The Aeronautical Journal (2022), **126**, pp. 907–931
doi:10.1017/aer.2022.8

REGULAR PAPER

Design of simulation-based pilot training systems using machine learning agents

J. Källström[1,*], R. Granlund[2] and F. Heintz[1]

[1]Department of Computer and Information Science, Linköping University, Linköping, Sweden and [2]RISE SICS East, Linköping, Sweden
*Email: johan.kallstrom@liu.se

Received: 16 June 2021; **Revised:** 22 November 2021; **Accepted:** 12 January 2022

Keywords: Air combat training; Flight simulation; LVC simulation; Machine learning; Reinforcement learning

Abstract

The high operational cost of aircraft, limited availability of air space, and strict safety regulations make training of fighter pilots increasingly challenging. By integrating Live, Virtual, and Constructive simulation resources, efficiency and effectiveness can be improved. In particular, if constructive simulations, which provide synthetic agents operating synthetic vehicles, were used to a higher degree, complex training scenarios could be realised at low cost, the need for support personnel could be reduced, and training availability could be improved. In this work, inspired by the recent improvements of techniques for artificial intelligence, we take a user perspective and investigate how intelligent, learning agents could help build future training systems. Through a domain analysis, a user study, and practical experiments, we identify important agent capabilities and characteristics, and then discuss design approaches and solution concepts for training systems to utilise learning agents for improved training value.

Nomenclature

a, r, s	action, reward, and state in a Markov decision process
A, R, S, T	the actions, rewards, states, and dynamics of a Markov decision process
AI	Artificial Intelligence
AP, DP, PP	Action, Decision, and Perception Points
CAP	Combat Air Patrol
d	distance
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q Networks
E	expectation
ESR, SER	Expected Scalarized Return and Scalarized Expected Return
FAOR	Fighter Area Of Responsibility
HDD	Head Down Display
I	the initiation set of an option
JCF	Joint Control Framework
JCF-S	Joint Control Framework Score
LACC	Levels of Autonomy in Cognitive Control
LVC	Live, Virtual, and Constructive
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
MCTS	Monte-Carlo Tree Search
MDP	Markov Decision Process
MLP	Multi-Layer Perceptron
MOMDP	Multi-Objective Markov Decision Process
MORL	Multi-Objective Reinforcement Learning

OTW	Out-The-Window
p	position
$Q(s, a)$	the state-action value function of a Markov decision process
G_t	the return received in a Markov decision process from time t
ReLU	Rectifier Linear Unit
t	time
u	the utility of an agent
U	the utility function of an agent
$V(s)$	the state value function of a Markov decision process
WVR	Within Visual Range

Greek symbol

α	learning rate in reinforcement learning
β	termination condition of an option
γ	the discount factor of a Markov decision process
ω	an option
Ω	a set of options
π	the decision-making policy of an agent

1.0 Introduction

Providing efficient and effective training solutions for fighter pilots is becoming increasingly challenging. Due to the high operational cost of aircraft, limited availability of air space and strict safety regulations, it is difficult to realise training scenarios with the desired contents and density in a live setting. Instead, virtual and constructive simulation resources must be used to a higher degree. Live, Virtual and Constructive (LVC) simulation aims to integrate real aircraft, ground-based systems and soldiers (Live), manned simulators (Virtual) and computer-controlled entities (Constructive) [1]. By using constructive simulation to augment the live and virtual aircraft operated by trainees, it is possible to improve training effectiveness by simulating scenarios with a large number of participating entities [2]. However, training value will depend on the quality of the agents used to control the constructive entities. Ideally, these agents should be able to act as synthetic instructors, and adapt their behaviour to the training needs of the human trainees. This would allow us to minimise the number of human support personnel required for conducting training, which would lead to lower costs and improved training availability.

As illustrated in Fig. 1, we can divide the users of training systems into two major categories: training audience and training providers. The training audience consists of those in training, e.g. pilots learning how to operate a new aircraft, while the training providers consist of those delivering the training, e.g. instructors, operators, and role-players. Instructors are responsible for the pedagogical contents of a training session, while role-players and scenario operators help deliver the training by participating as actors or controlling parts of the simulated scenario respectively. If synthetic agents were to become smarter, they could replace or augment human role-players, and reduce the amount of human input required for the training scenario to progress in the desired way. To further raise the level of autonomy of the system, agents could also assist instructors in evaluating the performance of the trainees, and in adapting the contents and characteristics of training scenarios. However, creating behaviour models for the agents is challenging, especially for end-users of training systems (e.g. instructors), who may not have the required expertise and experience [3]. In the past, this has constrained the use of agents in training. Now, with the recent advances in artificial intelligence (AI), there is hope that data driven methods will simplify the process of constructing intelligent agents, which could replace human support personnel in simulation-based training.

For learning sequential decision-making, reinforcement learning [4] has become the state of the art method. Guided by a human-designed reward signal, such agents can learn a policy purely by interacting

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

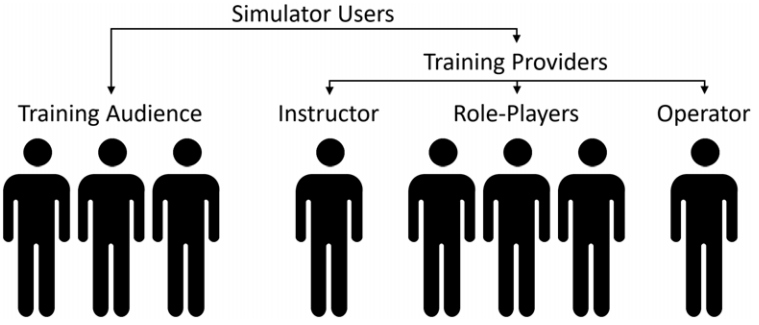

Figure 1. *Users of simulation-based pilot training systems (from Ref. [6]).*

with their environment. By leveraging deep learning [5], it has become possible to beat human champions in classic board games as well as multi-player computer games [7–10]. The results have sparked interest in investigating applications of reinforcement learning in many domains, including air combat simulation. However, the focus has, to a large extent, been on manoeuvre optimisation, rather than potential improvement of training value. To successfully design agents suitable for training, a good understanding of the domain and its actors is essential.

In this work, we proceed to study learning agents from a user perspective, with support from experienced fighter pilots. The goal is to learn more about how intelligent agents could be used to automate some of the tasks performed by human training providers. Our contributions and the structure of the paper can be summarised as follows.

- First, we perform an analysis of the domain of simulation-based training. The analysis is conducted from the perspectives of instructors and trainee pilots respectively. The purpose of the analysis is to identify constraints imposed on training providers when using different types of simulation resources, and to model the patterns of decision-making a synthetic agent must be capable of if it is to replace human role-players in air combat scenarios.
- Second, after a brief introduction to reinforcement learning, we conduct a user study consisting of repeated interviews and a written survey, with the purpose of finding out what experienced pilots consider important agent capabilities and characteristics in different types of simulation-based training scenarios.
- Third, we conduct a study of human-agent interaction in an air combat scenario, where agents trained with a state of the art reinforcement learning algorithm cooperate with humans to solve an air policing task. The purpose of the experiment is to study how aspects of the agent design affects the agent's performance.
- Finally, we discuss design approaches and solution concepts within the context of a system architecture for a simulation-based training system that incorporates learning agents. The purpose is to provide a breakdown of the problem into smaller sub-problems, and provide framing for future research efforts.

2.0 Domain Analysis of Simulation-Based Pilot Training

In this section we conduct an analysis of the domain of simulation-based pilot training. The aim is to identify and illustrate how different types of simulation resources affect the constraints imposed on actors that provide training, and what decision-making capabilities a learning agent would need to have to effectively participate in training scenarios, acting in a similar way as human role-players.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

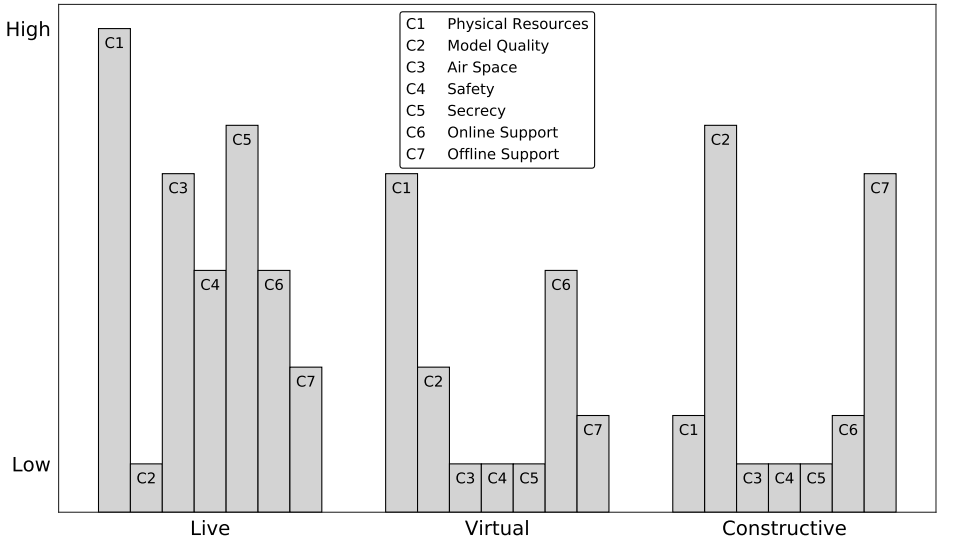

Figure 2. *Constraints affecting training for different types of simulation resources.*

2.1 Constraints when using different types of simulation resources

In interviews with pilots, we tried to identify constraints that they currently face when acting in the role of instructor. Based on these interviews, Fig. 2 shows a subjective measure of the relative importance of different types of constraints, for three different types of training simulations: Live, Virtual and Constructive. We discuss these constraints further in the text below.

Live training simulation provides the highest possible fidelity in terms of interaction with the aircraft and its subsystems. However, in the Live setting, training is highly affected by aspects of the physical world. For instance, the availability of vehicles and other types of systems may not be sufficient to realise complex scenarios. In particular, a military organisation may not have access to systems that have the same performance and characteristics as those that are used by the enemy. Furthermore, operation of physical vehicles, e.g. aircraft, is highly expensive, which limits the amount of training that can be delivered in this setting.

