Mastering Air Combat: Using Neural Fields for Self-Improving AI Agents

Orlando Avila-García, Joaquín Fernández León, Germán Pescador Barreto, Javier Rodríguez Vázquez
ARQUIMEA Research Center
Santa Cruz de Tenerife, Spain
oavila@arquimea.com, jfleon@arquimea.com, gapescador@arquimea.com, fjrodriguez@arquimea.com

George Hellstern, Rachael Shudde
Lockheed Martin Aeronautics Company
Fort Worth, Texas, USA
george.hellstern@lmco.com, rachael.shudde@lmco.com

ABSTRACT

Under the Department of Defense (DoD) Collaborative Combat Aircraft (CCA) effort, defense services are using reinforcement learning (RL) agents to develop autonomous capabilities for air-to-air engagements. Given the billions of dollars invested by DoD in artificial intelligence (AI) solutions for the warfighter, it is imperative to develop models for AI systems that allow agents to perform reliably and consistently. Current AI constructs leave out methodologies for deep neural-network-based systems to self-audit and review past performance in a self-critical way. Doing this is necessary for developing actions appropriate for increasingly difficult scenarios. Neural fields (NF) represent a capacity to replicate worlds, within which AI systems can review, rehash and grade past actions to create performance improvements. NFs represent a class of deep learning techniques to create compact, coordinate-based representations of complex signals gathered by conventional sensors.

This paper examines the application of NFs to mastering air combat by means of self-improving RL agents. We show how a trustworthy self-improvement process can emerge from the interplay among different AI capabilities using a common deep episodic memory (DEM) building block. Such a subsystem allows for the efficient and fast encoding, storage and recall of past experiences. In this paper, we introduce NF models (NFM) to encode high-fidelity representations of spatially aware observations (sensory inputs) from individual episodes. Also, we propose mapping these episodes into a latent space, where the distance between episodes relates to the similarity between sensorial experiences. We propose a blueprint for the two-layered architecture that we call the Neural Fields-based Hierarchical Episodic Memory (NF-HEM). NF-HEM is a fundamental building block for trustworthy self-improving agents, enabling different self-learning capabilities and safeguards for reliably removing the human from the development loop.

ABOUT THE AUTHORS

Orlando Avila-García holds a bachelor's degree in computer science from the University of La Laguna (Spain) and a Ph.D. in computer science–AI from the University of Hertfordshire (UK). He has over 20 years' experience in AI research and innovation. He has co-founded two startups. He is now the head of AI at ARQUIMEA Research Center. His research interests span AI research and engineering, with particular interest in AI safety engineering and trustworthiness.

Joaquín Fernández León holds a bachelor's degree in computer science from the University of Granada (Spain), a M.Phil. in data science from the Universitat Oberta de Catalunya (Spain) and a Ph.D. in computer science–AI from the Polytechnic University of Madrid (Spain). He currently works as team leader at ARQUIMEA Research Center, and his postdoctoral research there focuses on safe autonomy.

George Hellstern has over 30 years' experience with systems design, including AI solutions for air-to-air combat and sustainment. He is a program manager for autonomy and AI, uncrewed air systems command and control, and human performance. Previous experience includes operational, programmatic and technical experience from the Air Mobility Command, the Office of the Secretary of Defense and Lockheed Martin Aeronautics Company's Advanced Development Programs organization, informally known as Skunk Works®.

Rachael Shudde has a bachelor's degree in math and computer science and a Ph.D. in statistics. Her work at Lockheed Martin focuses on algorithm design for autonomous vehicles and developing methods of reliability to allow pilots to interact with autonomous and AI systems.

Germán Pescador Barreto has a bachelor's degree in computer science from the University of La Laguna (Spain) and a master's degree in AI, shape recognition and digital imaging from the Polytechnic University of Valencia (Spain). He works in ARQUIMEA Research Center as a machine learning engineer. His main research interest is deep RL (DRL).

Javier Rodríguez Vázquez holds a bachelor's degree in computer science from the University of Cádiz (Spain) and a Ph.D. in AI from the Technical University of Madrid (Spain). He has a robust background in onboard visual perception for robotic systems, utilizing deep learning techniques with a particular emphasis on semi-supervised settings. As a postdoctoral researcher at ARQUIMEA Research Center, he focuses his work on developing neural methods for detecting anomalies and novelty in visual perception systems.

Mastering Air Combat: Using Neural Fields for Self-Improving AI Agents

Orlando Avila-García, Joaquín Fernández León, Germán Pescador Barreto, Javier Rodríguez
ARQUIMEA Research Center
Santa Cruz de Tenerife, Spain
oavila@arquimea.com, jfleon@arquimea.com, gapescador@arquimea.com, fjrodriguez@arquimea.com

George Hellstern, Rachael Shudde
Lockheed Martin Aeronautics Company
Fort Worth, Texas, USA
george.hellstern@lmco.com, rachael.shudde@lmco.com

INTRODUCTION

DoD-funded efforts lay a firm foundation for CCA platforms to fly alongside crewed platforms in what is known as crewed-uncrewed teaming scenarios. The X-62 VISTA aircraft now flies with Open Mission Systems-compliant architectures, supporting aircraft control by onboard AI agents (Cotting et al., 2023). Under such programs as the Defense Advanced Research Projects Agency's (DARPA) Air Combat Evolution (ACE), air systems are being developed to leverage AI to create air intercept logic superior to human capabilities (DeMay et al., 2022). DoD investments and AI agents' dominance in air-to-air within-visual-range encounters (Plaks, 2021) have been drivers in aerospace and defense. The defense industry has pushed forward with developing aircraft capable of hosting AI in tactical scenarios, as well as RL agents with autonomous capabilities, to fluidly execute air-to-air engagements.

AI's success in challenging environments depends on strict performance parameters, dependable behaviors and self-optimization. This was demonstrated in the 2020 AlphaDogfight Trials (ADT), in which AI began to master basic flight maneuvers and outperform human pilots. In RL training, an episode is a sequence of states, actions and rewards that begins with an initial state and ends when reaching a terminal state. This terminal state can be a predefined condition, such as a goal, a completed task or exceeding a time limit. An episode represents a complete run of the agent's interaction with the environment, from start to finish. By analyzing these episodes, the agent can adapt its strategy to enhance performance gradually. Training involves many episodes to ensure that the agent has enough experience to learn effectively.

