo, M., & Richardson, M. J. (2023). Modeling Human Navigation in First-Person Herding Tasks. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 45, No. 45).

Appendix A

Error metrics and parametrisation algorithm used

A.1 Introduction and Motivation

The distance between two points on a plane can be calculated based on the chosen metric, such as the Euclidean or curvilinear metric for flat or curved spaces, respectively. Distances between two trajectories in flat spaces may be calculated naively by summing the pairwise Euclidean distances for each corresponding point in the two trajectories. However, problems arise when the trajectories are sampled at different frequencies or have different lengths. In this document, we consider various distance measures between two trajectories.

Once a distance measure is chosen, minimization algorithms can be used with this distance measure as a loss function to solve the optimization problem of ensuring that the simulated model trajectory is as close to the real trajectory as possible.

A.2 Hausdorff Distance

Named after Felix Hausdorff (1868-1942), the Hausdorff distance $H(A, B)$ between two finite sets of points $A = \{a_1, a_2, \ldots, a_N\}$ and $B = \{b_1, b_2, \ldots, b_M\}$ is defined as in [85] and [86]:

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{A.1}$$

where

$$h(A, B) = \max_{a \in A} \min_{b \in B} d(a, b) \tag{A.2}$$

with $d(\cdot, \cdot)$ representing the Euclidean distance between two points.

The Hausdorff distance measures the degree of dissimilarity between two trajectories as the larger of the maxima of the closest distances between the trajectories, considering each trajectory in turn.

A limitation of this distance measure is that it is unordered—it does not take into account the temporal evolution of trajectories. Thus, a modified distance, the Fréchet distance, was developed, inspired by the Hausdorff distance.

A.3 Fréchet Distance

The Fréchet distance $F(A, B)$ for A and B as above is defined in [87] as:

$$F(A, B) = \inf_{\alpha,\beta} \max_{t\in[0,1]} d(A(\alpha(t)), B(\beta(t))) \tag{A.3}$$

where d is defined as before, and α and β are reparametrizations of A and B, respectively. The Fréchet distance can be thought of as the minimum leash length needed to connect pairwise points that co-move along the parameter t across the two trajectories.

The limitation of both the Hausdorff and Fréchet distances is that they focus only on the extreme distances between the trajectories. In other words, they consider a single maximum distance and do not take into account all pairwise distances between trajectories. To capture all dissimilarities between two trajectories more comprehensively, we consider a different similarity measure: the dynamic time-warped distance.

A.4 Dynamic Time-Warped Distance

The Dynamic Time-Warped (DTW) distance, as presented in [88] and [89], provides a way to measure the similarity between two time series sampled at different time intervals or frequencies. This algorithm takes as input the sets of points A and B as above and constructs the distance matrix $C \in \mathbb{R}^{N\times M}$, which represents all pairwise distances between A and B. It finds the *alignment path* that runs through the lowest distances in the matrix, aligning the differently sampled trajectories to minimize the running Euclidean distance across corresponding points. The output $DTW(A, B)$ is the sum of distances along this alignment path.

Formally, the matrix C has elements

$$C_{i,j} = \|a_i - b_j\|, \quad i \in [1 : N],\ j \in [1 : M] \tag{A.4}$$

The alignment path p is a sequence of indices $p_1, p_2, \ldots, p_K$ with $p_l = (p_i, p_j) \in [1 : N] \times [1 : M]$ for $l = [1 : K]$ that satisfies certain conditions:

1. Boundary: $p_1 = (1, 1)$ and $p_K = (N, M)$
2. Monotonicity: The indices $p_{i,j}$ are strictly non-decreasing, where $p_k = (p_i, p_j)$, preserving the time-ordering of the points.
3. Step size: Indices do not jump by more than one.

The indices $p_{i,j}$ correspond to the trajectory points, allowing for repetitions. These repetitions are key to DTW, as they enable the alignment of paths sampled at different frequencies. The DTW distance is calculated as the minimal distance along these indices:

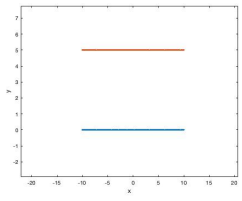
$$DTW(A, B) = \min_{p} c_p(A, B) \tag{A.5}$$

with

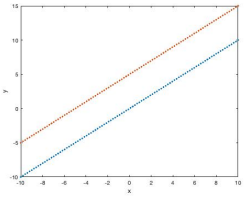
$$c_p(A, B) = \sum_{k=1}^{K} C_{p_k} \tag{A.6}$$

A.5 Test Cases

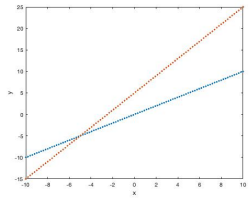
In this section, we generate different curves and evaluate how the metrics discussed above perform. Figure A.1 shows three sets of lines, and Figure A.2 shows two sets of curves. The distances between the y-intercepts of the parallel lines are 5, and the difference in radii of the concentric circles is 2.



(a) Two parallel lines with zero slope

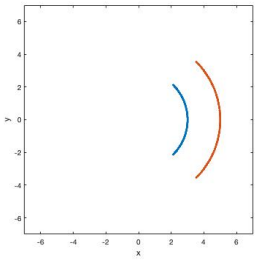


(b) Two parallel lines with non-zero slope

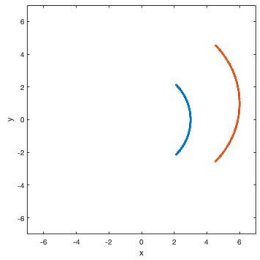


(c) Two non-parallel lines

Figure A.1: Simple lines



(a) Two concentric circular arcs



(b) Two non-concentric circular arcs

Figure A.2: Simple curves

Table A.1 reports the values for the different similarity measures applied to the various test cases.

Description	$H(\cdot,\cdot)$	$F(\cdot,\cdot)$	$DTW(\cdot,\cdot)$
Two parallel lines with zero slope	5	5	500
Two parallel lines with non-zero slope	5	5	408.76
Two non-parallel lines	15	15	533.28
Two concentric circular arcs	2	2	200
Two non-concentric circular arcs	3.41	3.41	308.33

Table A.1: Table of distance measures

We conclude that DTW is the most informative measure, as it decreases when the curves move closer to each other. In the case of parallel lines with non-zero slope, the Hausdorff and Fréchet distances do not exhibit the expected behavior.

A.6 Parametrization Technique - SLSQP

Our parametrization algorithm, Sequential Least Squares Quadratic Programming (SLSQP), was introduced in 1988 by Dieter Kraft ([90], [91]). The problem statement is as follows:

Consider minimizing a function $f(x)$ in n dimensions:

$$\min_{x \in \mathbb{R}} f(x) \tag{A.7}$$

subject to the constraints

$$g_j(x) = 0, \quad j = 1, \ldots, m_e \tag{A.8}$$

$$g_j(x) \geq 0, \quad j = m_{e+1}, \ldots, m \tag{A.9}$$

$$x_l \leq x \leq x_u \tag{A.10}$$

where

$$f : \mathbb{R}^n \to \mathbb{R} \tag{A.11}$$

and

$$g : \mathbb{R}^n \to \mathbb{R}^m \tag{A.12}$$

are both continuous and differentiable.

The algorithm is described as follows: Given an initial set of parameters $x^{(0)}$, the $(k+1)$-th iteration produces the following set of parameters:

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)} \tag{A.13}$$

where $\alpha^{(k)}$ is the step size at iteration k, and $d^{(k)}$ is the corresponding search direction. The search direction is determined using quadratic programming. First, we define the Lagrangian:

$$L(x, \lambda) = f(x) - \sum_{j=1}^{m} \lambda_j g_j(x) \tag{A.14}$$

with the Lagrange multipliers λ. The quadratic programming (QP) algorithm then seeks to minimize:

$$\text{QP}: \quad \min_{d \in \mathbb{R}^n} \frac{1}{2} d^T B^{(k)} d + \nabla f(x^{(k)}) d \tag{A.15}$$

subject to the constraints:

$$\nabla g_j(x^{(k)}) d + g_j(x^{(k)}) = 0, \quad j = 1, \ldots, m_e \tag{A.16}$$

and

$$\nabla g_j(x^{(k)}) d + g_j(x^{(k)}) \geq 0, \quad j = m_e + 1, \ldots, m \tag{A.17}$$

where

$$B := \nabla^2_{xx} L(x^{(k)}, \lambda) \tag{A.18}$$

In simple terms, with the appropriate step size, the algorithm moves in the parameter space to minimize the Lagrangian—i.e., it minimizes the function f subject to constraints while considering changes up to second-order, hence the name quadratic programming.

Appendix B

Single herder - multiple targets

Figures B.1, B.2, and B.3 include trajectory, weighted, and binary trace maps for all trials used in the Multi-Target Herding Experiment. The panels on the left display the human trajectories (in black), the initial positions of the TAs (in green), the initial positions of the HAs (in yellow), and the simulation (in red). The middle panels show the artificial HA simulation overlaid on the nonlinear heatmap generated by binning the spatial frequencies of human trajectories. The nonlinear heatmap's colours represent relative spatial frequency, with yellow indicating higher frequencies (hotter) and blue indicating lower frequencies (colder). The panels on the right show the simulation overlaid on the binary map extracted from the nonlinear heatmap by filtering values higher than a specified threshold.

As detailed in the main text, to further validate that the Successive Collinear Angle (SCA) policy best captured the participants' target selection decisions in the Multi-Target Herding Experiment, a second set of initial conditions were designed and tested. To test the successive collinear angle versus the successive collinear distance, the initial conditions in Figure B.4 were used. To test whether TA selection was best defined by the initial location (starting position) of the HA or as a successive function of the angular distance of the TA, the initial conditions in Figure B.5 were used. Finally, Figure B.6 corresponds to the initial conditions used to test whether the participants used collinear, cluster identification or not. The initial position of the HA is in yellow and that of the TA is in green. TAs are labelled A, B, and C, and their distances from the HA, as well as the angle from the HA subtended at the centre of the containment zone, are denoted.

The link to the online repository containing the simulation and analysis code, as well as the experimental data collection and simulation builds can be found through the link here.