Training in the Live setting is also constrained by the limited availability of air space, as well as safety regulations, which makes it difficult to realise scenarios with many entities, who are operating over a large geographical area. A large number of support personnel may also be required to plan and conduct such exercises. In addition, when acting in the open secrecy may be compromised, and there is a risk that systems' performance and tactics are revealed to opponents.

By using ground-based, Virtual simulators, the constraints imposed by the physical environment are lifted, and training delivery becomes easier. Still, there is a considerable cost related to populating complex scenarios with a large number of high-fidelity simulators. Constraints regarding model fidelity increase in this setting, in particular for within visual range (WVR) combat, where the effects of, e.g. g-forces is an important factor for pilot performance. To populate scenarios with only Virtual participants, some humans must act on the opponent's side. If they use high-fidelity pilot stations to play this role in the scenario, they can learn to understand the opponent's systems and tactics. However, if the simulators used for this type of role-play do not have sufficient fidelity, the training value will be low.

Constructive simulation makes it possible to realise large scenarios, populated by synthetic entities, which can replace human role-players. This reduces the need for physical resources, so that only the computation hardware for running the simulation software is required. Instead, the constraints are shifted to the fidelity of the simulation models, and the available offline support for building the models, as well as the simulation scenarios. In particular, it becomes challenging to construct behaviour models for the synthetic entities, and adapting models to the training needs of individual trainees. Since the expertise

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

required for such tasks may not be available locally, at each training facility, the turn-around time for updating training contents may be long. Instead, it may be necessary to have scenario operators manually control the flow of the tactical scenario to some extent.

Learning agents have the potential to reduce the constraints of constructive simulation, by simplifying the construction of high-quality behaviour models. This would enable training providers to keep simulation contents in pace with training needs, and to reduce the dependency on human role-players. Data-driven methods can also provide objective evaluations, on a machine-readable format, which can support automated adaptation of simulation contents, so that training scenarios are always in pace with training needs.

2.2 Human-machine interaction for decision-making in air combat

In this section, we study aspects of human-machine interaction in air combat scenarios. The aim is to illustrate to what extent perception, decisions, and actions are supported by the automation of the aircraft or pre-planned procedures, and which parts of aircraft control that must be handled by the pilot alone. This information gives insight regarding requirements that must be fulfilled by synthetic agents that are to replace human pilots in training scenarios, and how to design the interface between the agent and the aircraft model, including its tactical systems. When creating synthetic pilots, the high-level tactical decisions made by human pilots will be the most challenging ones to model using AI. Therefore, the information available to support human decision-making at this level of abstraction should also be incorporated in AI algorithms to maximise their performance.

In support of our study, we use the Joint Control Framework Score (JCF-S) notation [11]. JCF-S is intended to support modelling of temporal aspects of human-machine interaction, at different levels of autonomy in cognitive control (LACC). The levels are summarised in the list below.

1. The Physical level, which shows constraints related to physical actions
2. The Implementation level, which shows constraints related to implementation properties
3. The Generic level, which provides generic plans for common situations
4. The Value level, which handles trade-offs among the system's objectives
5. The Effect level, which deals with the system's purpose and goals
6. The Framing level, which identifies the situation and context for control

Levels 1 and 2 determine how control is realised, levels 3 and 4 deal with what is done, and levels 5 and 6 are related to why the system exists. To model the joint control of human and machine, perception points (PP), decision points (DP) and action points (AP) are placed on six timelines, each of them representing one of the LACC levels. As a result, a pattern of the control loop of the joint system emerges. When agents are used as synthetic pilots in training scenarios, they should display a similar decision-making pattern as human pilots.

To illustrate how the capabilities of the pilot's tactical control loop (Observe-Orient-Decide-Act) are mapped to different levels of cognitive control, we study the engagement of two pilots in offensive and defensive counterair operations, i.e. the quest for a favourable air situation, air superiority or air supremacy. For this study, we use the scenario illustrated in Fig. 3 to set the context. In this scenario, two aircraft are flying a Combat Air Patrol (CAP) directed towards the south to protect their assigned Fighter Area Of Responsibility (FAOR), which is illustrated by the large circle in the figure. The FAOR contains three high-value assets, which are illustrated by the smaller circles in the figure. Approaching from the west are two hostile aircraft, which intend to perform an opportunistic attack on the high-value assets, and must therefore first deal with the defending aircraft of the CAP. We assume that the defensive fighters are trainees, while the offensive fighters are human role-players, who try to support the training of the trainees.

The engagement is modelled using the JCF score notation, and the result is illustrated in Fig. 4. To simplify the notation we only present the engagement of two of the aircraft in the scenario.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

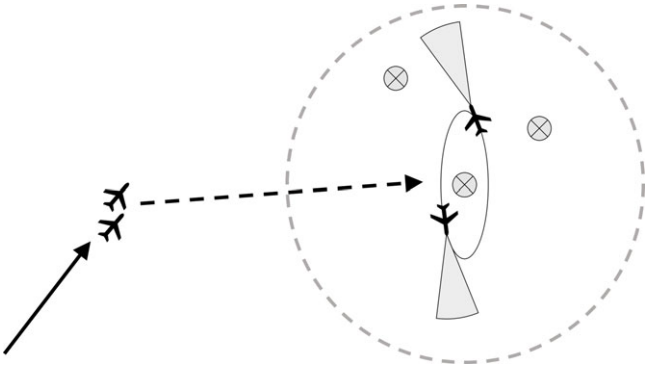

Figure 3. *Hostile entities approaching a Combat Air Patrol (CAP).*

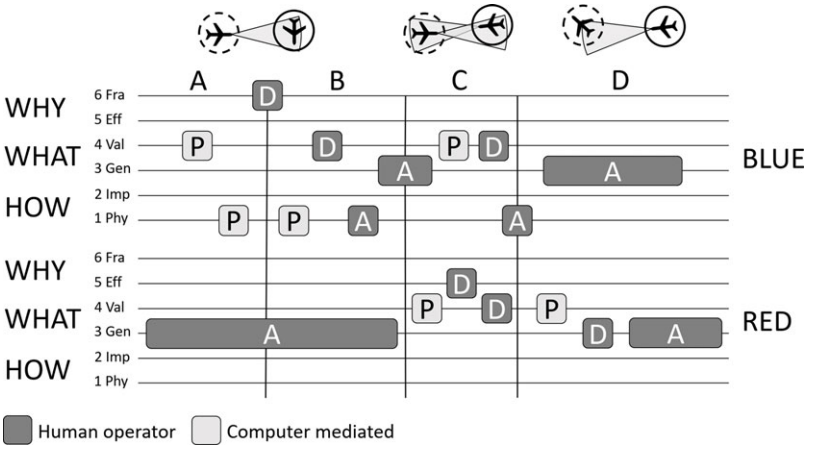

Figure 4. *Score for blue (in solid circle) and red (in dashed circle) forces in a counterair operations scenario.*

The timeline of the scenario is divided into four sections (A-D), where significant events occur. The behaviour of the defending agent is presented in the top score (labelled *BLUE*), and the behaviour of the attacking agent is presented in the bottom score (labelled *RED*). At the top of the figure, the geometry between the two aircraft in different sections of the scenario is illustrated.

In section A of the scenario, a hostile RED aircraft is approaching the FAOR of the opposing BLUE aircraft. The approach is carried out according to a pre-planned procedure (AP on level 3 Gen). The pilot of the BLUE aircraft is informed by the decision support system that it is in the radar field-of-view of the RED aircraft, and uses the head-down displays (HDDs) to update himself regarding the scenario geometry (PP on levels 4 Val and 1 Phy respectively). He then makes a decision regarding the current threat level (DP on level 6 Fra).

In section B of the scenario, the pilot of the BLUE aircraft once again refers to the HDDs, to assess how to best deal with the threat (PP on level 1 Phy). The decision is then made that the most valuable course of action is to engage the target (DP on level 4 Val). After the decision has been communicated to the tactical air unit (AP on level 1 Phy), the pilot proceeds with target engagement according to doctrine (AP on level 3 Gen).

In section C of the scenario, the pilot of the RED aircraft is informed by the decision support system that it is in the radar field-of-view of the BLUE aircraft, which it is tracking (PP on level 4 Val). The pilot considers desirable effects related to tactical mission goals as well as trainees' training goals (DP

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

on level 5 Eff) and decides to proceed into the BLUE aircraft's FAOR, with the hope of attacking a high-value asset (DP on level 4 Val). In the meantime, the pilot of the BLUE aircraft observes that the RED aircraft is now within range (PP on level 4 Val), and decides to fire a missile (DP on level 4 Val followed by AP on level 1 Phy).

In section D, after firing the missile, the pilot of the BLUE aircraft guides it towards the target according to doctrine, until handover (AP on level 3 Gen). The pilot of the RED aircraft is informed by the decision support system that there is an incoming missile (PP on level 4 Val) and performs an evasive manoeuvre to avoid the threat (DP followed by AP on level 3 Gen).

We can see that the pilot is supported by refined, abstract information, provided by the decision support system, to form his situational awareness. We can also see that several actions are pre-defined to handle a certain situation, and have a temporal extension, e.g. target approach procedures, missile guidance procedures and evasive manoeuvres. Finally, decisions on how to handle the situations that occur are often taken at the higher levels of cognitive control, were full automation may not currently be available. Therefore, pilots still play a vital part in the outcome of missions. They must have the capability to comprehend the situation, to identify and rank potential threats and targets. Then, when acting upon their situational awareness, pilots must carefully choose how to use the tactical systems of the aircraft.

3.0 Constructing Agents for Simulation-Based Training

In this section, we first discuss limitations of current methods for constructing synthetic agents for simulation-based training. Then we provide a brief introduction to a machine learning technique called reinforcement learning, which could help simplify the construction of high-quality agents that could act as synthetic role-players in training scenarios.

3.1 Limitations of current agent technologies

In interviews with experienced pilots, we discussed to what extent agents could currently be used to provide high quality training, and what challenges instructors were currently facing. Typically, instructors are provided with synthetic entities and behaviour models from their support organisation. However, these models may not fit all relevant training cases, especially as time passes, and aspects of the operational environment change. Therefore, it would be good if instructors could adapt training contents on their own, without the support of simulator engineers. However, instructors feel that this is difficult when using the tools that are currently available for behaviour modelling. This is not surprising, since the construction of behaviour models for multi-agent systems is a highly challenging task and a very active area of research. When simulator engineers must be involved, the turn-around time increases. It is also challenging to translate human domain expertise to model parameters that engineers can base their implementations upon.