Despite RL agents' previous success in combat autonomy, the human effort required to engineer such safety-critical systems remains enormous. Human experts, from experienced pilots to engineers and scientists, need to collaborate carefully to design training environments, scenarios and settings that ensure effectiveness and trustworthiness in agent behaviors. The goal of *self-improving* agents is to minimize human involvement in the RL agent development life cycle. This means the system could learn to successfully overcome RL challenges, such as sparse rewards, sample efficiency and the exploration-exploitation dilemma, independently. This transition shifts the role of humans from active participants in the RL agent's improvement (i.e., human in the loop) to supervisors of the process (i.e., human on the improving loop). In such an arrangement, agents autonomously self-optimize their training conditions throughout their life cycle.

In this paper, we propose a specific type of action-conditioned, spatially aware episodic memory to enhance the self-improvement of RL agents throughout their life cycle, spanning from simulation environments to full-scale aircraft. We introduce NF-HEM as a two-way self-improvement process, merging the concept of hierarchical episodic memory (HEM) with NFs to create an innovative, implicit, neural-representative technique. We propose using NFs as a lower-level episodic memory to quickly and efficiently reconstruct the episode's observation space, encoding the episode's observations in a continuous implicit neural representation (INR). HEM is compatible with current RL algorithms and can work to remove direct human intervention in the development of RL systems. It achieves this by supporting experience replay, reward shaping (loss function design) and exploration.

The organization of the paper is as follows. First, we discuss self-improvement in more detail. Next, we discuss the concept of episodic memory and how it can be paired with RL agents to improve outcomes. Then, we discuss combing NFs and HEM to create NF-HEM. We end with a discussion on how NF-HEM could impact the ADT scenario.

SELF-IMPROVING ARTIFICIAL INTELLIGENCE

Self-improving agents represent an advanced class of autonomous systems capable of learning and enhancing their performance through interaction with their environment with minimal human intervention. They differ from traditional AI systems in that those systems rely heavily on predefined programming and extensive human intervention in such tasks as reward function design and environment resetting. Self-improving RL agents, by contrast, have more control — beyond the conventional RL cycle — to span the learning process. This approach allows RL agents to continuously gather and use data during learning, resulting in more effective, reliable and safe performance in uncertain and dynamic environments.

Although the term "self-improvement" is present in AI literature, a formal, widely accepted definition is still lacking. In this paper, we draw inspiration from the three main components proposed by Zhu et al. (2020): learning without resets, learning directly from the agent's sensor information, and learning from rewards inferred by the RL framework itself. We conceive of self-improvement as a continuous process with several key stages, integrating experiences between simulation and real-world environments. Before introducing the NF-HEM that enables such a process, we hypothesize a self-improvement process for agents: experiences from simulation are used during operational events, and experiences from operational events are used in training. This bidirectional flow of episodic experiences, diagramed in Figure 1, allows agents to continuously refine policies and adapt to new challenges. This ensures faster and efficient convergence to safe and effective behaviors in dynamic environments.

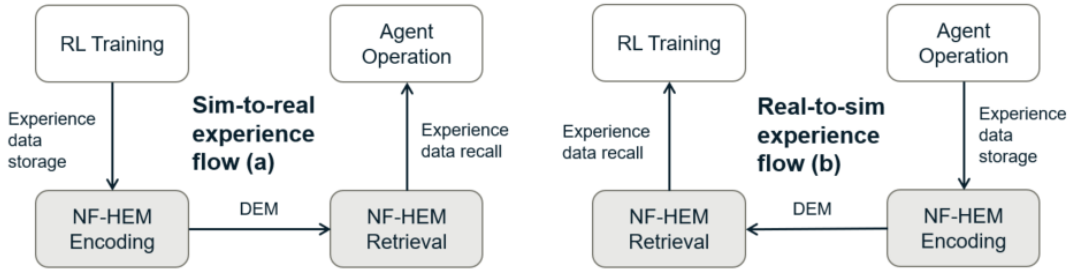

Figure 1. Two-Way Self-Improvement Process Using NF-HEM. *Experiences flow from simulation to agent operation as shown on the left, and from agent operation to training as shown on the right.*

Each experience flow serves different self-improving purposes across sensing, acting and learning. The following five capabilities can be created using sim-to-real self-improvement. The first is novelty detection, which involves detecting unfamiliar scenarios (e.g., agent out of its domain of competence) to enhance safety. The second is identifying repetition, through which familiar scenarios are recognized to optimize exploration. The third is virtual sensing, which accesses past sensory details not currently in perceptual reach. The fourth is safe exploration, which utilizes episodic memory to predict success/failure outcomes to explore safety. The fifth is managing goals, which results in keeping the long-term goal of the current episode in focus without overburdening working memory.

Going the other way, the real-to-sim episodic experience creates five capabilities of its own. The first is action modeling, which recalls and analyzes specific episodes to understand action effects. The second is environment modeling, which uses sequences of episodes to predict environmental changes and simulate them during training. The third is self-evaluation, for which live experiences are analyzed later to enhance understanding and self-evaluation (i.e., retroactive learning). The fourth is explicability, which entails replaying past behavior for explanation with humans and knowledge sharing. The fifth is experience reply, which enhances the reply buffer mechanism by optimally self-managing the most valuable episodes from which to learn.

By integrating these stages and leveraging the flows of experience between simulation and real-world environments, self-improving agents can achieve higher levels of autonomy, efficiency and adaptability. At the same time, they can support prioritizing trustworthy AI capabilities (robustness, safety and explainability) as the first consideration within the architecture. This dual-context approach ensures that RL agents are well prepared to handle the complexities of real-world applications with minimal human intervention. Moreover, the agents can continually refine skills and knowledge through a comprehensive learning process.

EPISODIC MEMORY IN ARTIFICIAL INTELLIGENCE

The term "episodic memory" was first introduced by Endel Tulving (1972) to distinguish between remembering events from the past (episodic memory) and knowing factual information (known semantic memory). Semantic memory focuses on general knowledge about the world and includes facts, concepts and ideas. By contrast, episodic memories are long-term memories of specific experiences or events, such as what you did yesterday or during your high school graduation.