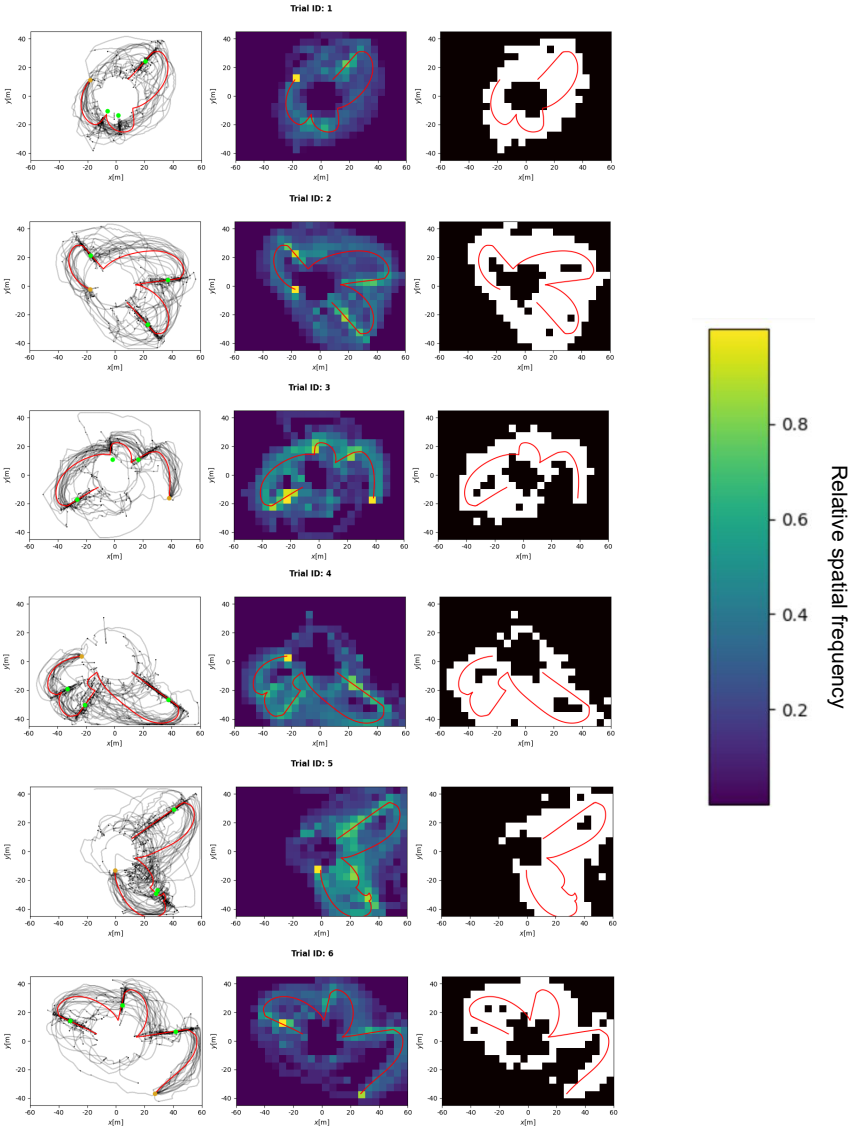


Figure B.1: Multi-Target Herding Experiment: Trials 1 through 6

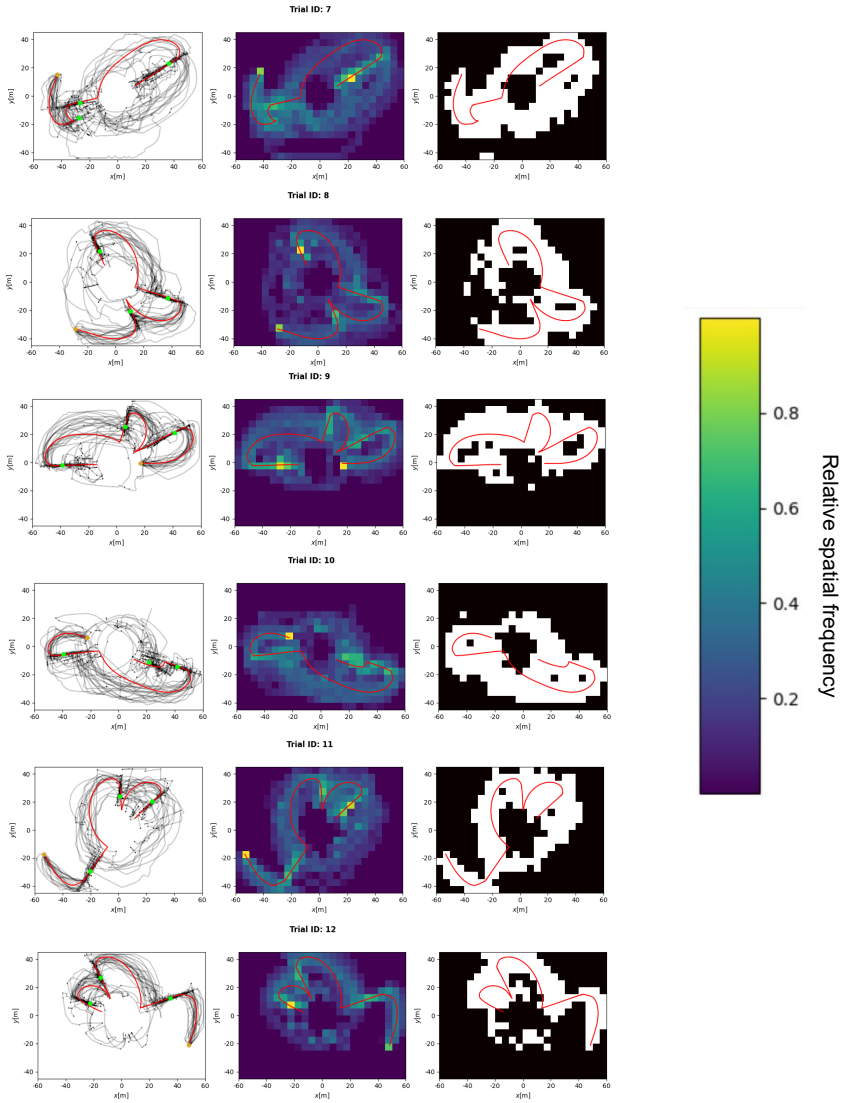


Figure B.2: Multi-Target Herding Experiment: Trials 7 through 12

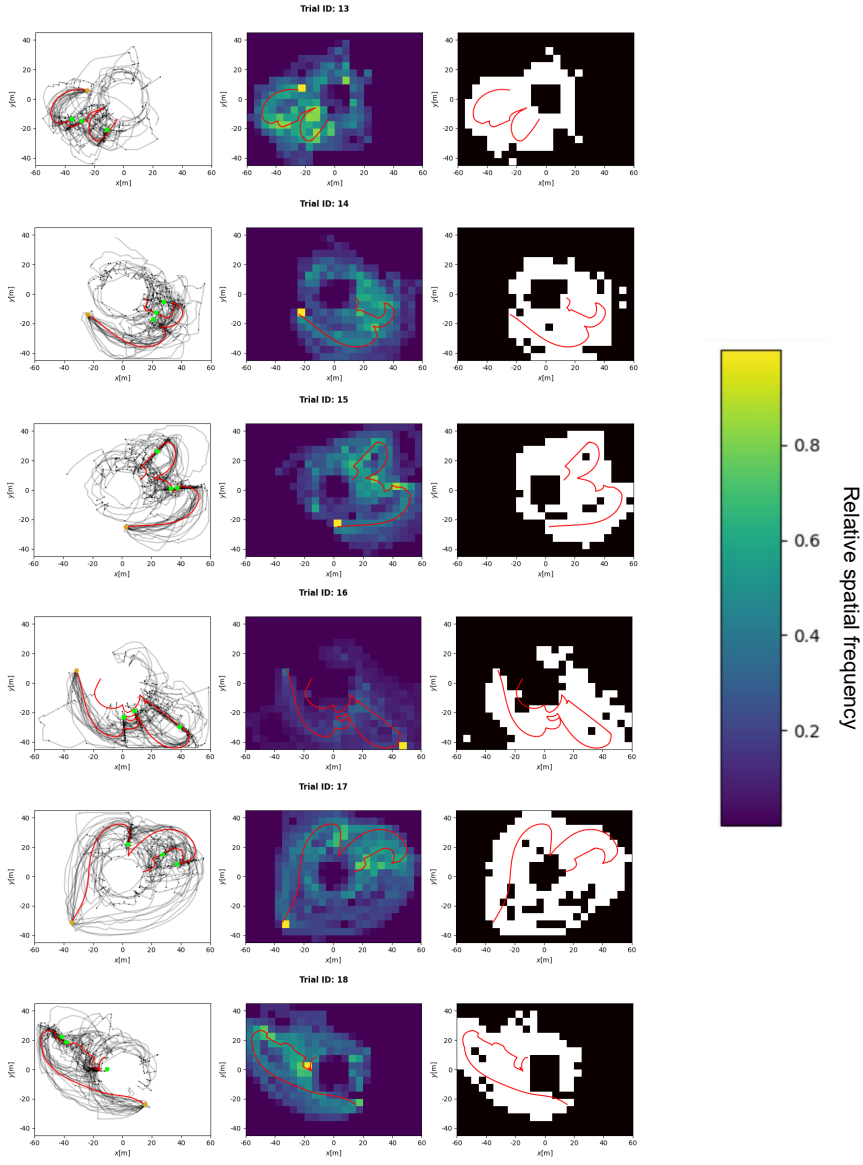


Figure B.3: Multi-Target Herding Experiment: Trials 13 through 18

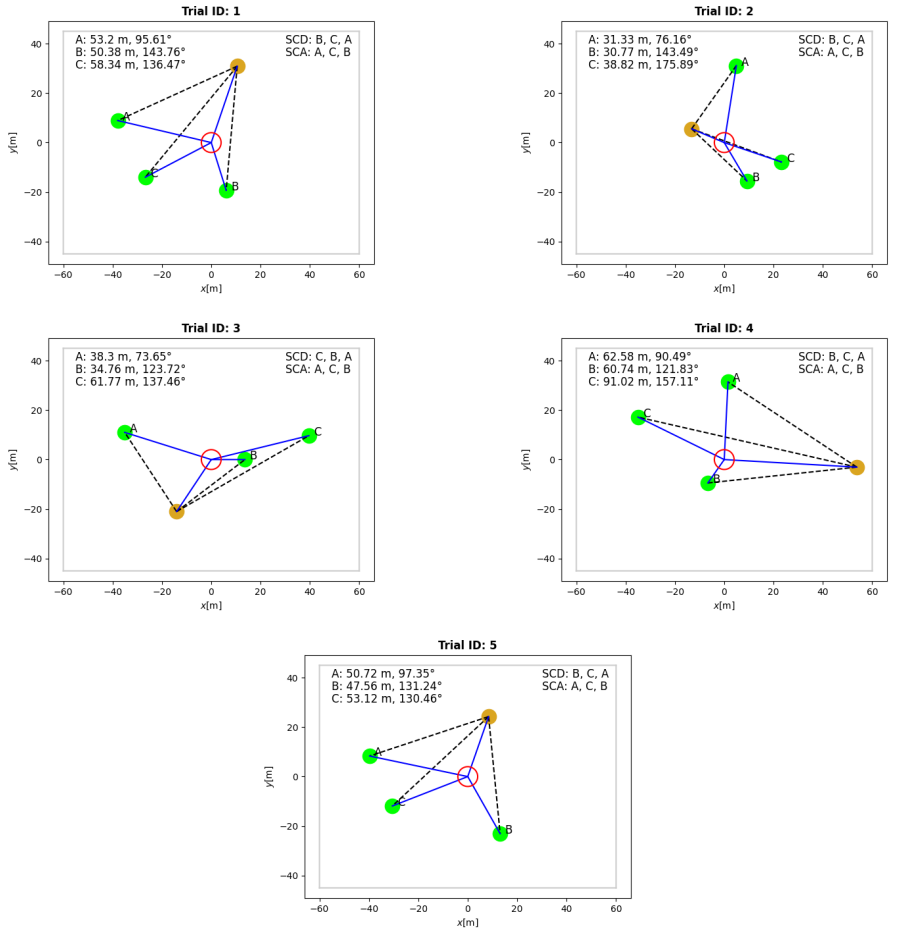


Figure B.4: Multi-Target Herding Experiment Validation trials in which successive collinear distance (SCD) vs successive collinear angle (SCA) was tested. SCD had a score of 22% and SCA, 78% (scores not mutually exclusive). Distances and angles between the HA and the respective TA are also denoted (with angles measured from the centre of the containment zone), along with the expected prediction of SCD vs. SCA.

.

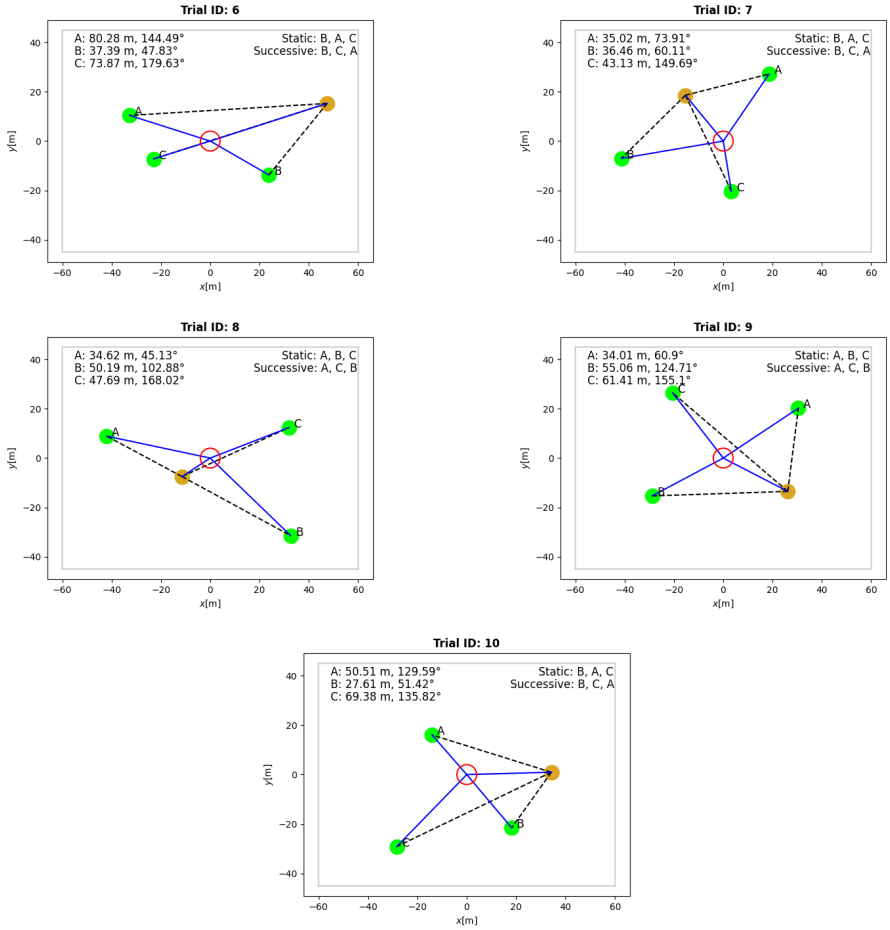


Figure B.5: Multi-Target Herding Experiment Validation trials in which initial vs successive collinear angle was tested. The former had a score of 12% and the latter, 78% (scores nonmutually exclusive). The expected predictions of these two campaigns are also noted.

