Human Factors in Robots, Drones and Unmanned Systems, Vol. 138, 2024, 1–10
https://doi.org/10.54941/ahfe1005003

Integrating Episodic and Semantic Memory in Machine Teammates to Enable Explainable After-Action Review and Intervention Planning in HAA Operations

Eric Davis[1] and Katrina Schleisman[2]

[1]Galois, Inc., Portland, OR 97205, USA
[2]Galois, Inc., Minneapolis, MN 55401, USA

ABSTRACT

A critical step to ensure that AI systems can function as effective teammates is to develop new modeling approaches for AI based on the full range of human memory processes and systems evidenced by cognitive sciences research. In this paper we introduce novel techniques that integrate episodic and semantic memory within Artificially Intelligent (AI) teammates. We draw inspiration from evidence that points to the key role of episodic memory in representing event-specific knowledge to enable simulation of future experiences, and evidence for a representational organization of conceptual semantic knowledge via self-organizing maps (SOMs). Together, we demonstrate that these two types of memory working in concert can improve machine capabilities in co-learning and co-training scenarios. We evaluate our system in the context of simulated helicopter air ambulance (HAA) trajectories and a formal model of performance and skill, with interventions to enable an AI teammate to improve its capabilities on joint HAA missions. Our modeling approach contrasts with traditional neural network training, in which specific training data is not preserved in the final trained model embedding. In contrast, the training data for our model consists of episodes containing spatial and temporal information that are preserved in the model's embedding. The trained model creates a structure of relationships among key parameters of these episodes, allowing us to understand the similarity and differences between performers (both human and machine) in outcomes, performance, and trajectory. We further extend these capabilities by enhancing our semantic memory model to encode not just a series of episodes, but labeled directed edges between regions of semantic memory representing meta-episodes. These directed edges represent interventions applied by the performer to improve future episodic outcomes in response to identified gaps in capability. These interventions represent the application of specific co-training strategies as a labeled transition system, linking episodes representing pre-intervention and post-intervention performance. This allows us to represent the expected impact of interventions, simulating improvements and skill decay, providing the machine with team-aligned goals for self-improvement between episodes to positively impact future teamwork.

Keywords: Human-machine teaming, Co-learning, Co-training, After-action review, Cognitive modeling, AI, Episodic memory, Semantic memory

INTRODUCTION

As the field of artificial intelligence (AI) has rapidly advanced in recent years, so has the need for machines to interact with human users in ways that are less like an appliance and more like a teammate. State-of-the-art techniques in AI such as neural networks and large language models (LLMs) allow modern systems to model complex statistical relationships between features and outcomes of artificial reasoning. While successful at producing high-accuracy predictions based on statistical learning, they fall short of true contextual reasoning displayed by human teammates and generally fail to capture the knowledge, skills, and strategies necessary for teams to perform effectively (Johnson et al., 2019). When human teammates face a decision-making challenge, they are able to leverage knowledge about similar prior experiences and apply that knowledge to novel experiences. The ability to apply knowledge from contextually similar, but not identical, experiences enables teams to evaluate the quality of potential future decisions in light of past outcomes (Deutsch et al., 2008) and is critical to co-learning and co-training.

In this paper we present a novel algorithm that demonstrates contextual reasoning abilities in an AI system through the implementation of 1. An *episodic memory*-like process that relies on the spatiotemporal context of training data, and 2. A topographic *semantic memory*-like process based on the concept of representational geometry (Kriegeskorte & Kievit, 2013). Events stored in the AI's episodic memory represent a model of the world experienced across the series of episodes. The collection of episodes are embedded in the machine's topographic semantic memory, grouped and clustered via relationships important to co-performance and co-training goals. The semantic memory structure is generated using self-organizing maps (Kohonen, 1990), which provides a simple inversion mechanism to simulate future episodes and the results of interventions suggested by after-action review. The overarching goal is to support new capabilities in AI teammates to learn alongside both human and other AI teammates, to participate in after-action review, and to set and achieve goals for their own skill development as a result of these new algorithms and an integrated memory system.