Allen and Fortin (2013) use an operational definition of memory, which is "events in context." Their definition emphasizes that, in the episodic memory system, information about specific events is tied to the spatial, temporal and other situational contexts in which they occurred. Specifically, the system comprises content, structure and flexibility. Content is the information that the individual remembers about the event, including the event's context (time and location) of occurrence. Structure is the integration of the information about the event and its context into a single representation. Flexibility enables expressing the memory to support adaptive behavior in novel situations.

Since Tulving proposed in 1972 that episodic recall involves the ability to "mentally time travel" to reexperience specific events, the phenomenology has been related to only humans and animals. More recently, episodic memory research has been linked to case-based reasoning (CBR) in AI research, as proposed by Kolodner in 1993. In CBR systems, a case represents the solution to a previously encountered problem that can be retrieved and adapted for new problems. The structure of cases, including specific fields within each case, is typically designed by humans for specific tasks or sets of tasks. However, this approach can limit the cases' generality, or applicability across different scenarios. To overcome those limitations, continuous CBR (Ram and Santamaría, 1997) relies upon cases that consist of the agent's sensory experiences and none of its internally generated abstractions.

Nuxoll and Laird (2007) highlighted that episodic memory shares similarities with CBR, viewing it as a versatile architectural method applicable to continuous CBR. They consider episodic memory as what you *remember*, and it includes contextualized information about specific events. In contrast, they describe semantic memory as what you *know*, consisting of isolated, decontextualized facts useful in reasoning about general properties of the world.

In their work, they developed task-independent mechanisms for encoding, storing and retrieving episodes without making assumptions about the structure or contents of these episodes. They outlined the design space for an episodic memory module (EMM), specifying the necessary criteria for their effective integration into cognitive architectures. Depicted in Figure 2 is the blueprint they provide for designing an episodic memory subsystem in a cognitive architecture with a series of design decisions to be made by the engineer.

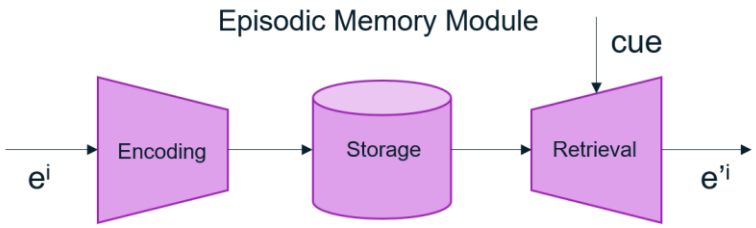

Figure 2. Episodic Memory. *The EMM is a component of the Soar cognitive architecture (Nuxoll and Laird, 2007).*

In Nuxoll and Laird's architecture, engineers must make decisions about encoding, storage and retrieval. Encoding requires deciding on when to initiate episode encoding, what to encode, and how to index it. For storage, the choices concern the structure of the episodes for storage efficiency and speed (e.g., how the episodes change over time) to enforce forgetting and generalization dynamics. Retrieval specifies when to initiate; which cue (query format) to use to match the stored episode; how the cue is matched with the most suitable episode; how the episode is actually loaded from the physical storage; how the episode is decoded (into its original form or not); and whether to use additional metadata about the episode in the matching cue.

The main challenge of modeling episodic memory is to build efficient (fast encoding and retrieval) and scalable storage for contextual (spatial and temporal) relationships (Chang and Tan, 2017). Episodic memory should allow for generalization when required and remain adaptable to new experiences. However, the memory model should support recalling stored events in real time in response to partial or noisy retrieval cues. Although there has been significant interest in the study of episodic memory in AI, most initial models of episodic memory rely heavily on symbolic representation. This limits their ability to encode spatial and temporal relationships, generalize, and recall using vague or incomplete cues (Nuxoll and Laird, 2007). Furthermore, the challenge of scaling memory storage in real-time environments arises, exacerbated when there is a lack of forgetting or a generalization dynamic.

Experience replay, also known as the replay buffer or replay memory, has been a staple of DRL algorithms since it was first introduced by Deep Q-Network (DQN) (Mnih et al., 2015). In brief, a replay memory is a data structure that temporarily stores the agent's observations, allowing the learning procedure to update them multiple times. The replay buffer accumulates and stores a vast dataset of experiences from various episodes. During subsequent offline training phases, miniature batches of experiences are randomly sampled from the replay buffer. This random sampling breaks the correlation between consecutive experiences, stabilizing and improving the learning process. By sampling from a wide range of past experiences, the agent can learn more effectively and generalize better to new situations. Over the last decade, the replay buffer has been a significant object of study in the context of DRL. Different designs and methods for utilizing it have been proposed.

More recently, there has been growing interest in introducing episodic memory into key components of DRL algorithms, such as exploration, experience replay and loss function. (Yang et al., 2020, provides a more extensive account of previous works.) These methods usually incorporate episodic memory in just one component of DRL. Moreover, they are complex, time consuming and storage intensive. Yang et al. (2020) proposed the Highly Efficient Episodic Memory DQN (HE-EMDQN). It is the first sample-efficient RL architecture with a new, highly efficient EMM, beyond the original, simplistic use of a KD tree. It also uses key components of DRL simultaneously: exploration, experience replay and loss function. However, one drawback remains: HE-EMDQN encodes the episodes utilizing a discrete (nonparametric) memory. This hampers its effectiveness in continuous control tasks and has limited its generalization ability to aggregate the experience across trajectories. More recently, Chen et al. (2022) proposed the Parametric Episodic Memory, which leverages the discrete memory by neural networks, enhancing both sample efficiency and generalization ability.

All these concepts build toward the ability of RL agents to use past experiences for self-improvement. A crucial missing component is the efficient and near-real-time representation and retrieval of high-fidelity (in the case of visual observations, photorealistic) episodic memory. NFs, particularly Neural Radiance Fields (NeRF) (Mildenhall et al., 2020), have gained attention in AI due to their significant representational advantages beyond computer vision and graphics. These advantages include a simplified mathematical model in which an artificial neural network is trained to map sensor coordinates to sensor readings. NFs enable a precise, spatially aware representation of observational data (sensor domain), allowing for efficient encoding and retrieval with a minimal memory footprint and low memory latency. The precision and efficiency are essential for implementing episodic memory in RL agents, facilitating better use of past experiences to refine and improve behavior. This is particularly important in EMMs that are to be integrated into the control systems onboard uncrewed aerial vehicles, where real-time processing and efficient memory usage are critical for optimal performance and safety.