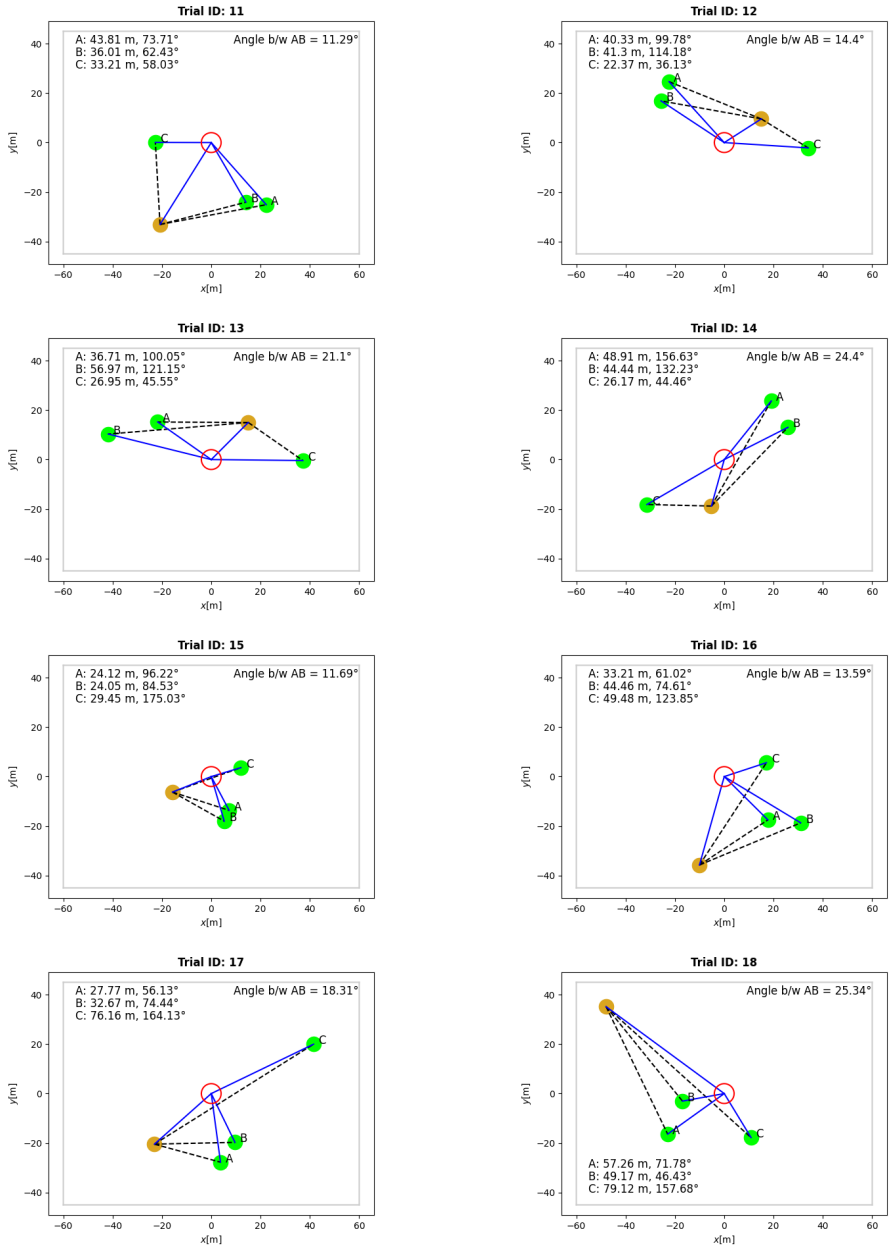
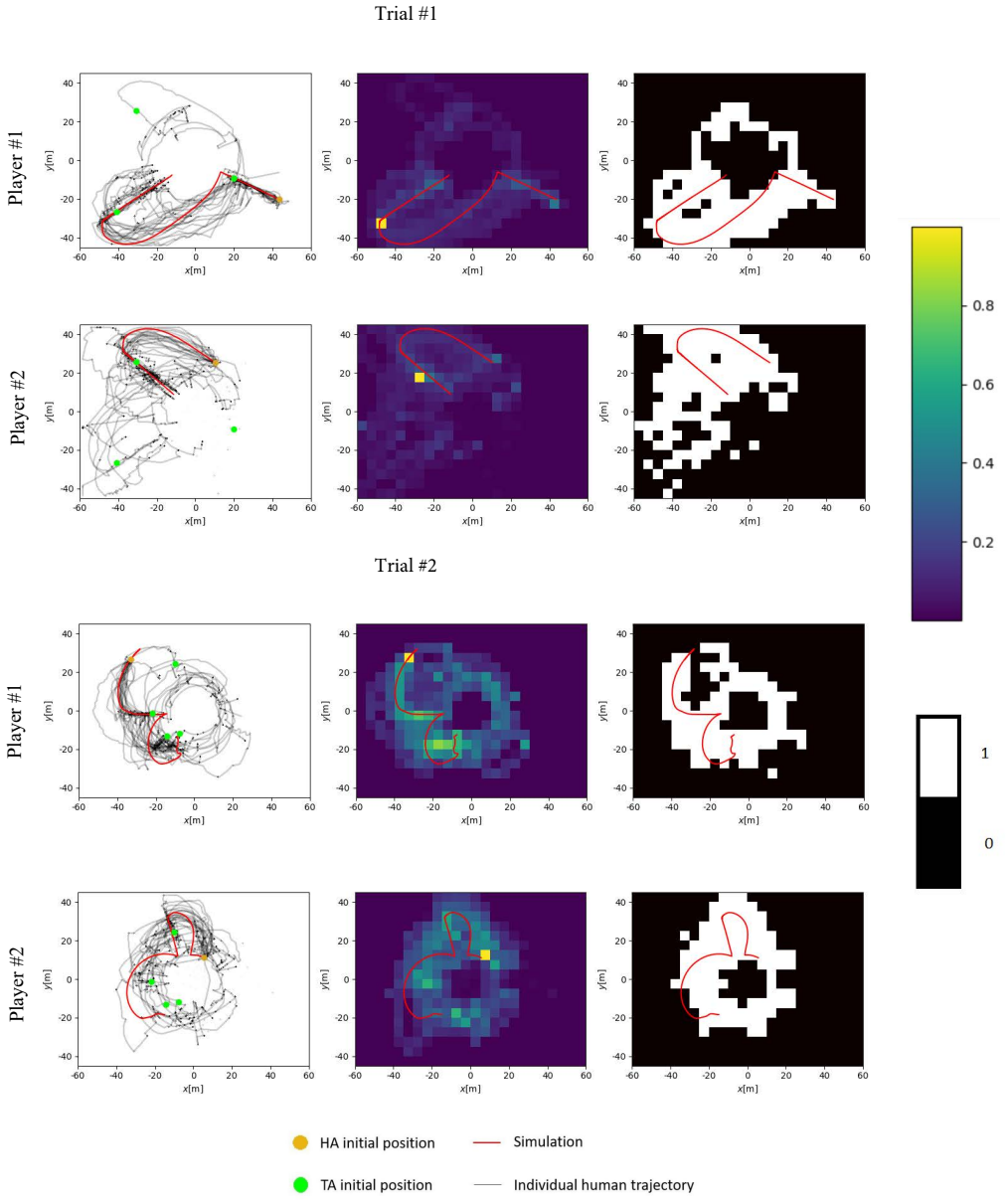


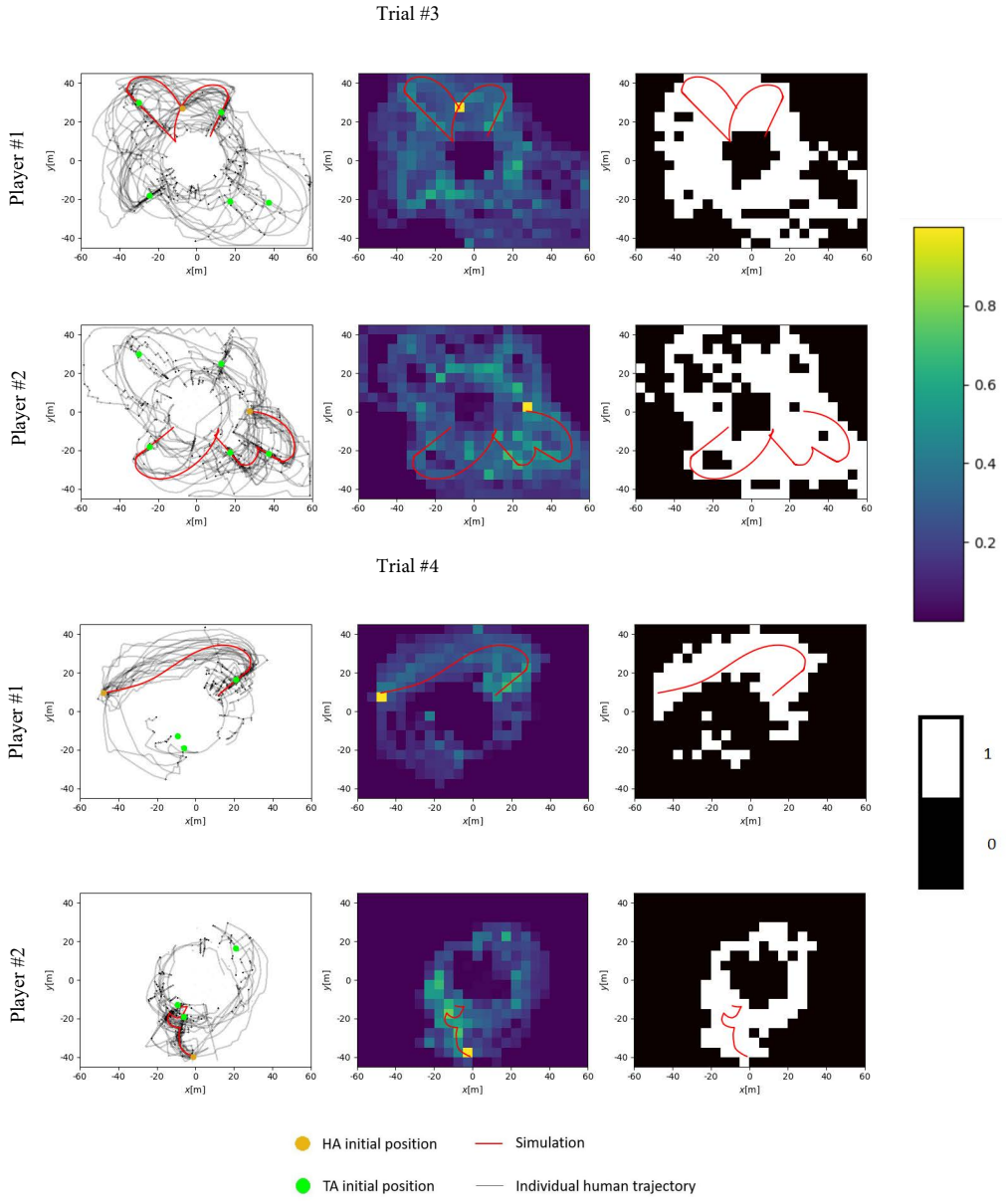
Figure B.6: Multi-Target Herding Experiment Validation trials in which successive angle versus successive collinear angle was tested. There was always a pair of TAs (A and B) with angles close to 10°, 15°, 20° and 25° (denoted in the figures). As varying participants had varying thresholds, an analysis of the scores has been presented in the main text, and the corresponding predictions have been omitted from the figures.