We explore these techniques in a case study of Helicopter Air Ambulance (HAA) operations from MIT's HAA encounter model (Weinert, 2020), validating our approach with a modified version of this data set including a full six degrees of freedom model assuming a standard platform used by the city of Boston, Massachusetts for HAA operations. We demonstrate the ability of our model to predict the results of interventions from a synthetic intervention model based on this data set, and show its ability to organize both individual episodes of HAA encounters, and meta-episodes employing interventions to shift capabilities to improve skill for future episodes.

SEMANTIC MEMORY

AI representational approaches, whether symbolic or statistical, most closely align to the type of human memory known as *semantic memory.* Semantic memory is stored long-term memory for conceptual knowledge

(Squire, 2004). Semantic memory representations code information in a way that is largely dissociated or abstracted from the contexts in which it was learned. This is similar to the way that neural networks do not store information about which training trials came first or last, but abstract all training data into a single representational embedding.

Evidence from cognitive neuroscience suggests that human semantic memory has a category-specific organizational structure, and that the emergence of this structure can be modeled using self-organizing maps (SOMs) (McClelland & Rogers, 2003). In the brain, concepts in semantic memory that are highly related to each other may be stored closer in topological spaces, while dissimilar concepts are further apart. Damage to regions of the brain that support semantic memory can result in selective impairments for specific categories of concepts such as living versus nonliving items (Warrington & Shallice, 1984). These results can also be temporarily induced in non-patient populations by disrupting activity in the same set of brain regions via transcranial magnetic stimulation (Pobrick, Jefferies & Ralph, 2010). Semantic deficits can present in patients in narrower sub-categories, such as "animals", "fruit/vegetables", and "artifacts" (Caramazza & Mahon, 2003). This evidence for the structure of semantic memory is consistent with evidence for other types of topological maps in the brain: retinotopic maps in primary visual cortex representing the spatial arrangement of the visual field (Wandell & Winawer, 2011), tonotopic maps in primary auditory cortex representing sound frequencies or tones (Formisano et al., 2003), the somatosensory and motor homunculi representing adjacent parts of the body in frontoparietal cortex (Nakamura et al., 1998; Schieber, 2020), and visual object maps representing categories of visual stimuli such as faces, places, bodies, and tools in occipitotemporal cortex (Grill-Spector, Kourtzi & Kanwisher, 2001; Downing et al., 2001; Haxby et al., 2001).

Our modeling approach implements a topographical structure of semantic memory, in which notions of distance and topology are used to represent conceptual similarity. We build this topology using **self-organizing maps (SOMs)** (Kohonen, 1990) as shown in **Figure 1.** Self-organizing maps use the computationally efficient process of competitive learning to perform unsupervised machine learning, producing a low dimensional representation of higher dimensional data while preserving the topological structure of the data. Input vectors of n-dimensional observations of the form $\{x_0, x_1, \ldots, x_n\}$ are fed into the map, and compared to the weighting functions associated with each codebook vector in the map. The vector is assigned to the neuron in the map associated with a codebook vector whose weights are initially most similar to the new input vector, then the weights of this neuron and those closest to it in the map are updated to become more similar to this input vector.

The result of the process of training a self-organizing map is a set of codebook vectors in the map, an embedding of each data point to the most similar codebook vector, and a **unified distance matrix** which represents the Euclidean distance between neighboring neurons, defining a topology as shown in the right side of **Figure 1.** This results in a topology of concepts

with good local Euclidean properties, providing a relationship between clusters of semantically similar data points, and whose codebook vectors can be interpreted as exemplars of neighboring semantic data.