At present time, the highest training value is achieved when using agents as synthetic opponents. This reduces the need for support personnel, who do not receive training, to participate in training scenarios. It also reduces the need for expensive equipment, e.g. aircraft or high-fidelity simulators. However, handcrafted behaviour models often result in behaviour that comes across as scripted, static and predictable. To get the variation required in a stimulating learning environment, a lot of manual work and time must be invested, and the cost of keeping in pace with training needs may be high. By using machine learning, it could become possible to construct behaviour models that continually adapt to changes in the training environment, e.g. encounters with new trainees, introduction of new aircraft systems and changes in trainees' tactics. Machine learning could also help simplify the interfaces of the tools used to build and control agent behaviour, so that explicit programming were no longer required. Instead, instructors could use their domain knowledge to specify the goals and characteristics of the agents.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

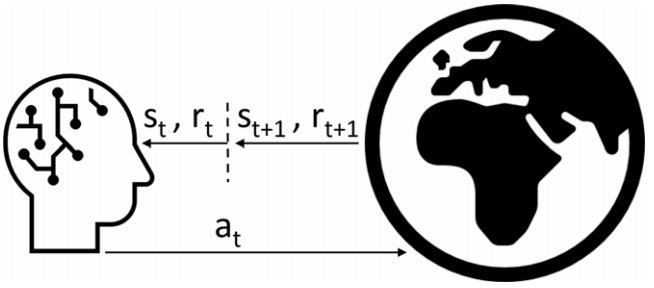

Figure 5. *Markov decision process.*

3.2 Learning sequential decision-making

In recent years, reinforcement learning [4] has come to be the state of the art method for learning sequential decision-making. By leveraging deep learning [5], it has become possible for synthetic agents to beat human champions in classic board games [7, 8], solve challenging robotics tasks [12–14] and learn how to play single and multi-player video games directly from pixel input [9, 10, 15]. The results have sparked interest in investigating applications of reinforcement learning in many domains, including air combat simulation.

Reinforcement learning allows an agent to learn a function for decision-making (policy π) by interacting with its environment in a form of trial-and-error learning [4]. A reinforcement learning problem is often modelled as a Markov Decision Process (MDP), or derivations thereof. A Markov Decision Process is defined by the tuple (S, A, T, R, γ), specifying:

- S: The set of states of the process
- A: The set of actions of the process
- T: The transition dynamics of the process
- R: The reward function of the process
- γ: The discount factor indicating the importance of immediate and future rewards respectively

The agent interacts with its environment by selecting actions according to its policy ($a_t = \pi(s_t)$) and observes the resulting environment state (s_{t+1}) and the received reward (r_{t+1}). When the agent executes an action that results in high reward, that action is reinforced, so that it will be taken more often in the future. During learning, the agent must balance between exploration and exploitation, which is one of the greatest challenges of reinforcement learning. Exploration means that the agent selects exploratory actions to learn more about the environment, while exploitation means that the agent uses the knowledge gained so far to gather reward. Learning can be on-policy or off-policy. For on-policy methods, the policy used for exploration (the behaviour policy) is the same as the policy being optimised (the target policy). For off-policy methods, the behaviour policy can be different from the target policy, e.g. the behaviour policy may sometimes take random actions to promote exploration. The process of reinforcement learning is illustrated in Fig. 5.

The goal of the agent is to maximise its future expected return G_t when starting in state s_t and then following policy π, which is captured in the state value function $V_\pi(s)$:

$$V_\pi(s) = E[G_t|s_t = s] = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s\right] \tag{1}$$

We can also define a state-action value function Q, which specifies the value of taking action a in state s and then following policy π:

$$Q_\pi(s, a) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s, a_t = a\right] \tag{2}$$

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

The Q function can be used as a policy, by greedily selecting the action with highest estimated value. The Q function can be learned through Q-learning [16] by representing the Q function as a table of Q values and applying the following update rule (with learning rate α) in each step of the episode:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \tag{3}$$

The tabular approach to reinforcement learning does not scale well to complex state and action spaces, which limits it applicability to many real-world problems. A breakthrough in reinforcement learning was the development of the Deep Q Networks (DQN) algorithm, which uses a neural network to represent the policy, making it possible for agents to learn how to play video games from pixels [15]. As one of the early scalable deep reinforcement learning algorithms, it has been evaluated in many application domains, including air combat simulation [6, 17–20].

The DQN algorithm can only learn policies for discrete actions, which may limit the applicability to, e.g. robotics problems. The Deep Deterministic Policy Gradient (DDPG) algorithm extended deep reinforcement learning to domains with continuous actions [12]. It is an actor-critic architecture, that uses a deep Q network (the critic) to estimate the values of actions, to guide updates of the agent's (the actor's) policy. In air combat simulation, continuous actions can be valuable for platform manoeuvring, and there have been a number of studies, primarily for WVR combat scenarios [21–25].

For success in air combat, agents need to learn how to cooperate with teammates in complex, competitive environments. This is a challenging task for reinforcement learning, since the environment becomes non-stationary when multiple agents are learning concurrently. Lowe et al. proposed the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm, an extension of DDPG to multi-agent environments [26] to address this challenge. The algorithm proposes to learn policies in a centralised fashion, allowing the critic of each agent access to the observations and actions of all other agents in the system. This simplifies determining what effect the behaviour of an individual agent has on the dynamics of the complete system. Though such approaches are valuable for air combat simulation, there has been surprisingly few research efforts in that direction, although interest seems to be increasing [6, 27–30].

While existing work on reinforcement learning for air combat simulation has covered some ground in investigating the applicability of different types of learning algorithms, the focus has been on optimisation of air combat manoeuvres, rather than potential added value for the users of simulation-based training systems. For this reason, the next section provides an analysis of desirable agent capabilities and characteristics from a user perspective to identify how learning agents could support instructors and trainees.

4.0 User Needs in Simulation-Based Training Using Agents

In this section, we present the results of a user study, which aimed at identifying how synthetic agents could help make simulation-based pilot training more efficient and effective. We discuss which capabilities and characteristics agents are expected to have from the perspectives of trainee pilots as well as instructors.

4.1 Organisation of the study

The study consisted of repeated user interviews and a follow-up written survey. The participants of the interviews and the survey were experienced fighter pilots from the Swedish Air Force and experienced test pilots from Saab Aeronautics. Three pilots participated in the interviews, while 25 pilots participated in the survey. Since this sample of pilots is quite small, the views expressed may not be fully representative of the whole population of pilots, but hopefully the feedback given by the participants can help identify important aspects related to the design of agents that are intended to act as synthetic pilots in training scenarios.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

The goal of the interviews was to allow pilots to describe current challenges in pilot training, and possible areas of improvement. In particular, the focus was on ways to automate training delivery to a higher degree using intelligent learning agents to reduce the dependency on support personnel such as role-players and scenario operators, and to improve the availability of high-quality training while reducing cost. Participants were initially asked to share their thoughts on training goals, training approaches, and training media to give an unbiased overview of how training is currently conducted. Thereafter, the interviewers took a more active part to identify the achievable training value when using agents in place of human role-players, to learn about challenges related to constructing training scenarios when using agents and to discuss what role learning agents could play in simulation-based training systems in the future.

The interviews with pilots revealed a set of important factors that would need to be considered in the design of synthetic agents. For the written survey, based on the information gathered through the conducted interviews, a number of statements regarding desirable agent capabilities and characteristics were presented to the participants. They were asked to rate to what degree they agreed with the statements for three different types of training: Basic Training, Tactical Procedure Training and Mission Training. In the Basic Training phase, a pilot with previous experience in flying a different type of aircraft, or a different edition of an aircraft, is trained in basic flight manoeuvres and system operation. In the Tactical Procedure Training phase, a pilot is trained in using, e.g. tactical sensors, data links and weapon systems in typical combat scenarios. In the Mission Training phase, pilots are trained to cooperate in teams to carry out typical operational missions. The intention was to identify how user needs differed for these three types of training.

The statements presented to respondees were divided into three categories: *Types of Agent Behaviour*, *Human-Agent Interaction*, and *Agent Behaviour in Training Scenarios*. Respondees were asked to give a score in the range one (low importance) to ten (high importance) for each statement, and they also had the possibility to add additional comments in free text. The results of the survey are presented as box plots (created using Matplotlib [31]) for each category of training. In these plots, the horizontal line of a box represents the median value, the upper and lower edges (hinges) of a box represent the upper and lower quartiles and the lines (whiskers) protruding from a box represent the upper and lower extreme values. Circles represent data points that are classified as outliers, if they exist. The notches surrounding the medians of the boxes can be used to judge the significance of the difference between two median values. If the notches of two boxes do not overlap, then the confidence level of the difference is roughly 95% [32]. If a notch would go outside of a box, protruding notches (with a "flipped" appearance) are plotted, which indicates a skewed distribution of values for that category. Median values are connected with a line for each statement to help identify trends.

4.2 Desirable agent capabilities and characteristics

For the category of *Types of Agent Behaviour*, the following statements were presented to respondees for rating:

- **Deterministic:** It is important that synthetic tactical entities can be given a deterministic behaviour.
- **Advanced:** It is important that synthetic tactical entities can display advanced tactical behaviour.
- **Doctrinal:** It is important that synthetic tactical entities can act according to doctrine.

The aim of this category of statements was to investigate the importance of different types of agent behaviour in different types of training. The scores given by respondees, which indicate to what degree they agree with the statements, are presented in Fig. 6.