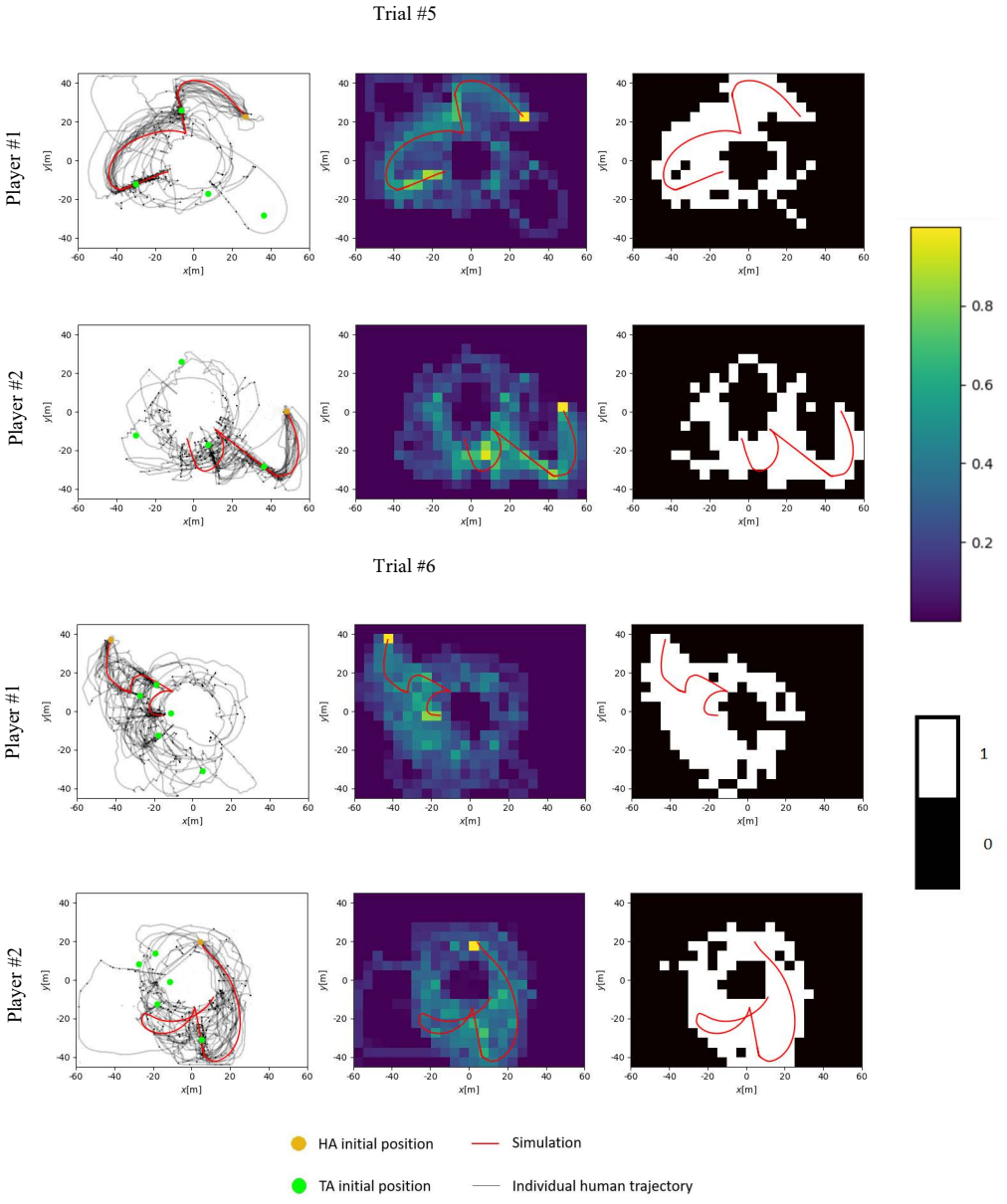
Appendix C

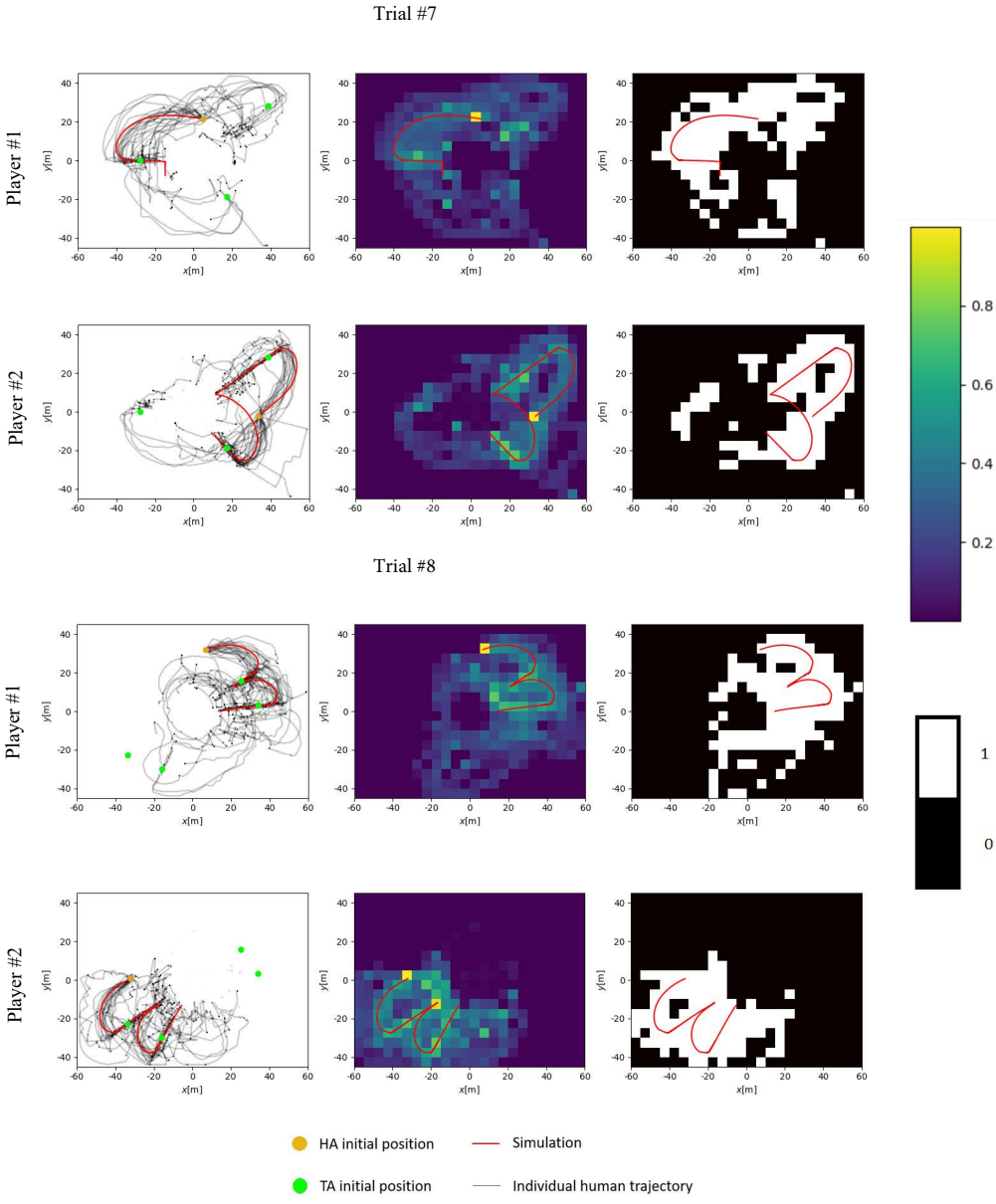
Multiple human herders - multiple targets

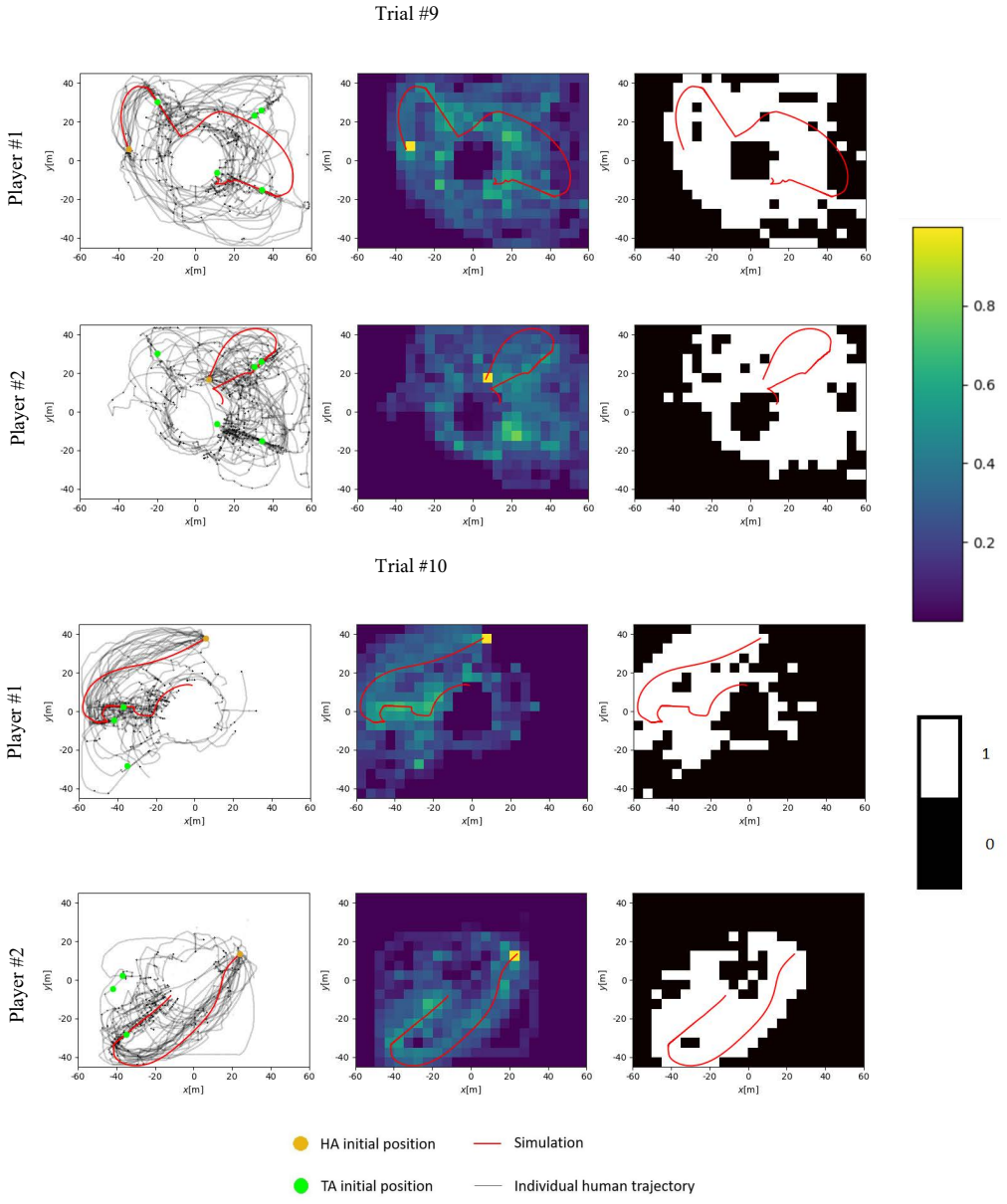
The figures in this Appendix represent the simulated trajectories overlaid on the human-human team data, alongside the simulations overlaid on the weighted and binary heatmaps. The panels on the left display the human trajectories (in black), the initial positions of the TAs (in green), the initial positions of the HAs (in yellow), and the simulation (in red). The middle panels show the artificial HA simulation overlaid on the nonlinear heatmap generated by binning the spatial frequencies of human trajectories. The nonlinear heatmap's colours represent relative spatial frequency, with yellow indicating higher frequencies (hotter) and blue indicating lower frequencies (colder). The panels on the right show the simulation overlaid on the binary map extracted from the nonlinear heatmap by filtering values higher than a specified threshold.

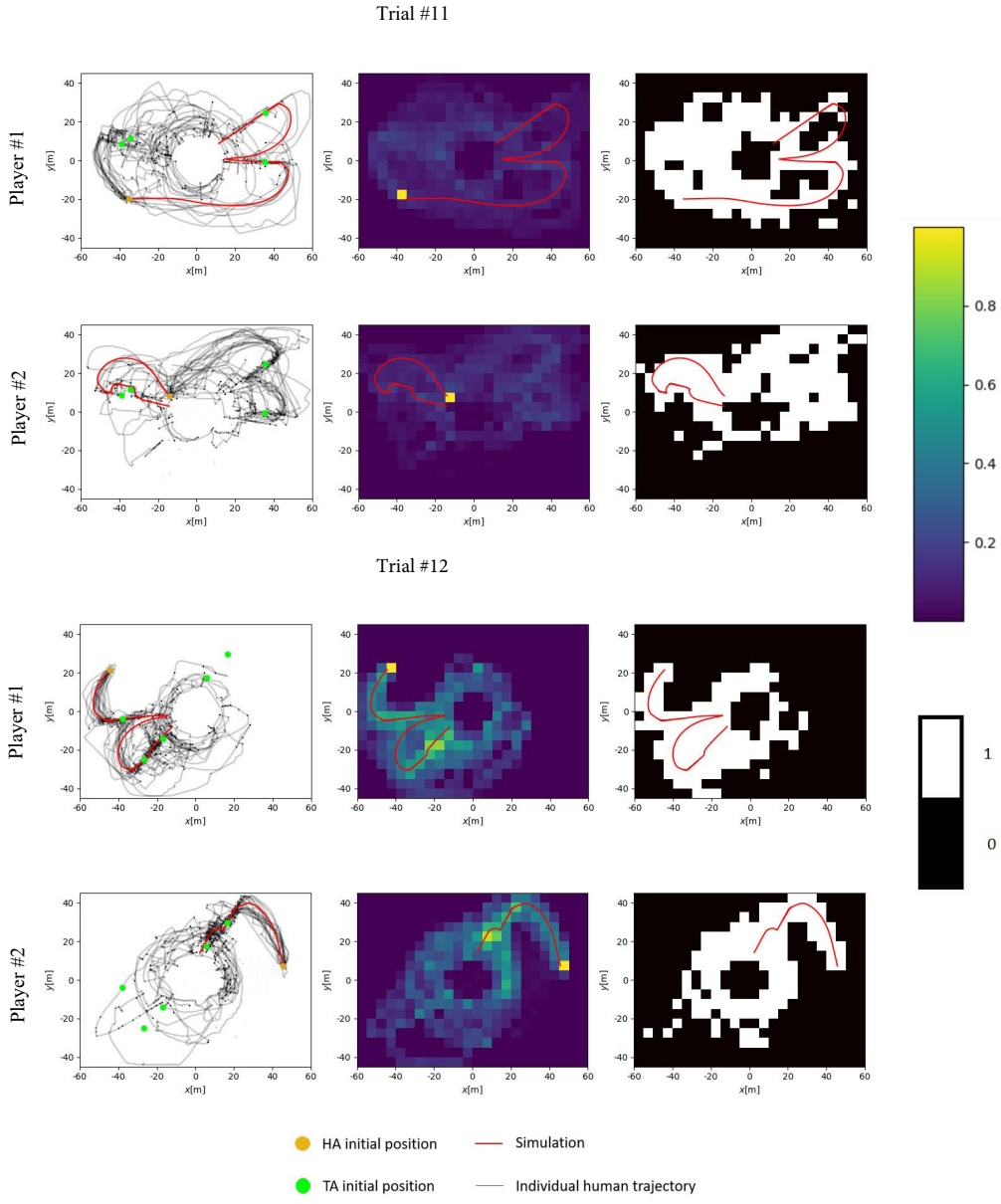

Trial #1
Player #1
Player #2
Trial #2
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory

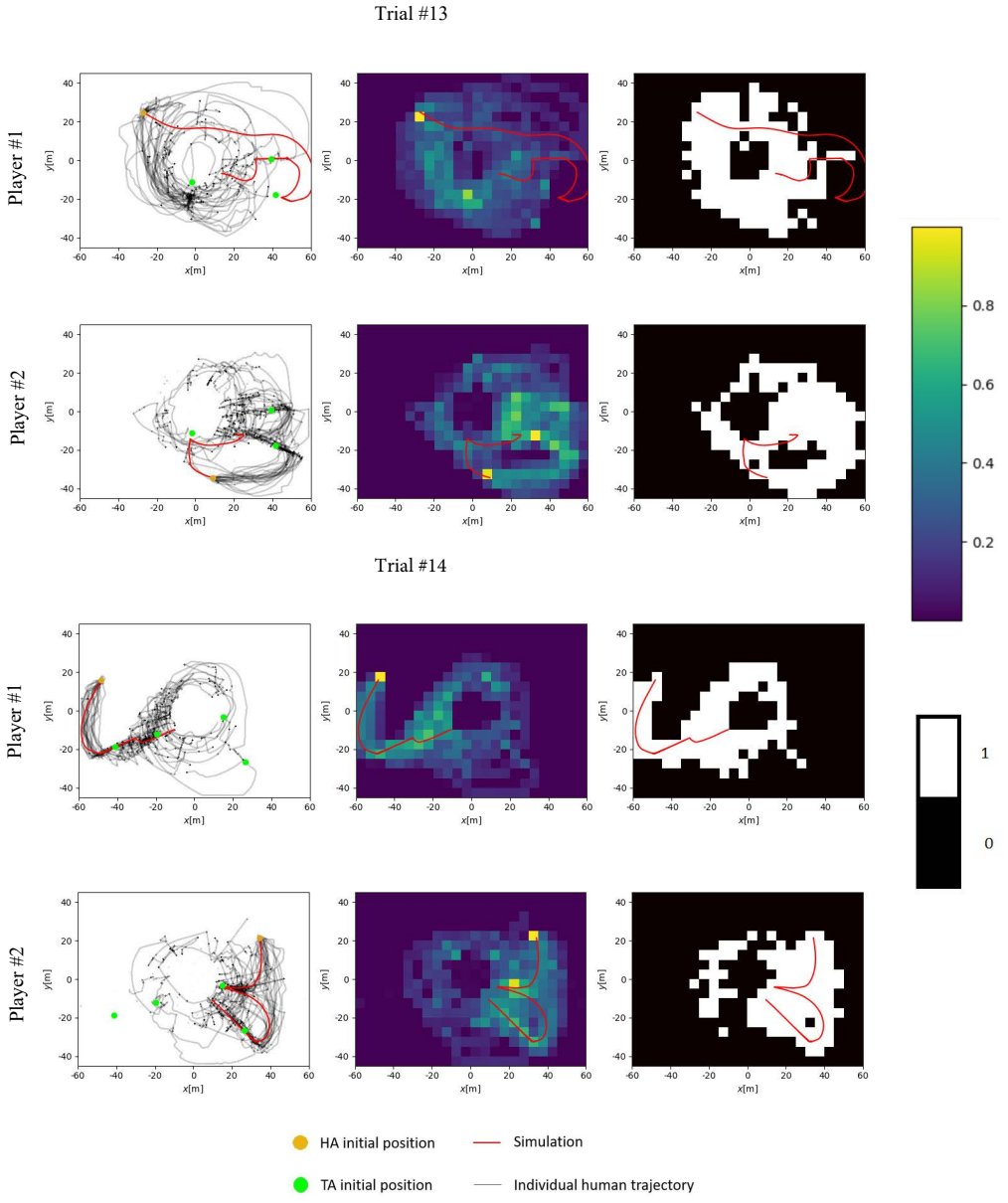

Trial #3
Player #1
Player #2
Trial #4
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory

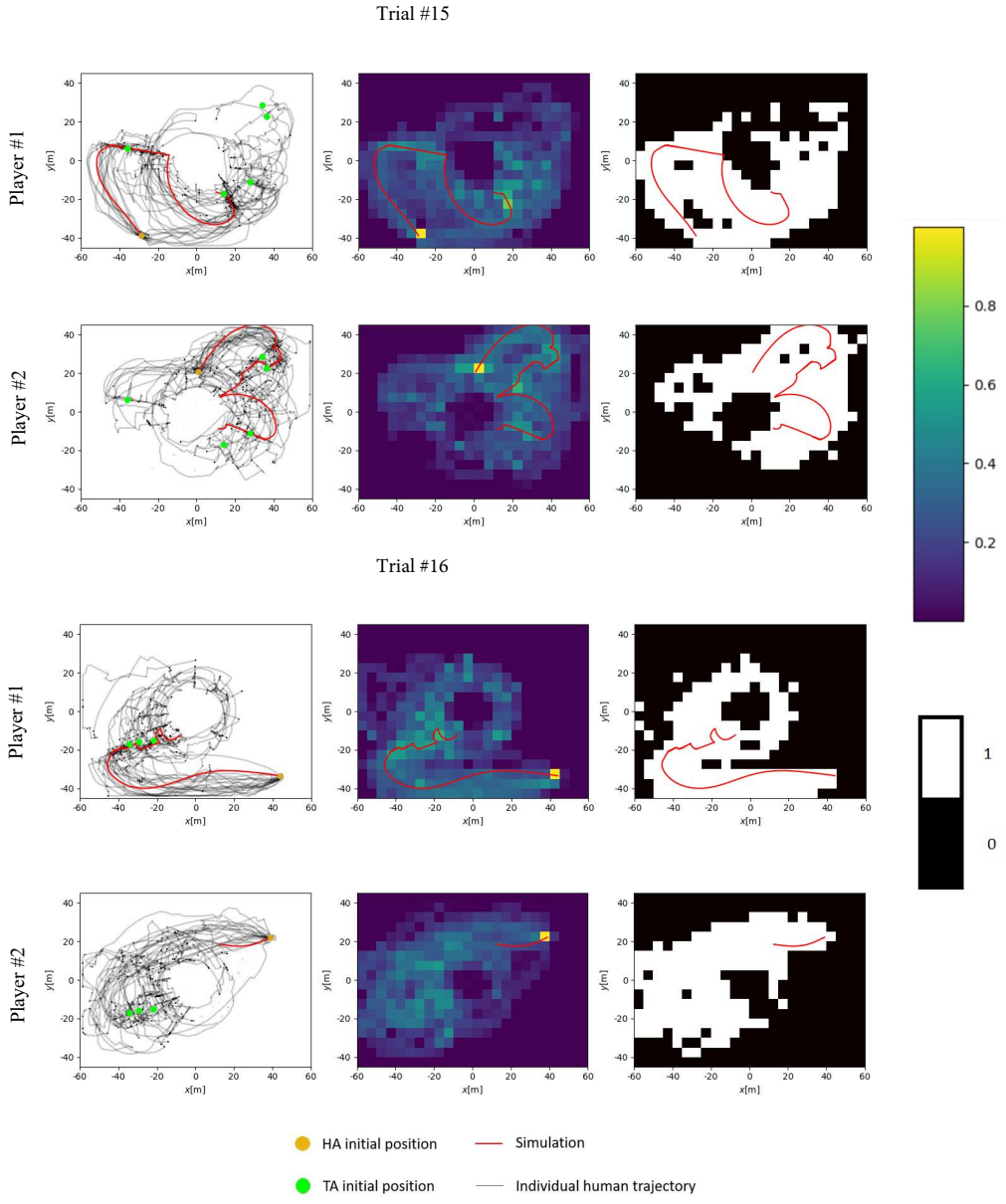

Trial #5
Player #1
Player #2
Trial #6
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory

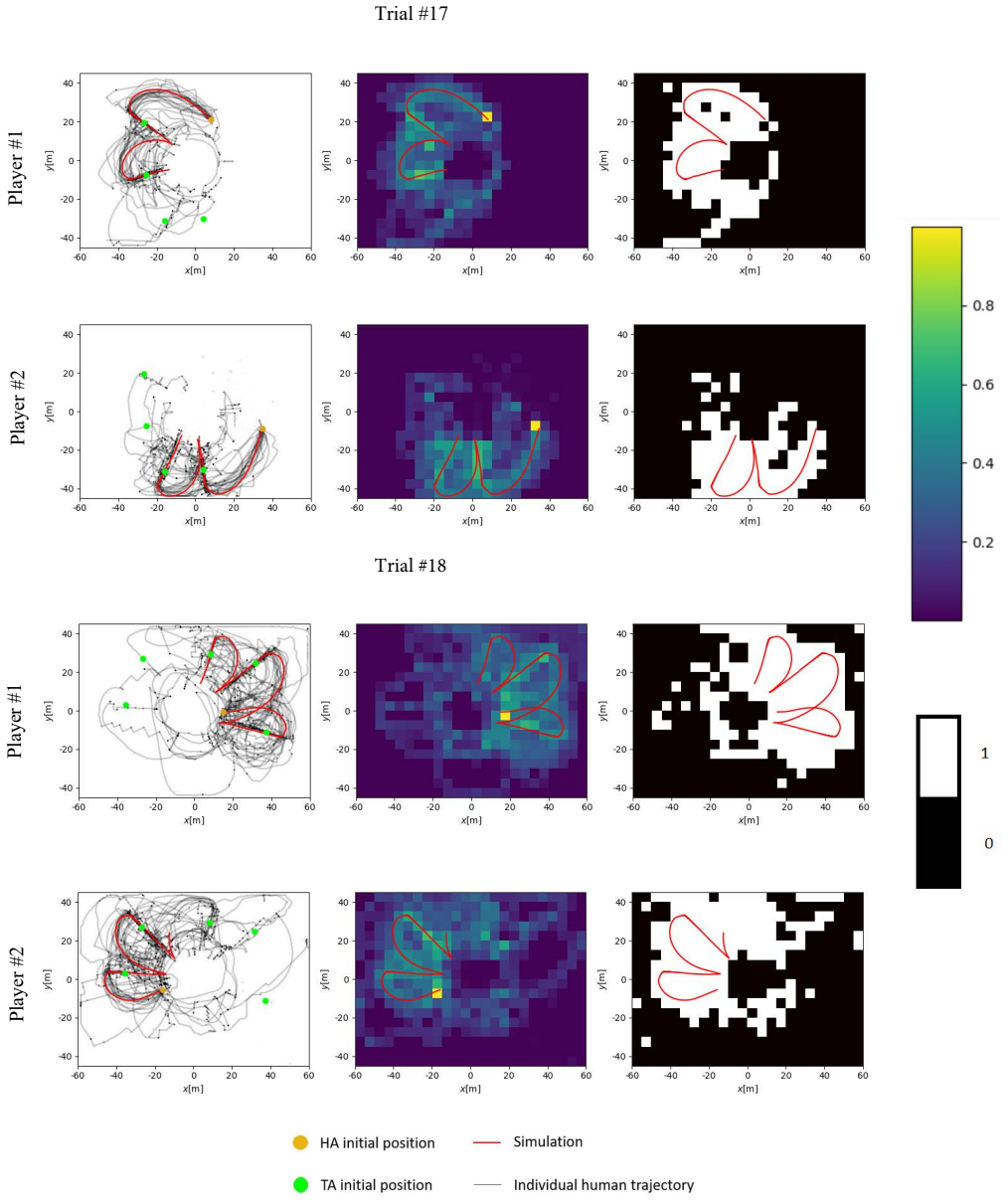

Trial #7
Player #1
Player #2
Trial #8
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory


Trial #9
Player #1
Player #2
Trial #10
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory


Trial #11
Player #1
Player #2
Trial #12
Player #1
Player #2
x[m]
y[m]
HA initial position
TA initial position
Simulation
Individual human trajectory