NEURAL FIELD HIERARCHICAL EPISODIC MEMORY

NFs, or INRs, are a class of deep neural networks (DNN) designed to map spatial-temporal coordinates to specific measures, such as scalars or vectors, making them coordinate-based neural networks (Tancik et al., 2020; Sitzmann et al., 2020). These methods are extensively used in computer vision, graphics and robotics to parameterize the physical properties of scenes or objects across space and time, enabling 3D or 4D scene reconstruction (Xie et al., 2022). Applications include 3D shape and image synthesis, human body animation, 3D reconstruction, and pose estimation. Typically parameterized as a multilayer perceptron (MLP) with low-dimensional inputs, NFs can learn high-frequency functions by using positional encoding or periodic activation functions like sinusoidal or wavelets. These techniques improve the MLP's ability to handle high-frequency, low-dimensional regression tasks. This was demonstrated by the enhanced image clarity achieved through Fourier feature mapping, which preprocesses input coordinates to capture finer details.

It is important to consider two unique properties of NFs. One is that NFs leverage neural networks to depict continuous signals by encapsulating them as acquired functions capable of interpolating and extrapolating based on the training data coordinates. This methodology fundamentally alters how data is traditionally represented, moving away from discrete samples to a continuous function that can be queried at any point in its domain. This is particularly advantageous when dealing with high-dimensional data like images or 3D shapes, for which the continuity and smooth transitions captured by NFs are vital.

The second unique property pertains to radiance fields, which produce a compact representation for efficiently and quickly storing the data of the original signal, a form of neural compression (Dupont et al., 2021-2022, Rivas-Manzaneque et al., 2023a). In other words, storing the weights of the radiance field model, which acts as a compressed form for a given signal (e.g., images, audio, medical and climate data), can be seen as a form of data compression. Moreover, they can be interpreted as generative models, learning distributions of radiance fields either implicitly or explicitly. In turn, they facilitate advanced algorithms for tasks like real-time 3D shape inference (Rivas-Manzaneque et al., 2023b) and the generative modeling of signals in general.

The most popular NF is NeRF, which parameterizes a 3D scene as a 3D NF mapping of 3D spatial coordinates plus the camera orientation (two extra parameters) to the magnitudes of radiance (RGB color of the point in the tridimensional space) and density. NeRF is a novel view-synthesis technique whose goal is to be able to render photorealistic 2D images of a tridimensional scene by training a radiance field (color plus density) of points in the 3D space. Figure 3 shows an overview of a NeRF scene representation and differentiable rendering procedure. Images are synthesized by sampling 5D coordinates (location and viewing direction) along camera rays (a), feeding those locations into an MLP to produce a color and volume density (b), and using volume rendering techniques to composite these values into an image (c). This rendering function is differentiable, so we can optimize our scene representation by minimizing the residual between synthesized and ground-truth-observed images.

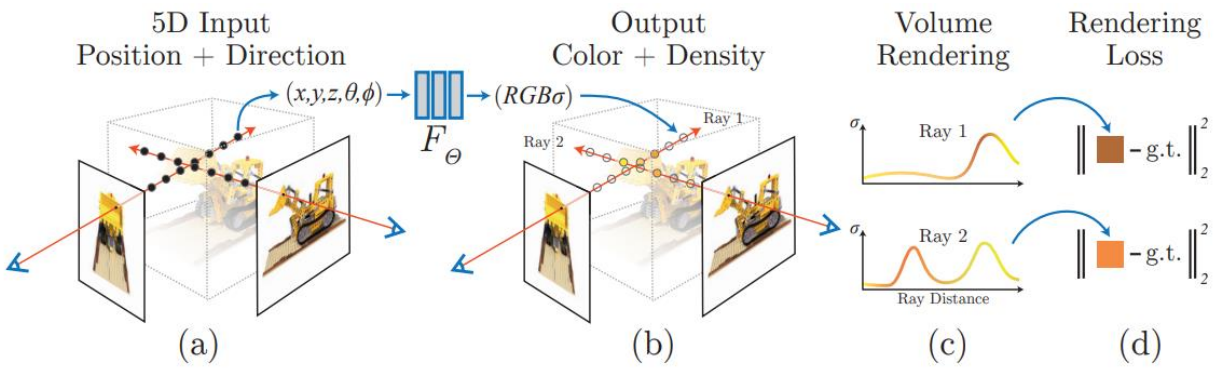

Figure 3. Neural Radiance Fields. *Introduced by Mildenhall et al. (2020), NeRF uses a radiance field to map radiance (RGB) and occupancy values to points in 3D space. It then employs a differentiable rendering technique to infer the RGB values of the pixels in the projected 2D view.*

More recently, a semantically aware NeRF was developed for visual scene understanding (Nguyen et al., 2024). These new methods explore the capability of NeRF representation not only to generate high-quality, new viewpoints, but also to complete missing scene details (inpainting), conduct comprehensive scene segmentation (panoptic segmentation), predict 3D bounding boxes, edit 3D scenes, and extract object-centric 3D models. A significant aspect of these methods is the application of semantic labels as viewpoint-invariant functions. These effectively map spatial coordinates to a spectrum of semantic labels, thus facilitating the recognition of distinct objects within the scene.

Recently, multiple works showed the tremendous potential of NeRF in RL. Driess et al. (2022) address the challenge of finding effective state representations for training RL agents. They introduce a novel approach called NeRF-RL that leverage NeRFs to supervise the learning of state representations. Their experiments show that NeRF-RL improves the performance of robotic object manipulation, compared to other methods. Wang et al. (2024) use NeRF in vision-based RL tasks to enhance an agent's ability to generalize about new images. Shim et al. (2023) combine semantic-aware NeRFs with a convolutional encoder to learn 3D representations from multi-view images to enable

model-free and model-based RL. There have been significant advancements in the use of NeRF in RL. However, none envisions this as a technology enabler for the generalized use of episodic memory in RL to support self-improving AI capabilities beyond the conventional RL loop.