Regarding the ability to assign agents a Deterministic behaviour, we can see that this is important in all phases of training, although the scores vary more for mission training. The scores for Advanced behaviour increases as we move from Basic Training, where the importance is modest, to Mission

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

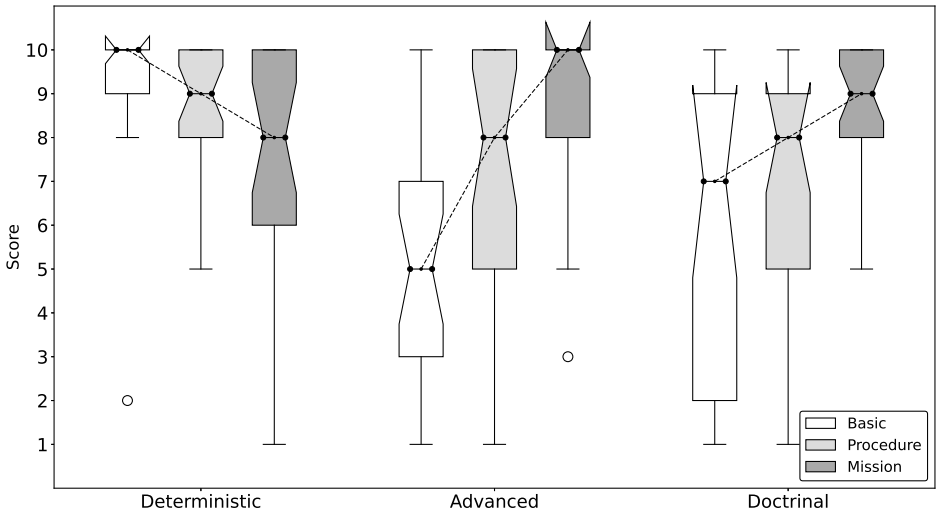

Figure 6. *Importance of different types of agent behaviour.*

Training, where the importance is very high. Here, there is some variance in the scores for Basic Training and Tactical Procedure Training, while the pilots are in high agreement regarding the importance of Mission Training with advanced synthetic opponents. Finally, the importance of Doctrinal behaviour shows a similar pattern as Advanced behaviour, becoming more important as we move towards Mission Training, although it is fairly important already for Basic Training. As for Advanced behaviour, the variance in the scores is higher for Basic Training and Tactical Procedure Training than for Mission Training.

In interviews, pilots expressed that in the initial phases of training, the requirements on synthetic opponents are rather modest. In this phase of training, it is important that the behaviour of synthetic opponents is predictable. The most important thing is to be able to create well defined, deterministic scenarios. For instance, when learning the functions and controls of a new sensor system, it may be distracting if opponents behave in an unpredictable manner. Instead, entities may move along predefined trajectories, or the positions of vehicles, including the trainee's own aircraft, may be frozen. In many scenarios, there are synthetic entities that are primarily used as background noise, and it is then desirable that they can perform simple tasks such as start and landing. For entities that play a tactical role in the scenario, there are also well established, standard manoeuvres that they are expected to be able to perform, such as straight flight, gimbal turn and pincer manoeuvre.

As the training progresses, more advanced synthetic pilots, who can take defensive as well as offensive roles, are required to realise scenarios that allow trainees to develop their tactical proficiency. One pilot reasoned that a good base requirement for entity behaviour is the ability to respond, in a believable way, to all orders available on the aircraft tactical data link. Furthermore, as explained by the participants in the study, providing entities that display advanced tactical behaviour is a necessary, but not sufficient condition. It is also required that synthetic pilots can follow a certain doctrine when acting in training scenarios, to prepare trainees for a variety of potential adversaries. That is, it must be possible to model important aspects of established procedures for completing a certain type of mission, e.g. standard manoeuvres, definitions of high-value targets and directions for when to attack or retreat. Such a modelling capability is a natural component of Mission Training, which is supposed to support preparations for specific missions. This is also supported by the results of the survey. This means that when behaviour models are developed using machine learning techniques, there must be a way to infuse domain knowledge in the learning process, so that the resulting behaviour fulfils rules encoded in a specific doctrine.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

For the category of *Human-Agent Interaction*, the following statements were presented to respondees for rating:

- **Challenging Opponent:** It is important that synthetic tactical entities can act as challenging opponents (e.g. by discovering and exploiting flaws in the human trainee's tactics and execution).
- **Wingmate:** It is important that synthetic tactical entities can act as wingmates of human trainees with intelligent behaviour.
- **Voice Communication:** It is important that a synthetic tactical entity that acts as wingmate can communicate with human trainees through radio voice communication.

The aim of this category of statements was to investigate the importance of having agents act in different types of roles in different types of training scenarios, as well as the importance of voice interaction with agents. The scores given by respondees, which indicate to what degree they agree with the statements, are presented in Fig. 7.

Regarding the ability of agents to act as Challenging Opponents, we can see a quite wide range of scores from Basic Training to Mission Training. The median score for Basic Training is low, while the median scores for Tactical Procedure Training and Mission Training are high. There is quite large variance in the responses for the two simpler categories of training, while respondees are more in agreement for the category of Mission Training. Being able to form teams with a mix of humans and agents, where agents act as intelligent Wingmates, is considered important for Tactical Procedure Training and Mission Training, but less important for Basic Training, where simpler scenarios are often used for training, and the main goal of the trainee is to become a proficient wingmate. The importance of Voice Communication received scores in the middle of the range and with high variance.

In Basic Training, having too Challenging Opponents may make it difficult to focus on learning how to, e.g. operate sensor and weapon systems. Instead, as noted previously, opponents may be configured to move along predefined routes, while acting according to predefined, predictable rules. For Tactical Procedure Training and Mission Training, having Challenging Opponents is essential to evaluate the performance of trainees, as well as to validate the effectiveness of developed tactics. Pilots reasoned that if agents had a learning capability, they could identify flaws in human-developed tactics and learn to exploit those flaws.

To be challenging opponents, agents need to possess similar capabilities as human pilots. Among other things, key to winning the fight is to coordinate with your teammates, achieve high time on station and to detect others while not being detected yourself. Together with teammates, synthetic pilots need to select good formations, maintain a favourable scenario geometry in relation to enemies (e.g. position, altitude, and movement) and keep enemies outside stand-off distance while the members of the team themselves move into stand-off distance. It is important to maintain pressure on the enemy and cover a lot of surface (depth and width). Synthetic pilots should also be able to learn to identify weak opponents and target them for attack in coordination with teammates. They must carefully consider when to engage an enemy based on its value and threat level so as to not take unnecessary risk or waste fuel and missiles. In a similar way, when using sensor or electronic warfare systems, emission management must be considered to balance the chance of detecting opponents while avoiding being detected by enemies.

Since pilots do not operate on their own in real-world missions, support for team training is of utmost importance, as indicated by the scores from the survey. In interviews, pilots expressed that having synthetic pilots that are intelligent enough to act as Wingmates of trainees is valuable, since it makes it possible to train as you fight even when there are not enough human pilots available to populate complex scenarios. At a minimum, self-paced training with 2-vs-2 fighters and a strike force should be supported. This requires that synthetic pilots can learn to understand the intentions of trainees as well as their own role in the mission. In basic training, the major goal of the trainee is to become a proficient wingmate, and the availability of a synthetic formation leader to support self-paced training may be of higher importance than a synthetic wingmate. Pilots also expressed that even if synthetic wingmates had human-level intelligence, it would still be important to conduct advanced tactical and mission training

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

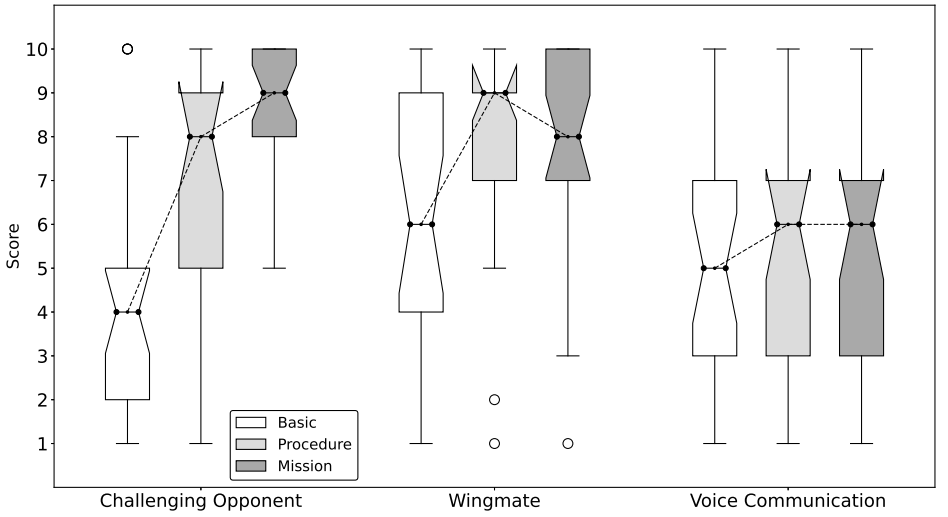

Figure 7. *Importance of different types of agent roles and voice interaction.*

in units populated by the other members of your wing, since these are the people you would cooperate with in real-world combat.

For success in air combat, it is important that the members of a unit coordinate their actions well. Therefore, some level of communication capability among human and synthetic agents may be required. The need for voice communication within mixed teams of human and synthetic agents was included in the survey since it is a rich form of communication, which in general may be challenging to realise in a believable way for synthetic agents. However, pilots reasoned that in air combat the information exchange over radio channels is often of a simple form, following a predefined protocol. Using domain knowledge makes it possible to predict what types of interaction will occur, which helps when building models for the speech understanding and speech synthesis of synthetic pilots. Pilots also argued that in many situations they know how to respond to teammates actions without communication, since the team is trained in executing coordinated manoeuvres. However, realising such a capability in a synthetic pilot may be challenging.

Populating training scenarios with mixed teams of human and synthetic agents can make training more efficient and effective. When using machine learning to build synthetic pilots, learning behaviour that supports interaction with humans is important, but also challenging, since during learning agents typically act in a simulation where no humans are present. Therefore, learning methods that result in behaviour that generalises to diverse environments and scenarios are important.

For the category of *Agent Behaviour in Training Scenarios*, the following statements were presented to respondees for rating:

- **Agent Performance:** It is important that synthetic tactical entities have realistic performance (e.g. do not always execute weapon delivery and evasive manoeuvres perfectly).
- **Element of Surprise:** It is important that there is an element of surprise in the tactical scenario (i.e. the scenario does not play out in the exact same way in each run).
- **Behaviour Explainability:** It is important that it is possible to explain the behaviour of synthetic tactical entities in debriefing sessions (e.g. why a missile was fired in a certain situation).