Trial #13
Player #1
Player #2
Trial #14
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory


Trial #15
Player #1
Player #2
Trial #16
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory


Trial #17
Player #1
Player #2
Trial #18
Player #1
Player #2
x[m]
y[m]
0.8
0.6
0.4
0.2
1
0
HA initial position
TA initial position
Simulation
Individual human trajectory

Appendix D

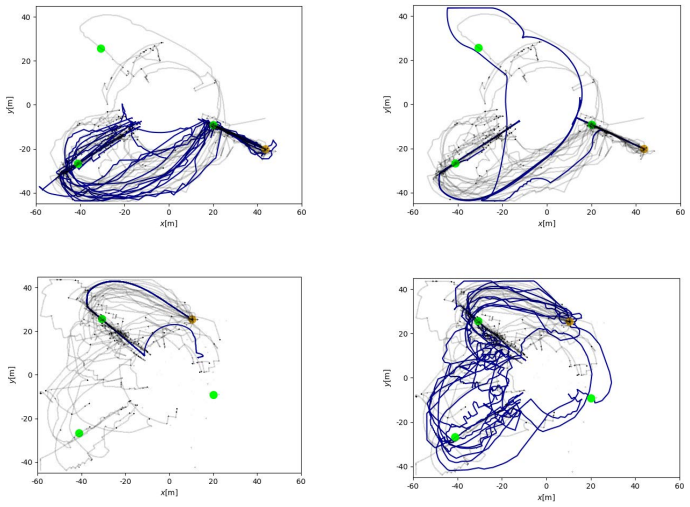
Human-AA team trajectories

The figures in this appendix represent the data from the human-human team (in black) and the data from the human-AA team, with the AA being the heuristic agent (in blue), As half of the participants were assigned to the initial conditions assigned to the AA in the other half, and vice versa, there are four panels per trial condition: player 1 human/AA and player 2 AA/human. Players 1 and 2 play simultaneously, collaborating with each other to complete the herding task. One may note the variability in the human data across both the human-human as well as the human-AA teams, while the AA picks up on the human trends. This has been quantified in the main text.

Trial #1

Player 1 is human

Player 1 is AA



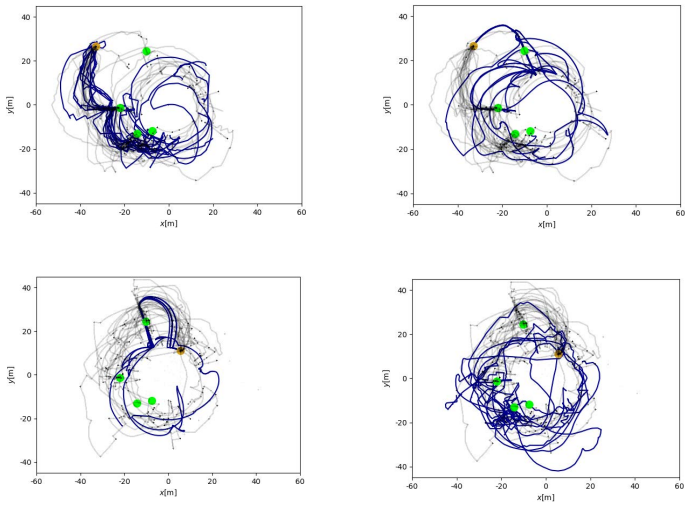
Player 2 is AA

Player 2 is human

Trial #2

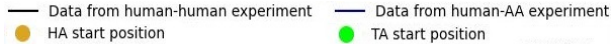
Player 1 is human

Player 1 is AA



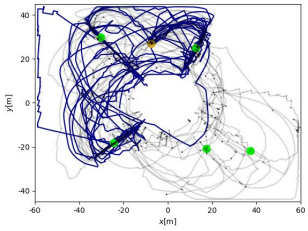
Player 2 is AA

Player 2 is human

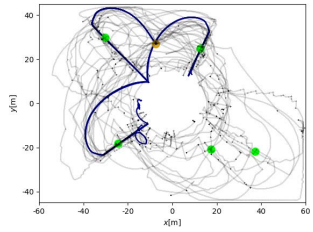


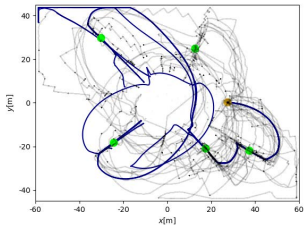
Trial #3

Player 1 is human

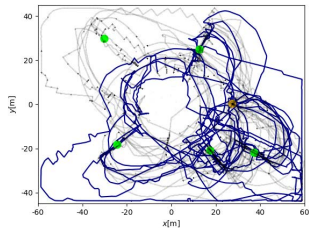


Player 1 is AA





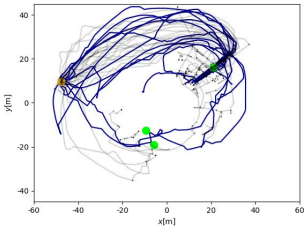
Player 2 is AA



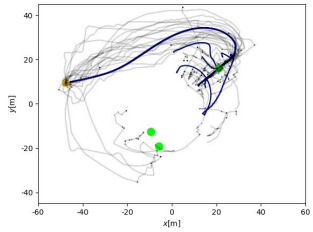
Player 2 is human

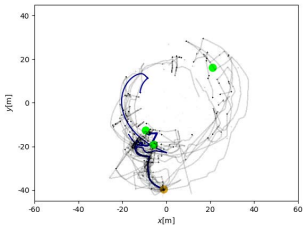
Trial #4

Player 1 is human

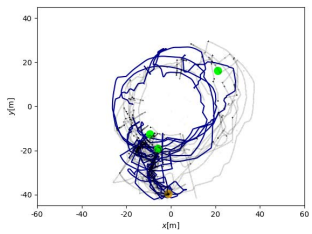


Player 1 is AA

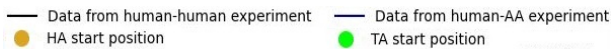




Player 2 is AA

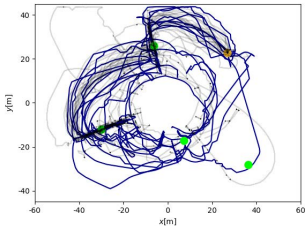


Player 2 is human

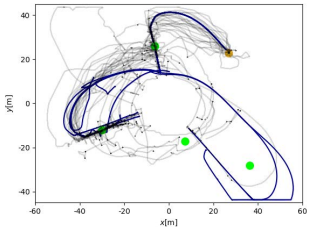


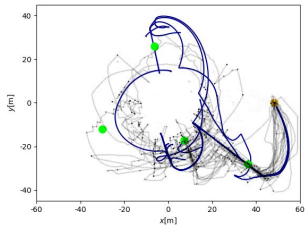
Trial #5

Player 1 is human

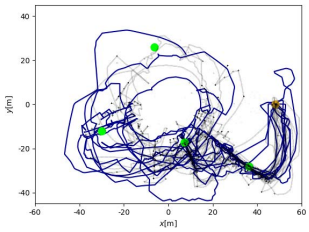


Player 1 is AA





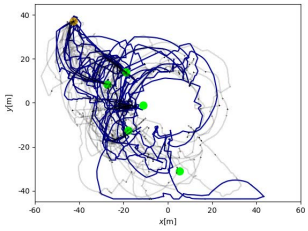
Player 2 is AA



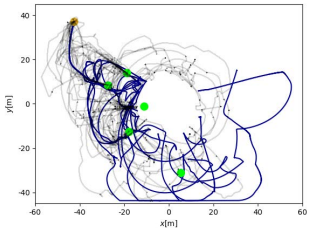
Player 2 is human

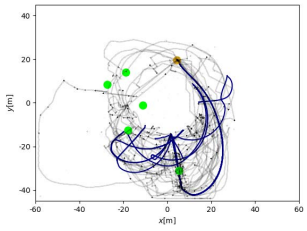
Trial #6

Player 1 is human

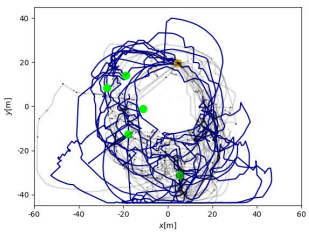


Player 1 is AA

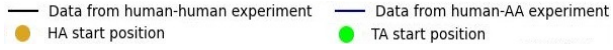




Player 2 is AA

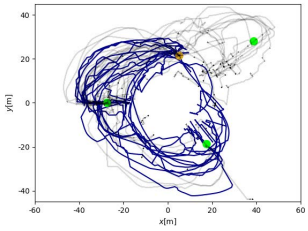


Player 2 is human

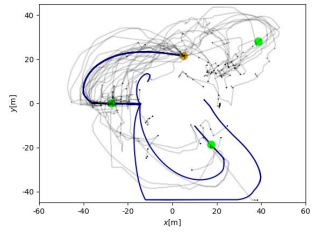


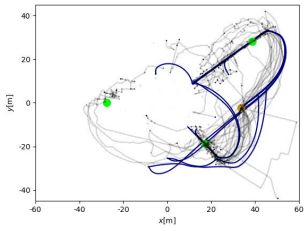
Trial #7

Player 1 is human

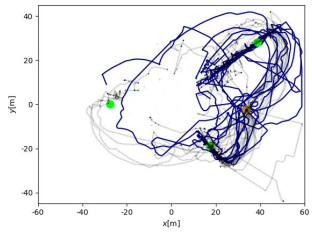


Player 1 is AA





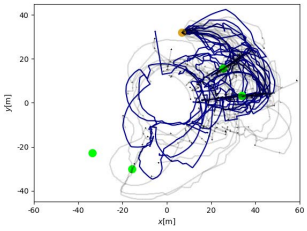
Player 2 is AA



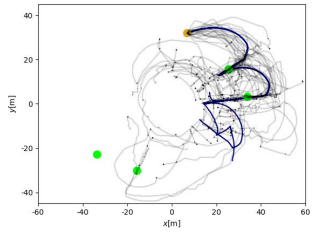
Player 2 is human

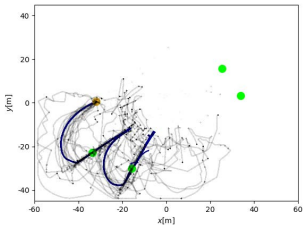
Trial #8

Player 1 is human

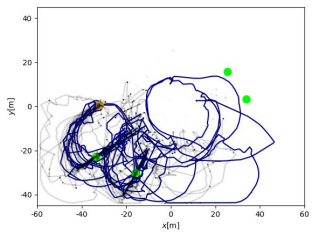


Player 1 is AA





Player 2 is AA

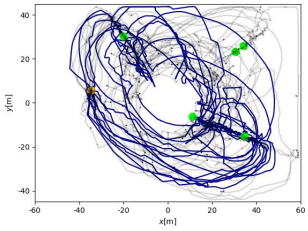


Player 2 is human

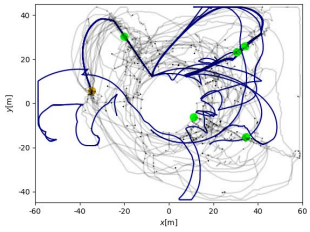
Data from human-human experiment
HA start position
Data from human-AA experiment
TA start position