NEURAL FIELDS-BASED HIERARCHICAL EPISODIC MEMORY

In our NF-HEM blueprint for self-improving agents we propose a two-layered architecture for a building block. The higher-level episodic memory (L2) embeds (partially or totally) the observation domain of each episode and uses the latent vector to index the storage of the episode data. The lower-level episodic memory (L1) encodes the observation data in a spatially aware manner. This allows random access to a specific observation to be retrieved by providing the coordinates (spatial and/or temporal) of that specific experience within the episode. Figure 4 shows the HEM architecture. In it, the L2 episodic memory allows for the utilization of any observation or sequence of observations as a cue to access the episode that closely aligns with the observation domain. Upon choosing the episode, its data is retrieved from storage. The episode data includes the NFM (L1 episodic memory), which effectively encodes all observations within the episode in a continuous and spatially aware compact representation.

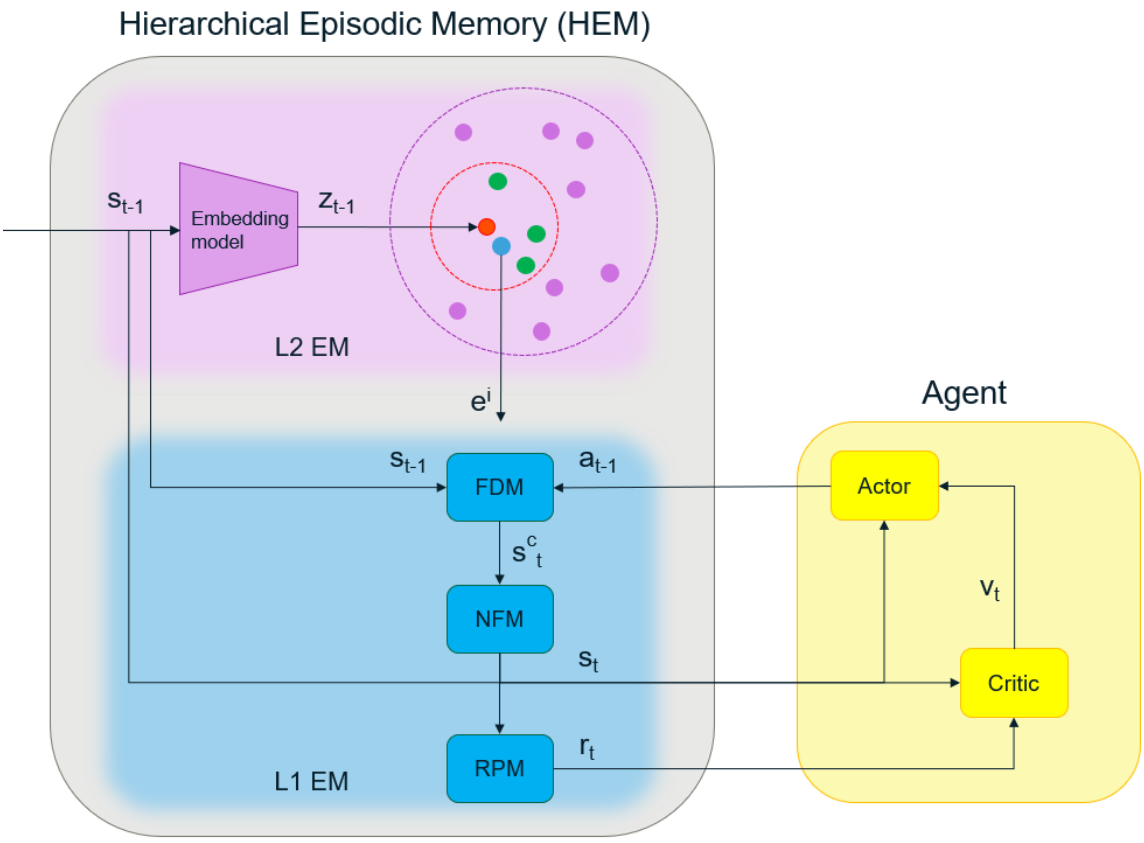

Figure 4. NF-HEM Architecture. *HEM consists of a two-layered hierarchical architecture with a higher-level episodic memory (L2) that maps observations to the closest episode (e^i), and a lower-level episodic memory (L1) taking the form of an NFM that reconstructs the detailed observation (S_t) space of the episode.*

L2 is a higher-level episodic memory. It is responsible for mapping observations, such as sensory inputs (e.g., a single image, a sequence of frames, audio signals or other sensory data), into an embedding (latent) space. These observations can be multimodal, encompassing various types of sensory input.

L2 comprises three components: an embedding model, observation encoding and a vector database. The first of these is a DNN model used to map a high-dimensional observation space into a lower-dimensional latent space. This allows for the efficient indexing of episode data. The second component refers to episode encoding, during which some or

all observations are used to create an observation embedding that captures the essential features of the data. For example, if a camera is the only sensor, the visual observation data used to generate the embedding to search for the closest episode can go from single images to a sequence of video frames. This can include a sliding window of observations to create multiple embeddings of the same episode. In the latter case, more than one vector is generated for the same episode. This results in a finer-grained representation of the observation domain of the episode in the L2 memory.

The resultant observation embeddings are stored in L2's third component, a vector database that indexes the episodes. It is important to note that the vector or episode embedding serves as an index to retrieve the episode later. The actual episode data retrieved are the weights of the NFM, which will then be used at L1 for finer-grained access to the observation space of the episode. The vector database encodes different episodes in memory to use to retrieve previous instance data for an agent.

A DNN model is used to generate vector embeddings from multimodal data. Various types of information (e.g., text, images, audio, tabular data) can be embedded using the DNN model. The data is run through the embedding model to get vector representations, and additional metadata can be stored with the vector embeddings for prefiltering or postfiltering search results for "approximate nearest neighbor" (ANN). To store the data, the vector embeddings and metadata are indexed separately using such methods as random projection, product quantization or locality-sensitive hashing.

To read episode data, the architecture executes a query against the vector database. This includes data for an ANN search (e.g., an image or audio clip to find similar ones) and a metadata query to exclude specific vectors based on known qualities. The metadata query is executed against the metadata index before or after the ANN search. The query data is embedded into the latent space using the same model used for writing data. The ANN search procedure retrieves a set of vector embeddings using similarity measures like cosine similarity, Euclidean distance or dot product.