The aim of this category of statements was to investigate the importance of different agent characteristics in the context of a training scenario. The scores given by respondees, which indicate to what degree they agree with the statements, are presented in Fig. 8.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

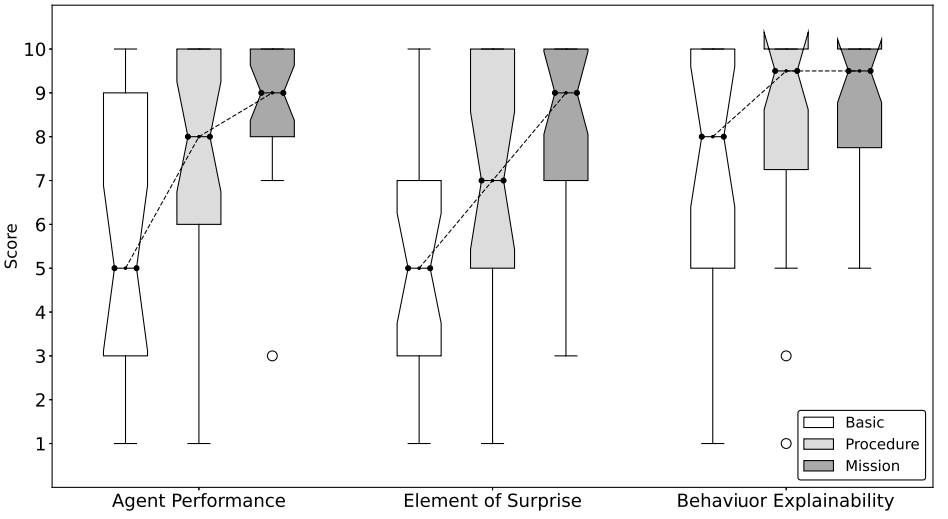

Figure 8. *Importance of different types of scenario characteristics.*

For the category of Agent Performance, we can see that there is high variance in the scores, but the median is at the lower half of the range for Basic Training. For Tactical Procedure Training and Mission Training, on the other hand, the importance of having realistic performance is high. The importance of having an Element of Surprise and variation in training scenarios increases as we move from Basic Training to Mission Training. Behaviour Explainability is of high importance in all types of training.

In interviews, pilots argued that it is important that it is possible to adjust the agents' performance to suit specific trainees and training scenarios. This imposes requirements on the algorithms used for learning the behaviour models of agents, requiring them to learn models that can be adjusted in a similar way as human role-players can be instructed how to act in a training scenario. In addition, in real-world air combat even experienced pilots will make mistakes, so it is important that this happens in training scenarios as well. As one pilot said, the learning agents should ideally be able to act as the perfect pedagogical instructor, adapting their behaviour to the current training needs of trainees. Synthetic pilots should not be perfect (e.g. performing perfect evasive manoeuvres or missile delivery); they should make similar mistakes as human opponents would, so that trainees can learn to take advantage of such mistakes. When using human role-players for training, these will try to adapt to the proficiency level of the trainees, and then sometimes make small, intentional mistakes, which the trainees are expected to exploit.

Training scenarios with variation are important for advanced tactical training, so that trainees do not simply learn how the scenario plays out each time and base their decisions on that information. Variation also makes it more difficult for trainees to exploit possible deficiencies in the behaviour models used to control synthetic pilots. The variation can be realised in different ways. As mentioned before, one way is to use stochastic instead of deterministic policies. Another, more expressive way is to introduce explicit variations in the scenario, e.g. by varying the goals of agents, by adapting the way in which agents try to achieve those goals, for instance by specifying different rules of engagement, and by varying the characteristics of the agents that populate the training scenarios, such as their proficiency, aggressiveness and level of risk-taking.

Training sessions are typically concluded with a debriefing session where the outcome of the training scenario is discussed to determine what went well, what went less well and areas for future improvement. In these sessions it is valuable if the decision-making process of the agents participating as synthetic role-players is transparent so that the decisions made at key points in the scenario can be understood by the human participants. When using traditional techniques for constructing behaviour models, e.g. scripts, state machines and behaviour trees, tools for analysis of behaviour can be constructed by extracting

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

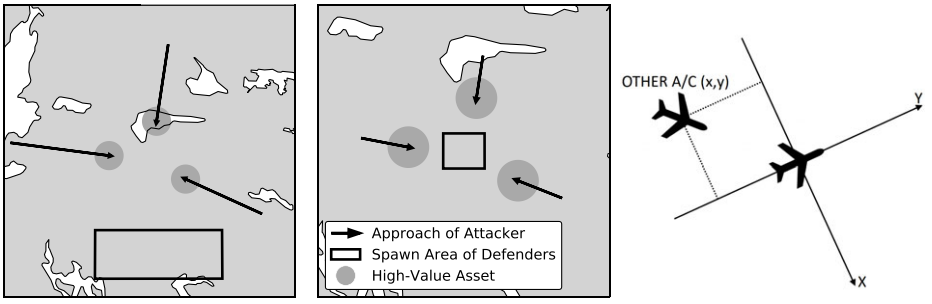

Figure 9. *Training scenario to the left, test scenario in the middle, and state space to the right.*

suitable information from those models. For learning agents that use neural networks to represent the decision-making policy, this process becomes more challenging, since neural networks are black-box models trained with data driven methods. This is true for deterministic as well as stochastic policies.

5.0 Human-Agent Interaction in an Air Policing Scenario

In this section, we present results from a practical experiment intended to illustrate aspects of human-agent interaction when using machine learning agents in an air combat scenario. In this experiment, human operators were teamed with agents trained using reinforcement learning to solve a task in an air policing scenario. The intention was to study how the mixed team of humans and agents performed on the task and how the behaviour of humans and agents differed. For these practical experiments, the participants were three experienced simulator engineers, and two of them also had extensive experience as pilots, although not as fighter pilots.

5.1 Experiment design

For the study of human-agent interaction, we used an air policing scenario developed in previous work [6] to train the agents. In this scenario, three aircraft controlled by learning agents should escort potential threats out of their air space, in order to protect three high-value assets. Incoming threats are controlled by handcrafted behaviour models, implemented using behaviour trees [33]. To escort a threat out of protected air space, an aircraft needs to fly within 5km of this threat. The challenge for the agents controlling the aircraft is to learn to allocate threats among themselves so that each aircraft can escort a threat out of protected air space. The scenario is illustrated to the left of Fig. 9. Before each episode of training, defending aircraft spawn in random positions and with random headings in the rectangle. Threats approach along the arrows towards the high-value assets illustrated by the circles.

To promote cooperation, in each time step of the simulation the learning agents received a shared reward defined as:

$$r_t = -\sum_{i=1}^{3} \min\left(\|p_{a_i} - p_{d_1}\|, \|p_{a_i} - p_{d_2}\|, \|p_{a_i} - p_{d_3}\|\right) \tag{4}$$

where p_{a_i} refers to the position of attacker i and p_{d_k} refers to the position of defender k. To maximise the shared reward, the group needs to minimise the distance between each attacker and its closest defender. The action space of an agent allowed it to fly forward, or turn left or right with a load factor of 2–4g. The observation space of the agent was defined as the relative position of all other agents, in a body-fixed coordinate system, as illustrated to the right of Fig. 9. To help an agent predict where the other agents in the scenario were going, it was given a stack of observations from the last 4 time steps in the episode as input to its policy. The agents' policies were represented by multi-layer perceptrons (MLP), with

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

2 hidden layers, each with 64 neurons and the ReLU activation function. The policy was executed at a frequency of 1Hz.

We trained the agents over 90k episodes, with each episode lasting for 600 steps, i.e. 10min. When an episode ended, each attacker was reset to a position corresponding to the beginning of its attack route in Fig. 9 and with a heading corresponding to the heading of the attack route. At the same time, the position of each defender was sampled from the spawning area in Fig. 9, while the heading was sampled from the interval $[0^\circ, 360^\circ)$. The experience of each learning agent in each time step of the simulation (action a_t taken in state s_t, and the resulting next state s_{t+1} and reward r_{t+1}) was stored in a First-In First-Out replay buffer. Periodically, every time 1024 simulation steps had been completed, the policy of each agent was updated based on data sampled from the replay buffer. We used the MADDPG algorithm [26] for learning with a learning rate of $\alpha = 10^{-2}$, a discount factor of $\gamma = 0.95$ and trained using the Adam optimiser [34].

After training, we studied transfer of learning by evaluating agents in a slightly different scenario, illustrated in the middle of Fig. 9. This scenario has a more compact geometry, which can make it more challenging to decide which defender should approach which threat. Additionally, we replaced one of the synthetic agents with a human operator to investigate how human pilots and learning agents could coordinate their actions to solve a cooperative task. Human pilots controlled their aircraft with a standard gaming joystick mounted on a desktop, observed the environment in an out-the-window (OTW) view presented on a monitor, and could also observe the positions of other entities in the scenario through a map presented on a monitor to the right of the OTW view. The length of each episode was reduced to 120 steps, and we defined the criterion for mission success as intercepting all attackers before the end of an episode.

With this scenario, we wanted to analyse how the selected design of action space, observation space, reward system and training approach affected the performance of the agent. The scenario is simple, but it can still help us identify limitations of applications of reinforcement learning in the domain of simulation-based air combat training as well as important aspects of agent design for this domain. In addition, by using a simple scenario, there will be less variation in the performance of pilots among the individual iterations of the experiment. This is useful when running experiments with humans in the loop, since for such experiments it is costly and time-consuming to run many iterations. In this work, we ran five iterations of the experiment, with starting positions and headings of defenders' aircraft generated by random for each iteration. After human pilots had completed their five iterations, they were asked to answer the following questions, with scores in the range from 1 (for worse) to 10 (for better), based on their experience:

- **Q1:** How easy was it to determine an optimal target allocation?
- **Q2:** To what extent was the synthetic agents' behaviour reasonable?
- **Q3:** How valuable do you think the following modes of interaction would be for coordination in these (and similar) scenarios?
  - **Q3.1:** Situation awareness map on the aircraft head-down display that shows, e.g. the position and combat value of other aircraft.
  - **Q3.2:** Link text message with current high-level goals (e.g. priority target) of other aircraft.
  - **Q3.3:** Speech message with current high-level goals (e.g. priority target) of other aircraft.