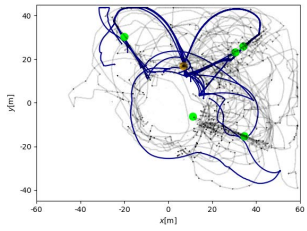
Trial #9

Player 1 is human

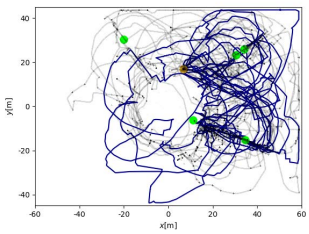


Player 1 is AA





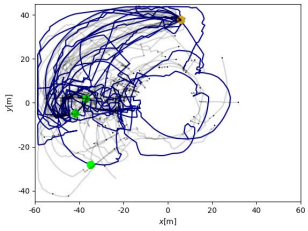
Player 2 is AA



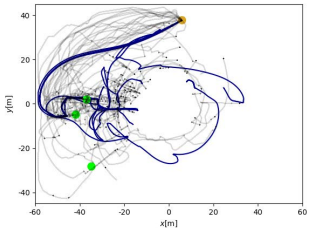
Player 2 is human

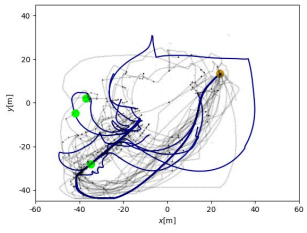
Trial #10

Player 1 is human

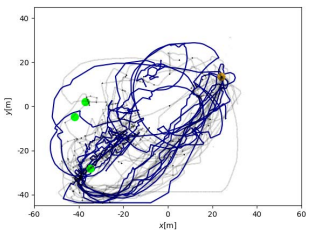


Player 1 is AA

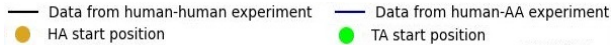




Player 2 is AA

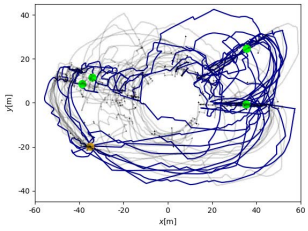


Player 2 is human

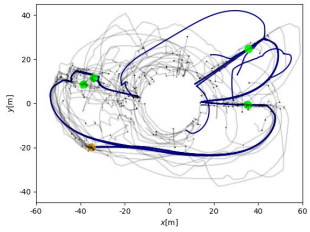


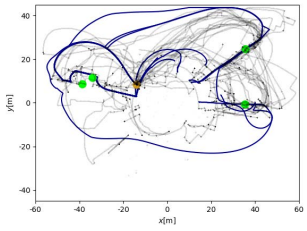
Trial #11

Player 1 is human

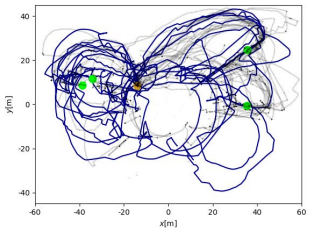


Player 1 is AA





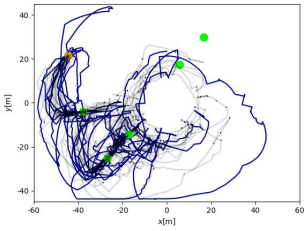
Player 2 is AA



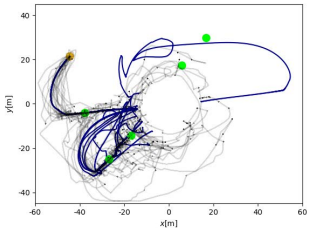
Player 2 is human

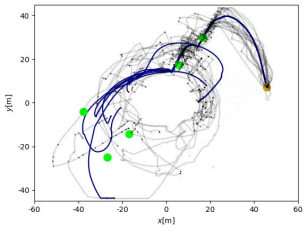
Trial #12

Player 1 is human

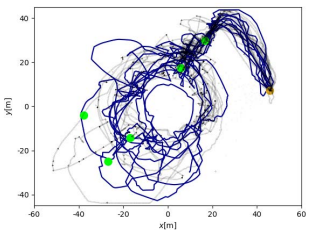


Player 1 is AA





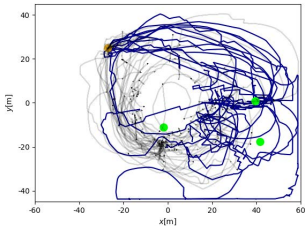
Player 2 is AA



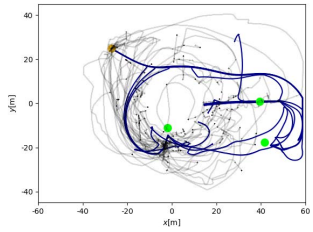
Player 2 is human

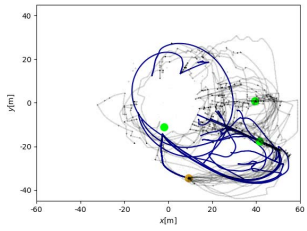
Trial #13

Player 1 is human

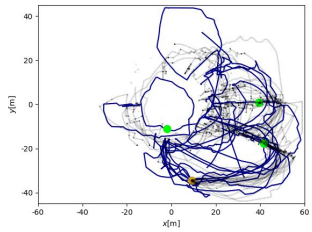


Player 1 is AA





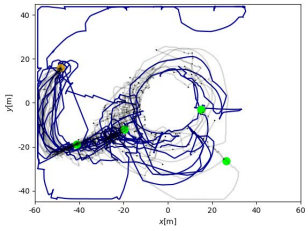
Player 2 is AA



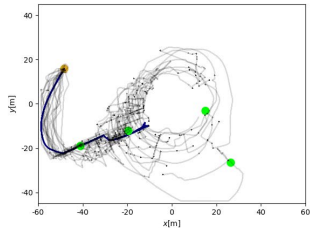
Player 2 is human

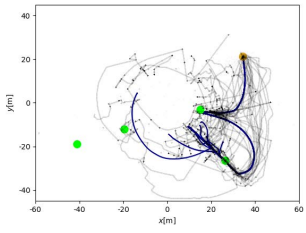
Trial #14

Player 1 is human

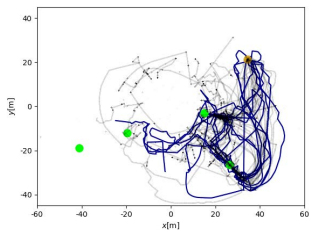


Player 1 is AA





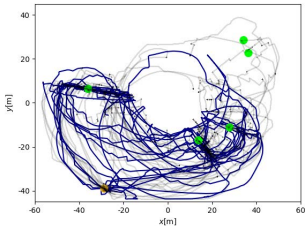
Player 2 is AA



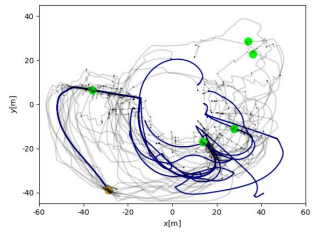
Player 2 is human

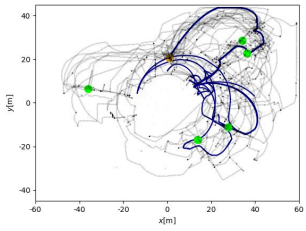
Trial #15

Player 1 is human

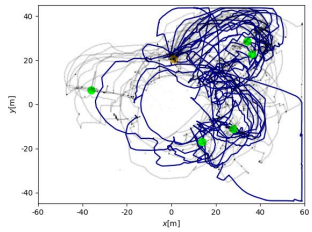


Player 1 is AA





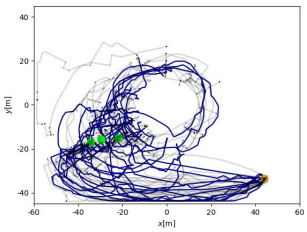
Player 2 is AA



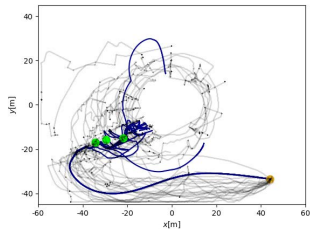
Player 2 is human

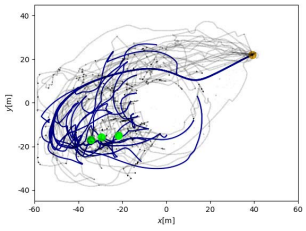
Trial #16

Player 1 is human

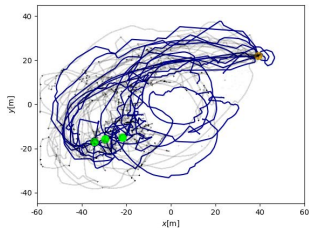


Player 1 is AA





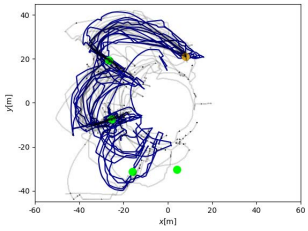
Player 2 is AA



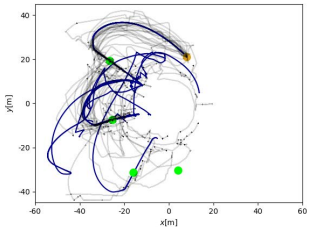
Player 2 is human

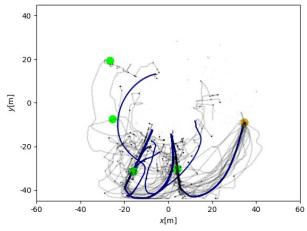
Trial #17

Player 1 is human

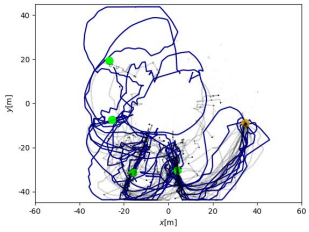


Player 1 is AA





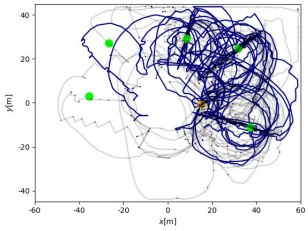
Player 2 is AA



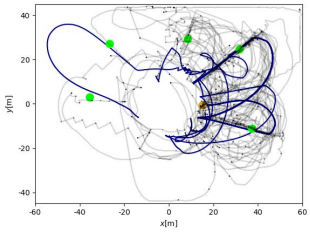
Player 2 is human

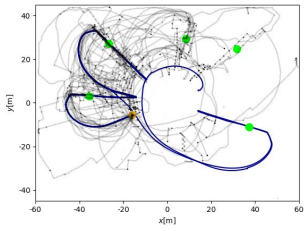
Trial #18

Player 1 is human

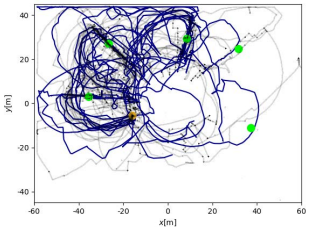


Player 1 is AA





Player 2 is AA



Player 2 is human

Bibliography

[1] P. W. Anderson, "More is different: Broken symmetry and the nature of the hierarchical structure of science." *Science*, vol. 177, no. 4047, pp. 393–396, 1972.

[2] E. Saltzman and J. Kelso, "Skilled actions: a task-dynamic approach." *Psychological review*, vol. 94, no. 1, p. 84, 1987.

[3] H. Haken, J. S. Kelso, and H. Bunz, "A theoretical model of phase transitions in human hand movements," *Biological cybernetics*, vol. 51, no. 5, pp. 347–356, 1985.

[4] P. N. Kugler and M. T. Turvey, "Self-organization, flow fields, and information," *Human Movement Science*, vol. 7, no. 2-4, pp. 97–129, 1988.

[5] B. R. Fajen and W. H. Warren, "Behavioral dynamics of steering, obstable avoidance, and route selection." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, no. 2, p. 343, 2003.

[6] B. Fajen and W. H. Warren, "Visual guidance of intercepting a moving target on foot," *Perception*, vol. 33, no. 6, pp. 689–715, 2004.

[7] W. H. Warren, "The dynamics of perception and action," *Psychological Review*, vol. 113, no. 2, p. 358, 2006.

[8] W. H. Warren and B. R. Fajen, "Behavioral dynamics of visually guided locomotion," in *Coordination: neural, behavioral and social dynamics*. Springer, 2008, pp. 45–75.

[9] W. H. Warren, "Collective motion in human crowds," *Current directions in psychological science*, vol. 27, no. 4, pp. 232–240, 2018.

[10] C. A. Coey, M. Varlet, and M. J. Richardson, "Coordination dynamics in a socially situated nervous system," *Frontiers in human neuroscience*, vol. 6, p. 164, 2012.

[11] R. P. van der Wel, C. Becchio, A. Curioni, and T. Wolf, "Understanding joint action: Current theoretical and empirical approaches," p. 103285, 2021.

[12] M. Varlet, L. Marin, J. Lagarde, and B. G. Bardy, "Social postural coordination." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 2, p. 473, 2011.

[13] D. Strömbom, R. P. Mann, A. M. Wilson, S. Hailes, A. J. Morton, D. J. Sumpter, and A. J. King, "Solving the shepherding problem: heuristics for herding autonomous, interacting agents," *Journal of the royal society interface*, vol. 11, no. 100, p. 20140719, 2014.

[14] R. A. Licitra, Z. I. Bell, and W. E. Dixon, "Single-agent indirect herding of multiple targets with uncertain dynamics," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 847–860, 2019.

[15] P. Nalepka, R. W. Kallen, A. Chemero, E. Saltzman, and M. J. Richardson, "Herd those sheep: Emergent multiagent coordination and behavioral-mode switching," *Psychological science*, vol. 28, no. 5, pp. 630–650, 2017.

[16] P. Nalepka, M. Lamb, R. W. Kallen, K. Shockley, A. Chemero, E. Saltzman, and M. J. Richardson, "Human social motor solutions for human–machine interaction in dynamical task contexts," *Proceedings of the National Academy of Sciences*, vol. 116, no. 4, pp. 1437–1446, 2019.

[17] J.-M. Lien, O. B. Bayazit, R. T. Sowell, S. Rodriguez, and N. M. Amato, "Shepherding behaviors," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 4. IEEE, 2004, pp. 4159–4164.

[18] R. Löhner, "On the modeling of pedestrian motion," *Applied Mathematical Modelling*, vol. 34, no. 2, pp. 366–382, 2010.

[19] H. Sun, L. Hu, W. Shou, and J. Wang, "Self-organized crowd dynamics: research on earthquake emergency response patterns of drill-trained individuals based on gis and multi-agent systems methodology," *Sensors*, vol. 21, no. 4, p. 1353, 2021.