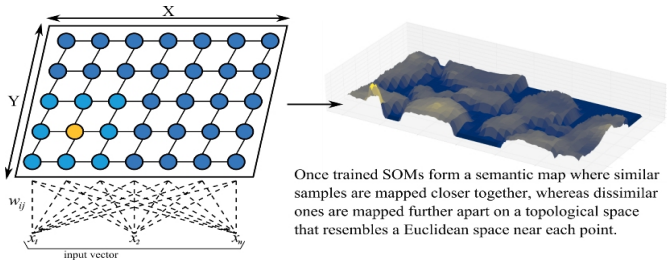

Figure 1: Self-organizing maps form the basis for our artificial semantic memory. Using competitive learning, high dimensional input vectors of features representing an episodic memory are embedded into a topographic map using competitive learning. This forms a manifold over the episodes which embeds them into a local Euclidean space similar to the cognitive neuroscience concept of representational geometry. Distances between episodes are represented in the topological space allowing concepts to be grouped by semantic similarity.

An AI system that can understand relative semantic distances and relational topologies between concepts can move closer to providing explainable representations of prior performance during after-action review and planned interventions to improve performance in future episodes. A topology of concepts, however, is by itself insufficient to support complex contextual reasoning. What is notably missing is a methodology for representing information for specific events in time and space. In cognition research this is known as *episodic memory*, which along with semantic memory comprises the larger declarative memory system in the human memory taxonomy (Squire, 2004).

EPISODIC MEMORY

The term *episodic memory* was coined by cognitive psychologist Endel Tulving and defined as "personal experience that is remembered in its temporal-spatial relation to other experiences" (Tulving, 1972). The concept was developed to directly contrast with the then-recently defined concept of semantic memory (Quillian, 1966). Tulving's model of the episodic memory system was developed based on insight into three key human functional abilities: a sense of *subjective time*, the ability to *mentally time travel* in the recollection of episodes, and a *self* (Tulving, 2002). In particular the work presented here focuses on our ability to mentally time travel, importantly not just to a recollected past, but also to an imagined future.

Episodic memories do not simply create a high-fidelity record of past experiences; in fact, the commonality of false memories suggests episodic memories are easily prone to distortion (Loftus & Pickrell, 1995). Instead, episodic memories are the basis for the ability of intelligent agents to perform *episodic future thinking* (Schacter, Benoit & Szpunar, 2017). Episodic future thinking is our ability to imagine, simulate, and plan for the future, based on what we have experienced in the past. Neuroimaging research shows

that common brain networks are active during both recollection of episodic memories and imagined simulations of the future (Schacter et al., 2012). Importantly, the active brain regions overlap with the *default mode network* of the brain (Raichle, 2015; Smallwood et al., 2021). The default mode network is an interconnected set of distributed brain regions whose activity is *lowest* when the mind is engaged in attention-demanding tasks, and *highest* when the mind is engaged in self-reflection and not constrained by sensory input. It is this latter "state of mind" that we are attempting to capture in the AI implementation described here, because it provides a model for a division of labor in intelligent systems between "mission-focused" activities in which specific goals must be met in real time, and "after-action review" activities in which reflection and integration of experiences can occur. It is notable that AI development has almost exclusively focused on systems that are designed to perform specific tasks, but that lay inert once that task is completed. This is very unlike the behavior of human teammates. We believe the next generation of AI systems will need an episodic memory-based, default-mode-network-like set of reflective capabilities in order to learn dynamically. By leveraging existing computing power during its "downtime" to recognize patterns in past episodes and simulate future episodes, an AI system can calibrate its future behavior to be a more successful teammate.

In our implementation individual episodes representing 120-second flight segments are stored as events, as shown below in **Figure 2.** Input vectors characterizing these episodes in a given domain are then generated to allow them to be embedded into a SOM representing their semantic distance from each other according to a high-dimensional vector, capturing metrics of interest generated for the episode. In our experiments we characterize these flights using a number of measures of human performance tied to individual skill including: velocity volatility, acceleration volatility, mean heading error, heading volatility, and roll volatility. These measures are generated both over the entire flight, and using sliding window estimates for 5-second and 10-second sub-segments of the entire trajectory. This allows us to embed episodic memories into our semantic memory structure so that nearest neighbors are those that are most alike in displayed performance.