5.2 Results

To the left of Fig. 10 we present the mean and standard deviation for rewards received by teams with only agents compared to the rewards received by teams with a mix of agents and humans. We can see that the performance of the two categories of teams is similar, although teams with a human participant perform slightly better. From a qualitative point of view, when observing the outcome of each iteration, it could be seen that both teams were trying to split up and have one pilot approach each threat, which

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

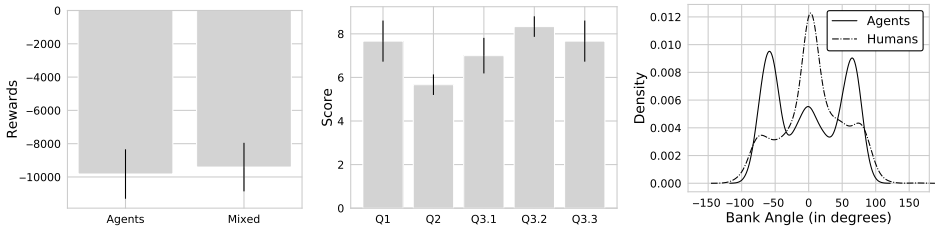

Figure 10. *Rewards to the left, survey results in the middle, bank angle distribution to the right.*

is the optimal tactic for the scenario. However, humans were better at quickly resolving conflicts when two pilots started approaching the same target. We found that mixed teams of humans and synthetic agents, as well as teams populated by only synthetic agents had a success rate of 80% over the complete experiment.

In the middle of Fig. 10 the results of the survey are presented as mean and standard deviation for the answers of the three pilots. Humans felt that in three of the five iterations it was clear how to allocate the threats among the pilots, while in two iterations it was not obvious what the optimal target allocation would be. The behaviour of agents received fair scores by the human pilots. The major complaint was that when there was a conflict in target allocation, with a human pilot and an agent approaching the same target, agents might not immediately realise this and select a new target. In the experiments with pure agent teams, it was noted that when conflicts arise agents may also have difficulties determining which agent has the most favourable position to keep pursuing the target.

Human pilots felt that the situation awareness map was valuable for coordinating with the synthetic pilots, but also reasoned that additional information presented in either text or speech messages, e.g. the targets selected for pursuit by synthetic pilots, would further simplify the task. This is most likely true for the synthetic pilots as well. To incorporate such functionality, the action and observation spaces of the learning agents could be modified. For instance, by letting each agent act by selecting which threat to engage, rather than acting by commanding the desired turn rate, information about the agent's selected target becomes available, and can be distributed over data link. This makes the agent's behaviour more explainable and transparent, and coordination with human pilots could be improved. In a similar way, by modifying the learning agent's observation space, and including information about targets that have been selected for pursuit by other pilots in the scenario, the decision-making task of the agent could be simplified.

To the right of Fig. 10 we can see the distribution of the aircraft bank angle for agents and human pilots over the iterations of the experiment. It can be seen that aircraft controlled by learning agents are turning frequently, while aircraft controlled by human pilots are more frequently flying straight. This is related to the design of the learning agents' action space as well as the design of their reward signals. The action space in this experiment is a low-level action space, with continuous actions. This makes it challenging for the agent to explore and to find the optimal action for each state. Furthermore, there is no component in the reward design that gives the agent an explicit incentive to avoid aggressive control of the aircraft. Either modifying the agent's action space, having it act using more abstract, high-level actions, or modifying the agent's reward signal to penalise aggressive manoeuvres, could help make the agent's behaviour more human-like and believable.

The results of these experiments illustrate that when designing learning agents it is important to use architectures, abstractions and training schemes that generalise over a wide range of environments, missions and adversaries. We further discuss ways of constructing agents in the next section.

6.0 Introducing Learning Agents in Training Systems

In this section, we first present an architecture for a simulation-based pilot training system, which automates parts of training adaptation and training delivery by incorporating learning agents. Based on this

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

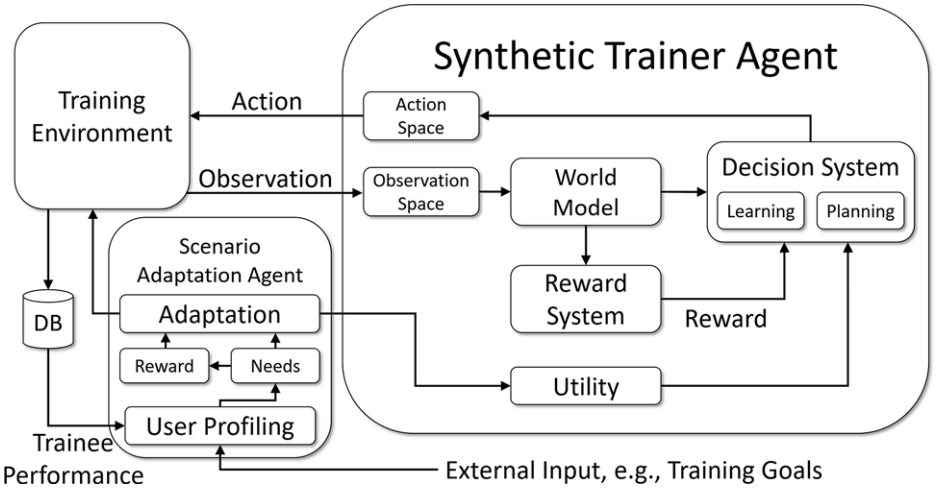

Figure 11. *System architecture for training system using learning agents.*

architecture, and the results of our domain analysis, user study and practical experiments, we then discuss design approaches and solution concepts that could produce agents with desirable capabilities and characteristics. The domain analysis identified perception, decision-making and actions at different levels of abstraction in an air combat training scenario. The user study discussed how to meet user requirements and revealed that it is desirable that agent behaviour can be adapted to fit the training needs of novice as well as experienced pilots, that the learning algorithms used allow us to let agents use standard doctrinal manoeuvres as well as learn novel tactics to challenge trainees and that the behaviour learned should be believable and explainable. Finally, the practical experiments illustrated how the chosen design of the observation space, action space and reward signal can affect the emergent behaviour of a learning agent.

6.1 System architecture for training systems using learning agents

A system architecture for a simulation-based pilot training system, which incorporates learning agents for improved efficiency and effectiveness, is illustrated in Fig. 11. The architecture is an extended and modified version of an architecture proposed in previous work [6, 35]. In this work we will discuss how to realise the components of the architecture, based on the findings of our domain analysis, user study and experiments.

The architecture integrates agents to support organisations and instructors in adapting training to trainees' training needs (Scenario Adaptation Agent) and delivering it in an efficient manner (Synthetic Trainer Agent). In training scenarios, Synthetic Trainer Agents participate as synthetic role-players with the same purpose as human role-players. These agents act in one of the roles of the scenario, either as opponents or teammates of the trainee pilots. Their major goal is to provide a stimulating training environment to the trainees. Offline, between training sessions, the Scenario Adaptation Agent takes the role of an instructor and analyses data generated in past sessions to identify trainees' weaknesses and strengths and then adapts training session contents to maximise future improvement of proficiency.

The goals of the Synthetic Trainer Agent are modelled through its Reward System, which captures important features for successful decision-making in air combat scenarios. An individual agent's preferences among reward features are determined by the agent's Utility function. The Synthetic Trainer Agent's ability to perceive the state of its environment is essential for decision-making in the complex domain of air combat. The agent's perception is formed by two components: the agent's interface to the fused information of the aircraft's sensors (Observation Space) in combination with the agent's internal beliefs regarding the state of the world, based on its past observations (World Model). The Decision System of the Synthetic Trainer Agent realises its capability to learn how to act (using the actions of

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

its Action Space) based on past experience as well as its capability to evaluate and plan future actions based on its learned understanding of the dynamics of its environment, e.g. its ability to predict future behaviour of other agents. Agents acting as synthetic role-players enter the training environment through computer generated forces software. Agents are first trained offline and then interact with human trainees in training sessions, which can potentially provide data that support further adaptation of the agents' decision-making policies.

Just like the Synthetic Trainer Agent, the Scenario Adaptation Agent learns its policy through reinforcement learning. The state considered when making decisions consists of inferred User Needs to meet Training Goals, based on User Profiling. The actions of the agent are related to Adaptation of the contents of training sessions, and the rewards received are related to the changes in Trainee Performance over time.

We provide further discussions regarding these components, and ways of realising them, in the sections below.

6.2 Reward system and utility function

For reinforcement learning agents, the goals that should be achieved are expressed through a reward signal. In the proposed architecture, the reward signal is generated by each agent's internal reward system, based on current and past states of the world, according to the agent's perception. In a standard MDP the reward signal is a scalar. To model the conflicting objectives of air combat training, we instead use a Multi-Objective MDP (MOMDP) [36], which provides vector-valued rewards. The values of the reward vector are based on important features of the scenario, which should affect the agent's decision-making over time. For instance, rewards could be given for achieving advantageous geometry relative opponents, for detecting other entities while avoiding being detected, for missile hits, and for sensible resource management. This information supports control at level 4 of the LACC. The overall value of states and actions are determined by applying the agent's utility function U over the vector returns:

$$u = U(V_{\pi}(s)) \tag{5}$$

This gives a scalar value u that supports ordering of policies.

The reward system design is one option for infusing domain knowledge in the behaviour of the learning agent, to bias the learning process towards desirable characteristics, e.g. making an agent learn behaviour which is in line with a certain doctrine. By using different combinations of reward components and utility functions, a diverse set of agents can be created, which can make training more varied and stimulating. The components of the reward vector represent the objectives of the agent, and finding optimal policies results in a multi-objective optimisation problem.

One way to construct the reward system is to let a human pilot demonstrate how to solve a certain task, and then inferring a reward signal from this information, or simply rewarding the agent for behaving in a similar way as the human pilot [37–39]. Demonstrations are often used to support training of human pilots, so this approach leads to a human-machine interface that feels natural for the instructor and reduces the need for explicit programming of agents.

Since dense reward systems, which give frequent feedback to the agent, as well as rewards based on demonstrations, introduce a bias in the agent's policy, they may prevent the agent from finding an optimal policy. To create very challenging opponents it may instead be desirable to use sparse reward signals, e.g. only rewarding the agents for winning a fight according to some metric. This allows agents to freely explore the world, and novel tactics and doctrines may emerge.