[20] R. C. Schmidt, C. Carello, and M. T. Turvey, "Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people." *Journal of experimental psychology: human perception and performance*, vol. 16, no. 2, p. 227, 1990.

[21] M. J. Richardson, K. L. Marsh, R. W. Isenhower, J. R. Goodman, and R. C. Schmidt, "Rocking together: Dynamics of intentional and unintentional interpersonal coordination," *Human movement science*, vol. 26, no. 6, pp. 867–891, 2007.

[22] V. S. Chipade and D. Panagou, "Herding an adversarial swarm in an obstacle environment," in *Proceedings of the IEEE 58th Conference on Decision and Control (CDC'19)*. IEEE, 2019, pp. 3685–3690.

[23] M. Haque, A. Rahmani, and M. Egerstedt, "A hybrid, multi-agent model of foraging bottlenose dolphins," *IFAC Proceedings Volumes*, vol. 42, no. 17, pp. 262–267, 2009.

[24] W. Lee and D. Kim, "Autonomous shepherding behaviors of multiple target steering robots," *Sensors*, vol. 17, no. 12, p. 2729, 2017.

[25] A. Pierson and M. Schwager, "Controlling noncooperative herds with robotic herders," *IEEE Transactions on Robotics*, vol. 34, no. 2, pp. 517–525, 2017.

[26] H. El-Fiqi, B. Campbell, S. Elsayed, A. Perry, H. K. Singh, R. Hunjet, and H. A. Abbass, "The limits of reactive shepherding approaches for swarm guidance," *IEEE Access*, vol. 8, pp. 214 658–214 671, 2020.

[27] L. Huber, J.-J. Slotine, and A. Billard, "Fast obstacle avoidance based on real-time sensing," *IEEE Robotics and Automation Letters*, 2022.

[28] M. Koptev, N. Figueroa, and A. Billard, "Neural joint space implicit signed distance functions for reactive robot manipulator control," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 480–487, 2022.

[29] A. Mörtl, M. Lawitzky, A. Kucukyilmaz, M. Sezgin, C. Basdogan, and S. Hirche, "The role of roles: Physical cooperation between humans and robots," *The International Journal of Robotics Research*, vol. 31, no. 13, pp. 1656–1674, 2012.

[30] T. J. Wiltshire, S. F. Warta, D. Barber, and S. M. Fiore, "Enabling robotic social intelligence by engineering human social-cognitive mechanisms," *Cognitive Systems Research*, vol. 43, pp. 190–207, 2017.

[31] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: learning attractor models for motor behaviors," *Neural Computation*, vol. 25, no. 2, pp. 328–373, 2013.

[32] A. A. Paranjape, S.-J. Chung, K. Kim, and D. H. Shim, "Robotic herding of a flock of birds using an unmanned aerial vehicle," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 901–915, 2018.

[33] F. Auletta, D. Fiore, M. J. Richardson, and M. di Bernardo, "Herding stochastic autonomous agents via local control rules and online target selection strategies," *Autonomous Robots*, vol. 46, no. 3, pp. 469–481, 2022.

[34] P. Nalepka, M. Lamb, R. W. Kallen, E. Saltzman, A. Chemero, and M. J. Richardson, "First step is to group them: Task-dynamic model validation for human multiagent herding in a less constrained task." in *CogSci*, 2017.

[35] S. Schaal, P. Mohajerian, and A. Ijspeert, "Dynamics systems vs. optimal control—a unifying view," *Progress in Brain Research*, vol. 165, pp. 425–445, 2007.

[36] P. G. Amazeen, "From physics to social interactions: Scientific unification via dynamics," *Cognitive Systems Research*, vol. 52, pp. 640–657, 2018.

[37] G. Patil, P. Nalepka, L. Rigoli, R. W. Kallen, and M. J. Richardson, "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: 19th International Conference, PAAMS 2021, Salamanca, Spain, October 6–8, 2021, Proceedings 19*. Springer, 2021, pp. 176–187.

[38] M. J. Richardson, S. J. Harrison, R. W. Kallen, A. Walton, B. A. Eiler, E. Saltzman, and R. Schmidt, "Self-organized complementary joint action: Behavioral dynamics of an interpersonal collision-avoidance task." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 41, no. 3, p. 665, 2015.

[39] J. S. Kelso, "Principles of dynamic pattern formation and change for a science of human behavior," in *Developmental science and the holistic approach*. Routledge, 2000, pp. 73–94.

[40] D. Babajanyan, G. Patil, M. Lamb, R. W. Kallen, and M. J. Richardson, "I know your next move: Action decisions in dyadic pick and place tasks," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, no. 44, 2022.

[41] M. Lamb, R. W. Kallen, S. J. Harrison, M. Di Bernardo, A. Minai, and M. J. Richardson, "To pass or not to pass: Modeling the movement and affordance dynamics of a pick and place task," *Frontiers in Psychology*, vol. 8, p. 1061, 2017.

[42] M. Lamb, P. Nalepka, R. W. Kallen, T. Lorenz, S. J. Harrison, A. A. Minai, and M. J. Richardson, "A hierarchical behavioral dynamic approach for naturally adaptive human-agent pick-and-place interactions," *Complexity*, vol. 2019, 2019.

[43] S. Ekdawi, G. Patil, R. W. Kallen, and M. J. Richardson, "Modelling competitive human action using dynamical motor primitives for the development of human-like artificial agents," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, 2022.

[44] G. Patil, P. Bagala, P. Nalepka, R. W. Kallen, and M. J. Richardson, "Evaluating human-artificial agent decision congruence in a coordinated action task," in *Proceedings of the 10th international conference on human-agent interaction*, 2022, pp. 327–329.

[45] G. Patil, P. Nalepka, H. Stening, R. W. Kallen, and M. J. Richardson, "Scaffolding deep reinforcement learning agents using dynamical perceptual-motor primitives," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45, no. 45, 2023.

[46] L. Rigoli, G. Patil, P. Nalepka, R. W. Kallen, S. Hosking, C. Best, and M. J. Richardson, "A comparison of dynamical perceptual-motor primitives and deep reinforcement learning for human-artificial agent training systems," *Journal of Cognitive Engineering and Decision Making*, vol. 16, no. 2, pp. 79–100, 2022.

[47] G. Patil, P. Nalepka, R. W. Kallen, and M. J. Richardson, "Hopf bifurcations in complex multiagent activity: the signature of discrete to rhythmic behavioral transitions," *Brain Sciences*, vol. 10, no. 8, p. 536, 2020.

[48] L. M. Rigoli, P. Nalepka, H. Douglas, R. W. Kallen, S. Hosking, C. Best, E. Saltzman, and M. J. Richardson, "Employing models of human social motor behavior for artificial agent trainers," in *Proceedings of the 19th International Conference on Autonomous Agents and Multi-agent Systems*, 2020, pp. 1134–1142.

[49] F. Auletta, M. di Bernardo, and M. J. Richardson, "Human-inspired strategies to solve complex joint tasks in multi agent systems," *IFAC-PapersOnLine*, vol. 54, no. 17, pp. 105–110, 2021.

[50] P. Nalepka, P. L. Silva, R. W. Kallen, K. Shockley, A. Chemero, E. Saltzman, and M. J. Richardson, "Task dynamics define the contextual emergence of human corralling behaviors," *PloS one*, vol. 16, no. 11, p. e0260046, 2021.

[51] V. Mnih, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[52] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse *et al.*, "Dota 2 with large scale deep reinforcement learning," *arXiv preprint arXiv:1912.06680*, 2019.

[53] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[54] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," *arXiv preprint arXiv:1909.07528*, 2019.

[55] F. Auletta, R. W. Kallen, M. di Bernardo, and M. J. Richardson, "Predicting and understanding human action decisions during skillful joint-action using supervised machine learning and explainable-ai," *Scientific Reports*, vol. 13, no. 1, p. 4992, 2023.

[56] C. J. Watkins, "Learning from delayed rewards," *Kings College PhD Thesis*, 1989.

[57] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[58] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[59] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[60] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[61] J. Orr and A. Dutta, "Multi-agent deep reinforcement learning for multi-robot applications: A survey," *Sensors*, vol. 23, no. 7, p. 3625, 2023.

[62] J. Dinneweth, A. Boubezoul, R. Mandiau, and S. Espié, "Multi-agent reinforcement learning for autonomous vehicles: A survey," *Autonomous Intelligent Systems*, vol. 2, no. 1, p. 27, 2022.

[63] P. Nalepka, J. P. Gregory-Dunsmore, J. Simpson, G. Patil, and M. J. Richardson, "Interaction flexibility in artificial agents teaming with humans," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, no. 43, 2021.

[64] M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan, "On the utility of learning about humans for human-ai coordination," *Advances in neural information processing systems*, vol. 32, 2019.

[65] M. J. Prants, J. Simpson, P. Nalepka, R. W. Kallen, M. Dras, E. D. Reichle, S. Hosking, C. J. Best, and M. J. Richardson, "The structure of team search behaviors with varying access to information," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, 2021.

[66] J. Simpson, P. Nalepka, C. L. Crone, R. W. Kallen, M. Dras, E. D. Reichle, S. G. Hosking, C. J. Best, D. Richards, and M. J. Richardson, "Tip of the Finger or Tip of the Tongue? The Effects of Verbal Communication on Online Multi-Player Team Performance," in *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, 2021, pp. 175–178.

[67] A. bin Kamruddin, H. Sandison, G. Patil, M. Musolesi, M. di Bernardo, and M. J. Richardson, "Modelling human navigation and decision dynamics in a first-person herding task," *Royal Society Open Science*, vol. 11, no. 10, p. 231919, 2024.

[68] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd international conference on knowledge discovery and data mining*, 1994, pp. 359–370.

[69] P. Senin, "Dynamic time warping algorithm review," *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, vol. 855, no. 1-23, p. 40, 2008.

[70] D. Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[71] J. Nocedal and S. J. Wright, "Quadratic programming," *Numerical Optimization*, pp. 448–492, 2006.

[72] L. M. Rigoli, G. Patil, H. F. Stening, R. W. Kallen, and M. J. Richardson, "Navigational behavior of humans and deep reinforcement learning agents," *Frontiers in Psychology*, p. 4096, 2021.

[73] E. B. Sandoval, J. Brandstatter, U. Yalcin, and C. Bartneck, "Robot likeability and reciprocity in human robot interaction: Using ultimatum game to determinate reciprocal likeable robot strategies," *International Journal of Social Robotics*, vol. 13, no. 4, pp. 851–862, 2021.

[74] A. M. Haith and J. W. Krakauer, "Model-based and model-free mechanisms of human motor learning," in *Progress in motor control: Neural, computational and dynamic approaches*. Springer, 2013, pp. 1–21.

[75] S. S. Mousavi, M. Schukat, and E. Howley, "Deep reinforcement learning: an overview," in *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016: Volume 2*. Springer, 2018, pp. 426–440.

[76] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman *et al.*, "Human-level performance in 3d multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019.

[77] F. Fuchs, Y. Song, E. Kaufmann, D. Scaramuzza, and P. Dürr, "Super-human performance in gran turismo sport using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4257–4264, 2021.

[78] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[79] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020.

[80] A. Shafti, J. Tjomsland, W. Dudley, and A. A. Faisal, "Real-world human-robot collaborative reinforcement learning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 11 161–11 166.

[81] N. J. McNeese, M. Demir, N. J. Cooke, and C. Myers, "Teaming with a synthetic teammate: Insights into human-autonomy teaming," *Human factors*, vol. 60, no. 2, pp. 262–273, 2018.

[82] N. Sebanz and G. Knoblich, "Prediction in joint action: What, when, and where," *Topics in cognitive science*, vol. 1, no. 2, pp. 353–367, 2009.

[83] R. Kelter, "Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to bayesian inference with jasp," *BMC Medical Research Methodology*, vol. 20, pp. 1–12, 2020.

[84] D. Pickem, P. Glotfelter, L. Wang, M. Mote, A. Ames, E. Feron, and M. Egerstedt, "The robotarium: A remotely accessible swarm robotics research testbed," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1699–1706.

[85] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Hausdorff1," *IEE Transactions on Pattern Analysis and Machine Intelligence*, 1993.

[86] M. P. Dubuisson and A. K. Jain, “A modified hausdorff distance for object matching,” vol. 1. Institute of Electrical and Electronics Engineers Inc., 1994, pp. 566–568.

[87] T. Eiter and H. Mannila, “Computing discrete fréchet distance computing discrete fréchet distance,” 1994.

[88] D. J. Bemdt and J. Clifford, “Using dynamic time warping to findpatterns in time series,” 1994. [Online]. Available: www.aaai.org

[89] P. Senin, “Dynamic time warping algorithm review,” 2008.

[90] D. Kraft, “A software package for sequential quadratic programming,” *Tech. Rep. DFVLR-FB 88-28, DLR German Aerospace Center – Institute for Flight Mechanics, Koln, Germany*, 1988.

[91] *Sequential Quadratic Programming*. New York, NY: Springer New York, 2006, pp. 529–562. [Online]. Available: https://doi.org/10.1007/978-0-387-40065-5_18