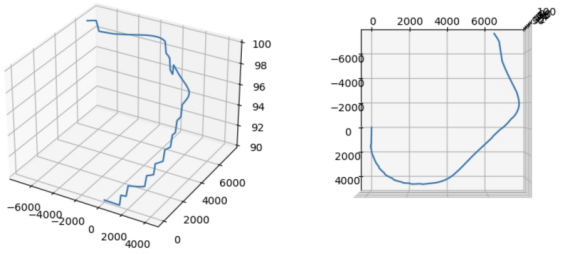

Figure 2: An example of an episodic memory from the HAA data set representing a single 120-second flight trajectory for a pilot. These episodes are embedded into a SOM on the basis of high dimensional features capturing human performance metrics derived from the flight, such as acceleration and velocity volatility, mean heading error, and roll volatility when correcting for heading error. These allow us to store not only the episodes themselves, but their relationships to each other in a given semantic space.

SEMANTIC CLUSTERING OF EPISODIC MEMORIES

An important feature of the artificial episodic memory is the grouping of those memories into neighborhoods of similar data, essentially creating a semantic memory-like topology representing categories of episodes. Using the unified distance matrix, the measured semantic distance between individual memories represents their similarity. Thus, given episodes can be said to be more like other episodes in their local neighborhood than episodes which embed into the map at neurons further away from this neighborhood. Furthermore, individual neighborhoods of similar memories can be compared as clusters, to understand the similarity or difference between different regions of episodes. Because each neuron in the SOM is also associated with a codebook vector trained on the embedded episodes, the SOM can be used to approximate *new* data that would be embedded near actual observations, and to understand the distribution of possible episodes that could embed in a given neuron of the SOM. This allows the memory to act as a vector database of episodes, and more importantly, a tool to simulate novel episodes, producing expectations of future memories that could be embedded into the SOM.

In our experiments with the HAA data set we used agglomerative clustering of episodes embedded into the SOM using the unified distance matrix for the SOM as a similarity measure between neighboring neurons. Similarity for non-adjacent neurons was calculated using Djikstra's algorithm (Djikstra, 1959) to find the shortest distance in the fully connected graph of all adjacent neurons. We then clustered individual neurons using agglomerative hierarchical clustering (Nielsen, 2016) estimating the appropriate number of clusters using a silhouette coefficient (Rousseeuw, 1987). An example of this process is shown below in **Figure 3.**

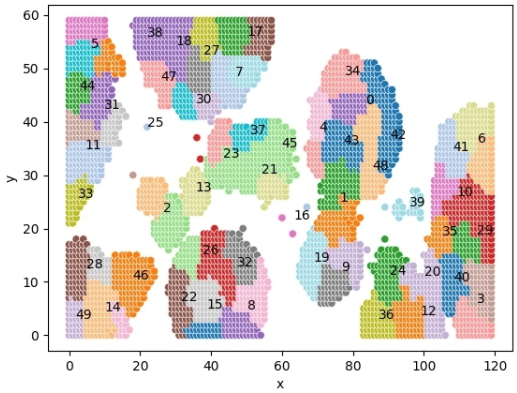

Figure 3: After individual episodes are embedded into a SOM, our method clusters the codebook vectors of neurons in the SOM on the basis of similarity using agglomerative clustering methods, determining the optimal number of clusters using the silhouette coefficient.

This clustering analysis allows us to understand the relationship between recorded episodes and to understand the features of these episodes which result in specific performance outcomes. For instance, when looking at pilot profiles which have low mean heading error, we find that episodes of higher mean heading error are associated with trajectories in which an obstacle

was present, forcing the pilot to change heading to maintain minimum safe distances required by FAA regulations. This analysis allows us to both understand the relationship between prior episodes and produce explanations for after-action review with human co-performers using contextual reasoning; and to understand and predict performance during future episodes by using the codebook vectors representing individual neurons in the map to create exemplars of future episodes.