Agents that are to act as synthetic trainers for humans can not only consider components of the tactical scenario, they must also include information about the trainees' learning needs in their reward system, so that the decisions made during a training session are based on reasoning about training effect at level 5 in the LACC. This includes observing the trainees' proficiency in different aspects of air combat, and identifying ways of giving the trainees the right stimulation to improve their proficiency over time. This process is supported by analysis of data from past training sessions offline, through the Scenario

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

Adaptation Agent's profiling of trainees, and inference of training needs to meet the organisation's training goals. Agent behaviour is then further adjusted during the progression of a training session, in a similar way as human role-players would adapt their behaviour to the performance of trainees.

6.3 Observation space and world model

The design of the agent's observation space determines which features of the environment will be considered when making decisions. As illustrated in our analysis of decision-making in air combat scenarios, human pilots use low level features (level 1 of the LACC) as well as more abstract value-based information (level 4 of the LACC) to support their decision-making. For efficient learning of policies, agents should be supported by similar information. This includes knowledge about the performance of own and opponents' vehicles, sensors, and weapon systems, for example, which human pilots would have acquired in theoretical study.

To realise intelligent behaviour, agents can not act only based on the immediate observation of the world, but must instead consider its whole history of observations. This functionality is realised by the agent's world model, which uses memory mechanisms to learn an abstract model of the state of the world, which can support decision-making. The model can be infused with domain knowledge by explicitly modelling such features that human pilots believe are important for success in air combat, for instance predictions regarding other agents' goals, beliefs and future behaviour. To support adaptation of behaviour to trainees' current training needs, the world model should also provide abstract information related to training effect, e.g. estimates of trainees' proficiency. The functionality of the world model enables the agent to frame the current situation, and to reason about the effect of its actions (levels 5 and 6 of the LACC).

One challenge for learning agents is that they typically learn their policies in an environment populated exclusively by other synthetic agents, i.e. they do not interact with humans. The reason for this is the large number of iterations required for learning algorithms to converge. As identified in our user study, agents that act as team members of trainees must learn to understand trainees' intentions, while agents that act as opponents should be able to identify and exploit flaws in human-developed tactics, which may change over time. Therefore, for agents to interact effectively with humans in training sessions, they need to have the capability of adapting their behaviour to a wide range of teammates and opponents. One way to achieve this is to maintain a diverse population of agents while learning new policies and to assemble teams of agents by random sampling from this population before each episode of learning [9]. Since this forces each agent to learn while interacting with several different types of other agents, the agent will hopefully learn a policy that is general enough to support interaction with humans as well. Another approach is to use meta-learning, where agents learn to model characteristics and behaviour of other agents based on few observations [40]. In meta-learning, an agent learns a meta-model of common properties of other agents in the environment, as well as an ability to use this model to make predictions regarding the characteristics and actions of an agent that it encounters for the first time. Provided that human behaviour is not to different from that of synthetic agents, a modelling capability learned from interactions with synthetic agents could be used for modelling of humans as well. During training sessions, such an approach could be used as a basis for modelling a specific human trainee.

6.4 Action space and decision system

The design of the agent's action space has great impact on its ability to explore, and will affect its final learned behaviour. For air combat simulation, it may be desirable to constrain the behaviour of the agent so that it resembles a certain opposing force. By using parametric action spaces, actions can be made available for selection only when certain conditions are fulfilled. Such approaches have been used to make sure that learning agents abide to the rules of games [8, 10]. In air combat simulation, for instance, rules of engagement can be encoded in the action space to restrict when and how target engagement is allowed.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

By including temporally extended actions in the design of the agent's action space, it becomes possible to learn a policy over actions at level 3 of the LACC. The options framework for hierarchical reinforcement learning provides a formalism for learning with temporal abstractions [41]. An option $\omega \in \Omega$ is defined as a tuple $(I_\omega, \pi_\omega, \beta_\omega)$, where:

- Ω is the set of available options
- I_ω is the initiation set, specifying in which states the option can be selected
- π_ω is the intra-option policy, i.e. the policy used once the option has been selected
- β_ω is the termination condition of the option, specifying the probability of the active option terminating in a state, to allow a new option to be selected

One benefit of temporal abstractions and hierarchical reinforcement learning is that agents' policies can become easier to understand [42]. Another benefit is that the performance of learning can be improved when a problem is broken down into a set of sub-problems, which are then dealt with in a decision-making hierarchy.

The options used for air combat simulation could be handcrafted to replicate how temporally extended actions are executed according to a nation's air combat doctrine. This is a natural approach when the designer has a clear idea about how an extended action should be performed, e.g. actions that have been optimised based on the laws of physics, such as missile guidance. Using handcrafted options as building block for learning tactics is also more likely to result in behaviour that is believable to humans than learning with low level actions alone, which can sometimes have undesirable effects, as illustrated in our practical experiments. For areas where there is greater uncertainty regarding how the agent should act to solve a task, there are also algorithms that make it possible to learn hierarchical policies from scratch, e.g. the option-critic architecture [43], the double actor-critic [44] and feudal reinforcement learning [45], which makes it possible to discover complex, novel forms of actions.

As noted in our discussion on reward systems and utility functions, designing air combat policies is a multi-objective optimisation problem. Multi-objective reinforcement learning (MORL) provides systematic methods for learning sets of policies that are Pareto optimal, meaning that for at least one objective there is no policy that gives higher return [36, 46, 47]. We believe that this is a natural approach for reinforcement learning in the air combat domain, where trade-offs between conflicting objectives are often required. The method supports decision-making at level 4 of the LACC. To adjust training to fit the needs of individual trainees, suitable agent policies can be selected from the set of Pareto optimal policies [6, 28, 48].

In MORL, there are two types of optimisation criteria that are used when learning policies, scalarised expected returns (SER) and expected scalarised returns (ESR):

$$V_u^\pi(s) = U\Big(E\Big[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\Big]\Big), SER \tag{6}$$

$$V_u^\pi(s) = E\Big[U\Big(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\Big) | s_t = s\Big], ESR \tag{7}$$

The SER criterion aims to optimise the average outcome of several episodes, while the ESR criterion tries to optimise the average outcome of each episode. For air combat training, the ESR criterion may be the most suitable one, since from a safety perspective pilots want to optimise their chances of survival in each mission, rather than their expected survival rate over a complete campaign. We can see that for linear utility functions the criteria are the same, since the positions of expectation and utility functions can be interchanged. However, we argue that the utility function of a fighter pilot is not a linear function. For instance, safety may be considered infinitely more important than other objectives up until a certain probability of survival, resulting in a utility function with non-linear thresholds.

To further adapt agents' behaviour to fit trainees' needs, planning algorithms can be used to adjust the agents' policies online while a training session is in progress. These algorithms use the agent's world

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

model to do simulated rollouts to explore the effects future actions would have on the outcome of the mission. One family of planning algorithms that has had great success in, e.g. games of various forms is monte-carlo tree search (MCTS) [49]. MCTS can be combined with the learned value functions of the agent to improve its performance [7, 8]. In training scenarios, planning could be used to adapt behaviour to maximise the current utility of the agent, which is related to the training effect of trainees, and level 5 of the LACC.

6.5 Adapting agent behaviour to inferred training needs

To support adaptation of simulation contents and agent characteristics to current training needs, the Scenario Adaptation Agent should learn a model of different aspects of trainees' proficiency, based on their performance in past training sessions. Performance measurements can include, e.g. measurements describing a pilot's flight path, risk exposure, resource management and success rate in engagements in missions. The model is used as input to the agents that participate as synthetic role-players in training sessions and affects their utility for different types of behaviour to achieve maximum training effect. This corresponds to having agents with a capability of perception and decision-making at level 5 of the LACC, which is, of course, a highly challenging task. However, recent advances in agent modelling using machine learning techniques have shown promising results.

In addition to supporting adaptation of agent behaviour, the Scenario Adaptation Agent should also be able to adjust the contents of training scenarios, so that suitable components for improving trainees' performance are included. Here there is an overlap between human training and training of synthetic agent's behaviour. For efficient learning, synthetic agents should be exposed to increasingly challenging problems at a rate determined by their rate of improvement. In reinforcement learning, this is called curriculum learning [50, 51]. It is possible that curriculum learning techniques that have proven effective for training of synthetic agents could be adapted for training of humans as well, but further research on the topic is required.

6.6 Training environments

The training environment is where agents acting as synthetic role-players interact with human trainees. Here, it is desirable to have high-fidelity models for the vehicles operated by the actors in the simulated scenario, as well as environment properties of various sorts, e.g. weather effects. However, when learning agent behaviour using current state of the art reinforcement learning techniques, many iterations of missions are required, leading to long simulation times if a complex simulator is used. For this reason, it is valuable to have the possibility to adjust the fidelity of the simulation in several steps. The initial learning can then take place in lower fidelity environments for many iterations, and the learned policies can then be successively transferred to environments with higher fidelity models for fine-tuning.

For our evaluations of learning agents, we use simulations of varying complexity and fidelity. Concepts are developed and analysed in desktop simulations with simple scenarios and then further developed for integration in the target environment. The target environment considered in this work is a high-fidelity tactical simulation used for training of fighter pilots that operate the Saab Gripen aircraft. In this environment, evaluations with multiple manned stations can be performed. In operational training, the simulations are intended for future use in ground-based simulators as well as embedded training solutions in the aircraft, which itself is integrated in a distributed simulation network using a data link.

7.0 Conclusion

In this work, we studied introduction of intelligent, learning agents in simulation-based pilot training systems from the users' perspective. We analysed how agent technologies relate to constraints imposed on actors in training systems, and what decision-making patterns should be supported by agent designs.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

Through interviews, a survey and practical experiments we learned about requirements on agent capabilities and characteristics, challenges and shortcomings of current agent technologies and aspects of human-agent interaction with agents constructed using state of the art reinforcement learning techniques. Finally, we discussed design approaches and solution concepts for a training system architecture that integrates learning agents.

We conclude that the ongoing revolution in artificial intelligence is providing great opportunities for improvement of training efficiency and effectiveness. While our focus in this paper was on military training, many of the discussed concepts have broader applicability, e.g. for training simulations of other sorts, including training of pilots for civilian flight, where agents could help realise dense air traffic and patterns of life, automated setting of adversarial weather conditions and malfunctions, as well as automated evaluation and profiling of trainees.