L1 represents a lower level in the hierarchical DEM system and is implemented as an NFM (also known as a coordinate-based MLP). This model encodes all the observations of an episode in a spatially aware manner, enabling random access to specific episode points in space and time. This allows for recovering the sensory input at precise moments. NFM reconstructs the detailed observation space of a single episode, encoding the episode's observations in a continuous INR. Spatially aware encoding is a parametrized model with a coordinate-based DNN that maps the spatiotemporal coordinates of an original observation (sensory input) taken in that point of the episode. Coordinate-based random access allows for querying specific points in space and time, enabling the retrieval of observations exactly where and when they were taken.

Figure 4 includes two additional building blocks to allow NF-HEM to become an interactive world model (Ha and Schmidhuber, 2018; Bond-Taylor, 2021). A flight dynamics model serves as a lightweight representation of the agent's physical location, predicting the next coordinate the agent will move to, based on the current coordinate and action. In the second block, a reward predictive model predicts the reward, given the predicted next state. The hypothesis that NF-HEM can serve as a world model will be formulated in depth in a separate paper and thoroughly investigated in the future; here, it is depicted merely as a thought-provoking idea.

The hierarchical organization of the DEM can be configured depending on the available storage and computational resources, as well as the time requirements. This flexibility allows the system to balance between computational efficiency and the level of detail in episodic memory representation.

In summary, L2 handles the encoding and storage of raw observations into an efficient latent space using a vector database enabled for accommodating multimodal data. For visual observations, this can include single images, sequences of video frames or sets of keyframes. The embeddings in L2 act as indices to retrieve the actual episode data in a vector database, which are the weights of the NFM used in L1. This lower-level memory leverages a coordinate-based DNN to provide a spatially and temporally aware encoding of an episode, supporting both efficient storage and the precise retrieval of episodic memories.

NF-HEM APPLIED TO THE ALPHADOGFIGHT TRIALS SCENARIO

The ADT, a precursor of ACE, were run in 2020 to test whether AI agents could effectively learn basic fighter maneuvers through simulated dogfights (DeMay et al., 2022). Figure 5 presents some examples from these tests. Eight diverse teams, ranging from small companies to large defense contractors, participated in a series of 1-v-1 combat simulations using a high-fidelity F-16 flight dynamics model. These competitions culminated in a three-day virtual event where AI agents faced off against each other and a human pilot.

Figure 5. AlphaDogfight Trials. *These screenshots show the typical behavior of the low-level policies driven by reward shaping: control zone (left); aggressive shooter, resulting in a head-to-head situation (middle); and conservative shooter (right). (Extracted from Pope et al., 2023.)*

Throughout ADT, teams found that the experience replay buffer was a crucial component to store an agent's experiences over time. These experiences are typically recorded as tuples <S, A, R, S'> consisting of the current state, the action taken, the reward received, the next state, and whether the episode has ended. Several benefits related to the use of past experiences during training were discovered. They provided stable and efficient training by breaking correlation between consecutive state, a situation that can lead to inefficient learning and poor generalization. The replay buffer breaks these correlations by randomly sampling experiences during training. Also, the use of past experience improved sample efficiency. This was particularly important in environments where obtaining new experiences is expensive or time-consuming – like dogfighting. Also, replay buffers facilitated an agent's ability to prioritize certain experiences based on significance or relevance. This enabled more focused learning and faster convergence toward optimal policies. Experiences were categorized as giving advantage or disadvantage to each one of the low-level policies, and they were categorized in terms of the behavior of adversarial pilots (Pope et al., 2023).

Most ADT teams overcame reward-shaping challenges by leveraging current observation space and expert pilot knowledge to offer more frequent reward signals to the agent. The reward function of each low-level policy became the sum of several independent rewarding or punishing components, each designed to encourage a specific behavior. They were carefully selected in an iterative and interactive time-consuming process in which the expert (former pilot, Air Force officer) watched playbacks to qualitatively assess the behavior. Thus, expert involvement is critical to designing and adjusting these reward components to ensure that the agent learns the desired behaviors. This process is labor intensive and can lead to brittle policies that can perform well in training environments but fail to generalize to slightly different or unforeseen scenarios. The reliance on expert knowledge also means that any biases or incorrect assumptions made during the reward design can significantly impact the agent's learning and performance (Pope et al., 2023).

Human engagement became critical in selecting the experience for each low-level policy to achieve sample efficiency. During ADT, teams observed that the performance of low-level policies was highly sensitive to the initial conditions of training episodes. Initial conditions, such as relative position, altitude and velocity, defined the positional advantage or disadvantage at the beginning of each episode. These conditions were classified into categories ranging from highly defensive to highly offensive positions, and were selected according to the policy's specific goals. The rationale was to optimize training by narrowing down the possible state-space, selecting the subset of episodes with initial conditions favorable to low-level policies. This process became an iterative, interactive and time-consuming task (Pope et al., 2023).

The goal of self-improving agents is to remove the involvement of humans in autonomous agent engineering, to have the agent self-optimize to overcome challenges, such as sparse rewards, sample efficiency or the exploration-exploitation dilemma. By developing agents that can autonomously adjust and optimize their RL program, we can enhance their adaptability and performance in dynamic scenarios without the extensive need for human intervention. Spatially aware episodic memories implemented as NF-HEM serve as a blueprint for self-improving agents.

In the ADT scenario, the implementation of self-improving AI capabilities, mediated by the NF-HEM module, can significantly enhance reward shaping and experience replay. NF-HEM proposes an autonomous, two-way continuous self-improvement process, leveraging the flows of experiences between simulation and real-world environments. NF-HEM ensures that AI agents can learn and adapt in both contexts, improving their performance over time in a self-optimizing schema that transcends the conventional RL loop. The architecture plays a fundamental role here, allowing for the high-fidelity encoding, efficient storage and fast, spatially aware retrieval of episodic experience (sensor data). It is important to notice that NF-HEM not only enables AI capabilities to self-optimize the learning program, but also enables critical AI safety capabilities (e.g., novelty detection, safe exploration, explicability and self-evaluation). As a result, these are built into the same self-improving system by design.

The sim-to-real experience flow with NF-HEM enables several key capabilities for enhancing RL agents: novelty detection, identifying repetition, virtual sensing, safe exploration and managing goals. The first of these pertains to detecting unfamiliar scenarios, which enhances safety by identifying when the agent is out of its domain of competence. For example, if the agent encounters an unexpected maneuver from an opponent, it can flag this scenario for further analysis and cautious exploration in the simulated environment. The second capability, identifying repetition, consists of recognizing familiar scenarios to optimize exploration by focusing on novel experiences. For instance, if the agent has repeatedly encountered a particular adversarial strategy, it can prioritize exploring new strategies to counter.