REASONING ABOUT CO-TRAINING STRATEGIES

In addition to their use in after-action review to provide explanations, and to predict the characteristics and performance of future episodes that are similar to ones stored in our machine co-performer's memory, this fusion of semantic and episodic memory can also be used to plan interventions to improve performance by annotating our SOM with information about applied interventions forming meta-episodes. We define a meta-episode as a transition between performance states mediated by a co-learning or co-training intervention. An autonomous system might, for instance, display higher than desired acceleration volatility due to a lack of training data for contested air spaces (those complicated with other platforms, causing more frequent speed adjustments) resulting in a less reliable flight path and higher deviation from the ideal mission trajectory. As an intervention which we will label α, we might introduce new training data about similar missions resulting in a new controller that produces more stable paths in similar conditions. If a given machine co-performer was initially flying trajectories in a way that resulted in those trajectories being embedded in cluster 8 (shown in Figure 3), and after the application of intervention α resulted in trajectories that embedded in cluster 15, we would represent this meta-episode as the transition 8 $\overrightarrow{\alpha}$ 15, indicating that the application of intervention α when in a performance state that produces trajectories in cluster 8 will result in a new performance state associated with the production of traces in the cluster 15, as shown in **Figure 4.**

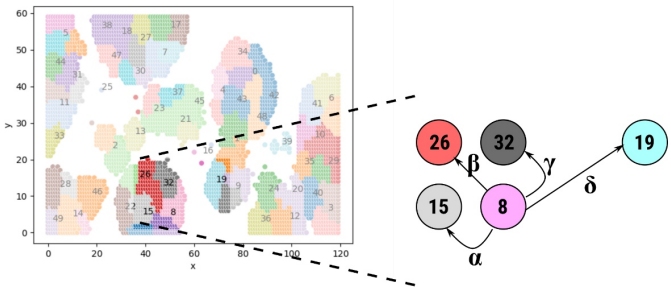

Figure 4: Once semantic clustering has been applied, we can analyze "meta-episodes" between episodic memories by building labeled transition systems from these clusters, as shown in the fragment above. If the agent experiences episodes that are embedded in the cluster labeled '8', we can examine the impact of possible interventions α, β, γ, and δ. For each of these interventions we look at the distribution of episodes post-intervention, and create a directed edge labeled with the intervention to the cluster where the post-intervention episode is embedded.

The fusion of episodic memory and semantic memory in this modeling approach allows an agent to build expectations of future performance states resulting not only from the interventions it has applied in the past, but also the interventions applied by other similar agents. Using the codebook vectors stored in semantic memory, it can even simulate likely future episodes after a proposed intervention, using the distribution of codebook vectors present at the expected future cluster to set expectations for co-performers, sharing possible intervention strategies in advance of their application for feedback from human co-performers.

RESULTS AND CONCLUSION

We tested our methods using data from MIT's HAA encounter model (Weinert, 2020) and profiled the demonstrated performance during a 120-second flight trajectory using measures of velocity volatility, acceleration volatility, mean heading error, heading volatility, and roll volatility. Synthetic interventions were introduced which provided mappings among witnessed performance states, trained over 100,000 randomly selected trajectories from the over 941,000 trajectories in the original data set. The interventions resulted in skill improvement in five performance measures as well as the potential for skill decay in other performance measures and were assigned randomly to a 50 performance state model over the trained data. This clustering was chosen based on the results of calculating the silhouette coefficient of various numbers of clusters, yielding the highest silhouette width for 40–60 clusters. We then simulated 1,000 instances of training over 50 new episodes, allowing our system to select interventions in order to improve performance in the next episode and measured the distance between the expectation on applying these interventions, and the resulting episode that occurred after the intervention was applied.

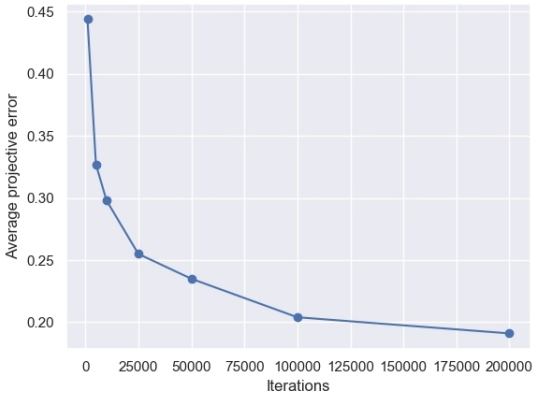

Figure 5: The average projective error of our methods in estimating trajectories resulting from an intervention is plotted against the number of iterations used to refine the SOM. This is the average distance from the cell predicted by our methods for a post-intervention trajectory to the cell in which the actual trajectory was embedded. Clusters representing similar trajectories had an average of 32.62 cells and standard deviation of 26.15 cells, meaning the difference was also, on average, within the same cluster of values.