For future research, we recommend further development of learning agents from the users' perspective, to steer progress in the most valuable direction. Real-world air combat scenarios provide many challenges to agents, e.g. cooperation and competition in scenarios with many agents, decision-making under partial observability and uncertainty, and the need to prioritise among multiple conflicting objectives, such as tactical mission goals, resource consumption and safety. Therefore, the domain of air combat training is an excellent benchmark for reinforcement learning algorithms, and there are many exciting directions of research left to explore.

Acknowledgements. This work was partially supported by the Swedish Governmental Agency for Innovation Systems (grant NFFP7/2017-04885), and the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. This work was supported by computation resources provided by the Swedish National Infrastructure for Computing (SNIC) at Tetralith/NSC partially funded by the Swedish Research Council through grant agreement no. 2020/5-230.

References

[1] Page, E.H. and Smith, R. Introduction to military training simulation: a guide for discrete event simulationists, *Proceedings of the Winter Simulation Conference*, 1998, vol. 1, pp 53–60.

[2] Roessingh, J.J. and Verhaaf, G.G. Training effectiveness of embedded training in a (multi-) fighter environment, Tech Rep, National Aerospace Lab Amsterdam (Netherlands), 2009.

[3] Gilbert, N. *Agent-based Models*, Sage Publications, Incorporated, 2019.

[4] Sutton, R.S. and Barto, A.G. *Reinforcement Learning: An Introduction*, MIT Press, 2018.

[5] Goodfellow, I., Bengio, Y. and Courville, A. *Deep Learning*, MIT Press, 2016.

[6] Källström, J. and Heintz, F. Multi-agent multi-objective deep reinforcement learning for efficient and effective pilot training, *Proceedings of the 10th Aerospace Technology Congress*, October 2019, pp 101–111.

[7] Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D. Mastering the game of Go with deep neural networks and tree search, *Nature*, 2016, **529**, (7587), pp 484–489.

[8] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. and Hassabis, D. Mastering the game of go without human knowledge, *Nature*, 2017, **550**, (7676), pp 354–359.

[9] Jaderberg, M., Czarnecki, W.M., Dunning, I., Marris, L., Lever, G., Castaneda, A.G., Beattie, C., Rabinowitz, N.C., Morcos, A.S., Ruderman, A., Sonnerat, N., Green, T., Deason, L., Leibo, J.Z., Silver, D., Hassabis, D., Kavukcuoglu, K. and Graepel, T. Human-level performance in 3D multiplayer games with population-based reinforcement learning, *Science*, 2019, **364**, (6443), pp 859–865.

[10] Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J.P., Jaderberg, M., Vezhnevets, A.S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T.L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C. and Silver, D. Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, 2019, **575**, (7782), pp 350–354.

[11] Lundberg, J. and Johansson, B. A framework for describing interaction between human operators and autonomous, automated, and manual control systems, *Cognition, Technology & Work*, 2020, pp 1–21.

[12] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D. Continuous control with deep reinforcement learning, *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, May 2016.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

[13] Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P. and Zaremba, W. Hindsight experience replay, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp 5055–5065.
[14] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347, 2017.
[15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D. Human-level control through deep reinforcement learning, *Nature*, 2015, **518**, (7540), pp 529–533.
[16] Watkins, C.J.C.H., and Dayan, P. Q-learning, *Mach. Learn.*, 1992, **8**, (3–4), pp 279–292.
[17] Rijken, R. and Toubman, A. The future of autonomous air combat behavior, *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp 3089–3094.
[18] Liu, P. and Ma, Y. A deep reinforcement learning based intelligent decision method for UCAV air combat, *Asian Simulation Conference*, 2017, pp 274–286.
[19] Ma, X., Xia, L. and Zhao, Q. Air-combat strategy using deep Q-learning, *Proceedings of the Chinese Automation Congress (CAC)*, 2018, pp 3952–3957.
[20] Zhang, X., Liu, G., Yang, C. and Wu, J. Research on air confrontation maneuver decision-making method based on reinforcement learning, *Electronics*, 2018, **7**, (11), pp 1–19.
[21] Yang, Q., Zhu, Y., Zhang, J., Qiao, S. and Liu, J. UAV air combat autonomous maneuver decision based on DDPG algorithm, *Proceedings of the IEEE 15th International Conference on Control and Automation (ICCA)*, 2019, pp 37–42.
[22] Kong, W., Zhou, D., Yang, Z., Zhao, Y. and Zhang, K. UAV autonomous aerial combat maneuver strategy generation with observation error based on state-adversarial deep deterministic policy gradient and inverse reinforcement learning, *Electronics*, 2020, **9**, (7), pp 1–24.
[23] Hu, Z., Wan, K., Gao, X., Zhai, Y. and Wang, Q. Deep reinforcement learning approach with multiple experience pools for UAV's autonomous motion planning in complex unknown environments, *Sensors*, 2020, **20**, (7), pp 1–21.
[24] Li, B., Gan, Z., Chen, D. and Aleksandrovich, S.D. UAV maneuvering target tracking in uncertain environments based on deep reinforcement learning and meta-learning, *Remote Sensing*, 2020, **12**, (22), pp 1–20.
[25] Wan, K., Gao, X., Hu, Z. and Wu, G Robust motion control for UAV in dynamic uncertain environments using deep reinforcement learning, *Remote Sens.*, 2020, **12**, (4), pp 1–21.
[26] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P. and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments, *Advances in Neural Information Processing Systems*, 2017, pp 6379–6390.
[27] Zhang, G., Li, Y., Xu, X. and Dai, H. Efficient training techniques for multi-agent reinforcement learning in combat tasks, *IEEE Access*, 2019, **7**, pp 109301–109310.
[28] Källström, J. and Heintz, F. Agent coordination in air combat simulation using multi-agent deep reinforcement learning, *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp 2157–2164.
[29] Xiao, J., Wang, G., Zhang, Y. and Cheng, L. A distributed multi-agent dynamic area coverage algorithm based on reinforcement learning, *IEEE Access*, 2020, **8**, pp 33511–33521.
[30] Kong, W., Zhou, D., Yang, Z., Zhang, K. and Zeng, L. Maneuver strategy generation of UCAV for within visual range air combat based on multi-agent reinforcement learning and target position prediction, *Appl. Sci.*, 2020, **10**, (15), pp 1–23.
[31] Matplotlib matplotlib.pyplot.boxplot, https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.boxplot.html.
[32] McGill, R., Tukey, J.W. and Larsen, W.A. Variations of box plots, *Am. Stat.*, 1978, **32**, (1), pp 12–16.
[33] Colledanchise, M. and Ögren, P. *Behavior Trees in Robotics and AI: An Introduction*, CRC Press, 2018.
[34] Kingma, D.P. and Ba, J. Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[35] Källström, J. Adaptive agent-based simulation for individualized training, *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2020, pp 2193–2195.
[36] Roijers, D.M., Vamplew, P., Whiteson, S. and Dazeley, R. A survey of multi-objective sequential decision-making, *J. Artif. Intell. Res.*, 2013, **48**, pp 67–113.
[37] Ng, A.Y. and Russell, S.J. Algorithms for inverse reinforcement learning, *ICML*, 2000, **1**, pp 1–8.
[38] Abbeel, P. and Ng, A.Y. Apprenticeship learning via inverse reinforcement learning, *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004, pp 1–8.
[39] Ho, J. and Ermon, S. Generative adversarial imitation learning, *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp 4572–4580.
[40] Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.A. and Botvinick, M. Machine theory of mind, *Proceedings of the International Conference on Machine Learning*, 2018, pp 4218–4227.
[41] Sutton, R.S., Precup, D. and Singh, S. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning, *Artif. Intell.*, 1999, **112**, (1–2), pp 181–211.
[42] Smith, M., Hoof, H. and Pineau, J. An inference-based policy gradient method for learning options, *Proceedings of the International Conference on Machine Learning*, 2018, pp 4703–4712.
[43] Bacon, P.L., Harb, J. and Precup, D. The option-critic architecture, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp 1726–1734.
[44] Zhang, S. and Whiteson, S. DAC: The double actor-critic architecture for learning options, *Advances in Neural Information Processing Systems*, vol. **32**, 2019.
[45] Vezhnevets, A.S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D. and Kavukcuoglu, K. Feudal networks for hierarchical reinforcement learning, *Proceedings of the International Conference on Machine Learning*, 2017, pp 3540–3549.

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press

[46] Hayes, C.F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L.M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A.A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P. and Roijers, D.M. A Practical Guide to Multi-Objective Reinforcement Learning and Planning, arXiv preprint arXiv:2103.09568, 2021.

[47] Rădulescu, R., Mannion, P., Roijers, D.M. and Nowé, A. Multi-objective multi-agent decision making: a utility-based analysis and survey, *Auton. Agents Multi-Agent Syst.*, 2020, **34**, (1), pp 1–52.

[48] Källström, J. and Heintz, F. Tunable dynamics in agent-based simulation using multi-objective reinforcement learning, *Proceeding of the Adaptive and Learning Agents Workshop (ALA-19) at AAMAS*, May 2019, pp 1–7.

[49] Browne, C.B., Powley, E., Whitehouse, D., Lucas, S.M., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S. and Colton, S. A survey of monte carlo tree search methods, *IEEE Trans. Comput. Intell. AI Games*, 2012, **4**, (1), pp 1–43.

[50] Florensa, C., Held, D., Wulfmeier, M., Zhang, M. and Abbeel, P. Reverse curriculum generation for reinforcement learning, *Proceedings of the Conference on Robot Learning*, 2017, pp 482–495.

[51] Czarnecki, W., Jayakumar, S., Jaderberg, M., Hasenclever, L., Teh, Y.W., Heess, N., Osindero, S. and Pascanu, R. Mix & match agent curricula for reinforcement learning, *Proceedings of the International Conference on Machine Learning*, 2018, pp 1087–1095.

Cite this article: Källström J., Granlund R. and Heintz F. (2022). Design of simulation-based pilot training systems using machine learning agents. *The Aeronautical Journal*, **126**, 907–931. https://doi.org/10.1017/aer.2022.8

https://doi.org/10.1017/aer.2022.8 Published online by Cambridge University Press