Virtual sensing accesses past sensory details not currently in view to enable the agent to recall specific environmental conditions or opponent behaviors encountered previously. This aids in anticipating opponent moves based on historical data, or imagining the advantage given by an arbitrary coordinate. Safe exploration utilizes episodic memory to predict success or failure outcomes, which ensures that the agent explores strategies likely to be safe. For example, if the agent recalls that a particular evasive maneuver often fails, it can avoid it altogether. The last capability, managing goals, refers to keeping the long-term goal of the current episode in focus without overburdening working memory. This helps the agent to maintain strategic objectives, such as achieving a positional advantage over the opponent throughout the engagement.

On the other side, the real-to-sim experience flow with NF-HEM offers additional critical capabilities for refining and advancing RL agents: action modeling, environment modeling, self-evaluation, explicability and experience replay. Through action modeling, by recalling and analyzing specific episodes, the agent can better understand the effects of its actions. For instance, replaying a successful dogfight can help the agent to learn which maneuvers were most effective. This would turn the NF-HEM into a world model, as was suggested earlier in this paper.

Environment modeling uses sequences of episodes to predict environmental changes and simulate them during training in order to help the agent adapt to dynamic combat environments. For example, simulating how weather conditions affect visibility and maneuverability can prepare the agent for real-world variations. Self-evaluation entails analyzing live experiences later to enhance understanding and retroactive learning. After a combat exercise, the agent can review its performance to identify strengths and weaknesses, leading to continuous improvement. Explicability involves replaying past behavior for explanation with humans and knowledge sharing, which ensures that the agent's decisions are transparent and understandable. This is crucial for building trust with human operators and refining tactics based on human feedback. The last capability on this side of NF-HEM, experience replay, enhances the replay buffer learning mechanism through the optimal self-management of the most valuable episodes to ensure efficient learning. By prioritizing episodes that provide the most learning value, the agent can accelerate its training and achieve better performance quickly.

Through this comprehensive approach, we aim to develop agents that can autonomously adjust and optimize their RL process, enhancing their adaptability and performance in complex, dynamic scenarios without extensive human

intervention. We also propose that NF-HEM facilitates shifting the role of humans to supervisors in the self-improving AI process, a new relationship between the human and the agent that we call "human on the improving loop."

CONCLUSIONS

The ACE program's achievements in dogfighting simulations, culminating in the ADT, illustrate the impressive progress of RL agents. However, the immense human effort required to meticulously design and oversee these training environments underscores the need for more autonomous self-improvement mechanisms. This paper proposes the adoption of an NF-HEM as a pivotal component in advancing RL agent self-improvement, reducing reliance on human intervention. This is critical for rapid and efficient adaptation to dynamic environments. Also, we argue that this same building block can be used to implement and embed AI safety capabilities (e.g., novelty detection, safe exploration, explicability and self-evaluation) into the self-improving AI architecture. In this respect, we propose the concept of human on the improving loop to describe the new human-agent relationship. In that, AI agents can self-optimize their learning program while remaining under the control of the human supervisor.

We propose a blueprint for self-improving AI capabilities: a two-way self-improvement process and NF-HEM. This continuous process, integrating the flows of experience between simulation and real-world environments, ensures that AI agents can effectively learn and adapt in both contexts, improving their performance over time. NF-HEM plays a fundamental role, allowing for the high-fidelity encoding, efficient storage and fast, situationally aware retrieval of episodic experience data. Specifically, we hypothesize a bidirectional flow of episodic experiences, allowing agents to refine their policies and adapt to new challenges. This ensures a faster and more efficient convergence into safe and effective behaviors across complex, dynamic and uncertain environments.

In the sim-to-real episodic experience flow, we explore capabilities, such as detecting unfamiliar scenarios to enhance safety and recognizing familiar scenarios to optimize exploration. The crucial aspects of this process are the use of past sensory details not currently experienced (virtual sensing), the utilization of episodic memory to predict success or failure outcomes for safe exploration, and the keeping of long-term goals in focus without overburdening working memory. In the real-to-sim episodic experience flow, we focus on recalling and analyzing specific episodes to understand action effects. We use sequences of episodes to predict environmental changes and simulate them during training, analyzing live experiences later to enhance understanding and self-evaluation through retroactive learning. Another key component is the replaying of past behavior for human review. This is done while enhancing the replay buffer learning mechanism through the optimal self-management of the most valuable episodes for efficient learning.

The proposed NF-HEM is a blueprint, and many questions regarding its implementation in software remain. For instance, we have proposed that L2 handles the encoding and storage of raw observations into an efficient latent space using a vector database capable of accommodating NFMs. We have also proposed that the NFM is a coordinate-based MLP. Recent studies suggest the potential use of different DNN architectures for the NF, such as transformers. Additionally, generative models could be used (as hypernetworks) to generate the NFM from the latent representation, rather than retrieving it from a vector database. These implementation details require thorough investigation to determine the most effective approach.

We hypothesize that NF-HEM can serve as a robust world model for training RL agents by leveraging its high-fidelity encoding, efficient storage and situationally aware retrieval of episodic experiences. This approach allows agents to train within a simulated hallucination/dream-like environment generated by NF-HEM, enabling faster and more resource-efficient learning. By training the agent's controller inside this internally generated environment and subsequently transferring the learned policy to the real world, NF-HEM could facilitate more effective and autonomous self-improvement. This, in turn, could reduce the reliance on actual environmental interaction during the training phase. This hypothesis needs to be further investigated in the future.

REFERENCES

AFRL (2023). DOD artificial intelligence agents successfully pilot fighter jet. *Air Force Research Laboratory*, published Feb. 13, 2023 (last visited, Jun. 2024). https://www.afrl.af.mil/News/Article-Display/Article/3297364/dod-artificial-intelligence-agents-successfully-pilot-fighter-jet/

Allen, T. A., & Fortin, N. J. (2013). The evolution of episodic memory. *Proceedings of the National Academy of Sciences of the United States of America*, 110 Suppl 2(Suppl 2), 10379–10386. https://doi.org/10.1073/pnas.1301199110

Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C.G. (2021). Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*, 7327-7347.