Figure 5 shows the results of these experiments, as a factor of the number of training iterations used to form the SOM (and thus impacting its quality). We measured the average distance from the new expected episode and the episode which actually resulted, showing an average error of under one cell. Given an average of 32.62 cells per performance cluster, even in the case of error, the results typically yielded a cell within the same performance cluster, indicating similar but not exact performance. Furthermore, these methods are computationally efficient, and our results were run on a single core of an Apple M1 Pro processor with 32GB of RAM, with run times between 7.6 seconds for the 1,000 iteration instance, and 79.0 seconds for the 200,000 iteration instance. This is primarily due to the use of competitive learning for SOM estimation, which is a particularly computationally efficient algorithm which requires no hardware acceleration, making our method suitable even for embedded platforms and autonomous vehicles.

In conclusion, the modeling work we have described here demonstrates the benefit of integrating concepts from human cognition research into artificially intelligent systems. In grounding our modeling approach in evidence-based cognitive theories of human memory we produced a system with a novel representational structure and enhanced capabilities for future planning and explainability. We anticipate that human-machine teaming will continue to improve as principles of human cognition continue to be integrated into the design of machine teammates.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. Steven K. Rogers and the AFRL ACT3 QuEST group for providing ongoing inspiration for the future of AI systems.

REFERENCES

Caramazza, A., & Mahon, B. Z. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, *7*(8), 354–361.

Deutsch, T., Gruber, A., Lang, R., & Velik, R. (2008, May). Episodic memory for autonomous agents. In *2008 Conference on Human System Interactions* (pp. 621–626). IEEE.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numerische Mathematik, 1, pp. 269–271.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473.

Federal Aviation Administration. (2009). *Pilot's handbook of aeronautical knowledge*. Skyhorse Publishing Inc.

Formisano, E., Kim, D. S., Di Salle, F., Van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, *40*(4), 859–869.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10-11), 1409–1422.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Johnson, M., & Vera, A. (2019). No AI is an island: the case for teaming intelligence. *AI Magazine*, *40*(1), 16–28.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, *25*(12), 720–725.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322.

Nakamura, A., Yamada, T., Goto, A., Kato, T., Ito, K., Abe, Y.,... & Kakigi, R. (1998). Somatosensory homunculus as drawn by MEG. *NeuroImage*, *7*(4), 377–386.

Nielsen, F. (2016). Introduction to HPC with MPI for Data Science. Springer.

Pobric, G., Jefferies, E., & Ralph, M. A. L. (2010). Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology*, *20*(10), 964–968.

Quillian, M. R. (1966). Semantic Memory. Carnegie Institute of Technology. Unpublished Ph. D dissertation.

Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience*, *38*, 433–447.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron*, *76*(4), 677–694.

Schacter, D. L., Benoit, R. G., & Szpunar, K. K. (2017). Episodic future thinking: Mechanisms and functions. *Current Opinion in Behavioral Sciences*, *17*, 41–50.

Schieber, M. H. (2020). Modern coordinates for the motor homunculus. *The Journal of Physiology*, *598*(23), 5305.

Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., & Margulies, D. S. (2021). The default mode network in cognition: a topographical perspective. *Nature Reviews Neuroscience*, *22*(8), 503–513.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Wandell, B. A., & Winawer, J. (2011). Imaging retinotopic maps in the human brain. *Vision Research*, *51*(7), 718–737.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829–853.

Weinert, A., Campbell, S., Vela, A., Schuldt, D., & Kurucar, J. (2018). Well-clear recommendation for small unmanned aircraft systems based on unmitigated collision risk. *Journal of Air Transportation*, *26*(3), 113–122.