Chang, P., & Tan, A. (2017). Encoding and Recall of Spatio-Temporal Episodic Memory in Real Time. *International Joint Conference on Artificial Intelligence* (IJCAI-17), 2017, 1490–1496.

Chen, K., Gan, Z., Leng, S., Guan, C. (2022). Deep Reinforcement Learning with Parametric Episodic Memory. *2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022*, pp. 1-7, doi: 10.1109/IJCNN55064.2022.9891902.

Cotting, M.C., Stephens, S., Cole, J., Barricklow, J., Gray, W. (2023). X-62 VISTA Capabilities and Architecture. *AIAA SciTech Forum*, January 23-27, National Harbor, MD. https://arc.aiaa.org/doi/pdf/10.2514/6.2023-1744

DeMay, C.R., White, E.L., Dunham, W.D., Pino, J.A. (2022). AlphaDogfight Trials: Bringing Autonomy to Air Combat. *Johns Hopkins APL Technical Digest, Vol. 36, No 2.* https://secwww.jhuapl.edu/techdigest/content/techdigest/pdf/V36-N02/36-02-DeMay.pdf

DARPA (2024). ACE Program Achieves World First for AI in Aerospace. Published online (darpa.mil), April 17, 2024 (acceded May 13, 2024). https://www.darpa.mil/news-events/2024-04-17

Driess, D., Schubert, I., Florence, P.R., Li, Y., & Toussaint, M. (2022). Reinforcement Learning with Neural Radiance Fields. Advances in Neural Information Processing Systems 35 (NeurIPS), 2022.

Dupont, E., Goliński, A., Alizadeh, M., Teh, Y. W., & Doucet, A. (2021). COIN: COmpression with Implicit Neural representations. *Neural Compression Workshop, The 9th International Conference on Learning Representations* (ICLR), 2021.

Dupont, E., Loya, H., Alizadeh, M., Golinski, A., Teh, Y. W., & Doucet, A. (2022). COIN++: neural compression across modalities. *Transactions on Machine Learning Research*, 2022(11).

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. Annual review of psychology, 68, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Ha, D.R., & Schmidhuber, J. (2018). Recurrent World Models Facilitate Policy Evolution. *Neural Information Processing Systems*, 31:2451-2463.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. *In Proceedings of the Europeean Conference on Computer Vision* (ECCV), 2020, pp 405–421.

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., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

Nguyen, T., Bourki, A., Macudzinski, M., Brunel, A., & Bennamoun, M. (2024). Semantically-aware Neural Radiance Fields for Visual Scene Understanding: A Comprehensive Review. ArXiv, abs/2402.11141.

Nuxoll, A.M, Laird, J.E. (2007). Extending cognitive architecture with episodic memory. *Proceedings of the 22nd National Conference on Artificial Intelligence* (AAAI'07), vol. 2. AAAI Press, pp. 1560–1565.

Pope, A.P., Ide, J.S., Mićović, D., Diaz, H., Twedt, J.C., Alcedo, K., Walker, T.T., Rosenbluth, D., Ritholtz, L., & Javorsek, D. (2023). Hierarchical Reinforcement Learning for Air Combat at DARPA's AlphaDogfight Trials. *IEEE Transactions on Artificial Intelligence*, 4, 1371–1385.

Rivas-Manzaneque, F., Ribeiro, A., Avila-García, O. (2023a). ICE: Implicit Coordinate Encoder for Multiple Image Neural Representation. *IEEE Transactions on Image Processing*, vol. 32, 5209–5219, 2023, doi: 10.1109/TIP.2023.3299501.

Rivas-Manzaneque, F., Sierra-Acosta, S., Penate-Sanchez, A., Moreno-Noguer, F., Ribeiro, A. (2023b). NeRFLight: Fast and Light Neural Radiance Fields using a Shared Feature Grid. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12417–12427, doi: 10.1109/CVPR52729.2023.01195.

Rothfuss, J., Ferreira, F., Aksoy, E.E., Zhou, Y., & Asfour, T. (2018). Deep Episodic Memory: Encoding, Recalling, and Predicting Episodic Experiences for Robot Action Execution. *IEEE Robotics and Automation Letters, vol. 3, no. 4*, pp. 4007-4014.

Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., & Wetzstein, G. (2020). Implicit Neural Representations with Periodic Activation Functions. *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NeurIPS), December 2020, pp 7462–7473.

Sharma, A., Ahmed, A., Ahmad, R., & Finn C. (2023). Self-Improving Robots: End-to-End Autonomous Visuomotor Reinforcement Learning. *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

Shim, D., Lee, S., & Kim, H.J. (2023). SNeRL: Semantic-aware Neural Radiance Fields for Reinforcement Learning. *Proceedings of the 40th International Conference on Machine Learning* (ICML), 2023, PMLR 202:31489–31503.

Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., & Ng, R. (2020). Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NeurIPS), December 2020, pp. 7537–7547

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson, Organization of memory. Academic Press.

Tulving, E. (2002). Episodic memory: from mind to brain. Annual review of psychology, 53, 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Yang, D., Qin, X, Xu, X, Li, C., Wei, G. (2020). Sample Efficient Reinforcement Learning Method via High Efficient Episodic Memory. *IEEE Access, vol. 8*, pp. 129274-129284, 2020, doi: 10.1109/ACCESS.2020.3009329.

Yang, Z., Moerland, T.M., Preuss, M., and Plaat, A. (2023) "Two-Memory Reinforcement Learning," *2023 IEEE Conference on Games (CoG)*, Boston, MA, USA, 2023, pp. 1-9, doi: 10.1109/CoG57401.2023.10333174.

Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., & Sridhar, S. (2022). Neural Fields in Visual Computing and Beyond. *Computer Graphics Forum*, 41(2), May 2022, pp 641-676. https://doi.org/10.1111/cgf.14505.

Zhu, H., Yu, J., Gupta, A., Shah, D., Hartikainen, K., Singh, A., Kumar, V., & Levine, S. (2020). The Ingredients of Real-World Robotic Reinforcement Learning. *The 8th International Conference on Learning Representations* (ICLR), 2020.

Wang, G., Pan, L., Peng, S., Liu, S., Xu, C., Miao, Y., Zhan, W., Tomizuka, M., Pollefeys, M., & Wang, H. (2024). NeRF in Robotics: A Survey. *ArXiv, abs/2405.01333.*