Modeling Autobiographical Memory for Believable Agents

Andrew Kope, Caroline Rose, Michael Katchabaw
Department of Computer Science, The University of Western Ontario
London, Ontario, Canada
{akope2, crose34, mkatchab}@uwo.ca

Abstract

We present a multi-layer hierarchical connectionist network model for simulating human autobiographical memory in believable agents. Grounded in psychological theory, this model improves on previous attempts to model agents' event knowledge by providing a more dynamic and non-deterministic representation of autobiographical memories. From this model, a Java-based proof-of-concept prototype system was created for use as an enabling technology in video games. This prototype was leveraged in the creation of a Minecraft modification (mod) implementation of the model that is able to demonstrate context-dependent recall and the effects of recency on memory recall. Wider implications of the model in agent and game design are discussed.

Motivation

In video games, agent event knowledge has traditionally been represented using one of a few arguably naive techniques. One classic technique for example, is to use a game-time related schedule, where the agents' behavior and dialogue advance along a pre-determined path, with progress along that path determined by checkpoints in game-clock time. Another technique is stimulus driven, where in-game events activate event knowledge objects in a predetermined master event list for a given agent (King 2002).

These techniques, although easy to implement and conceptually straightforward, each provide a deterministic and oversimplified representation of event knowledge as compared to the human capability to represent event knowledge with autobiographical memories. The ability of an agent to represent knowledge of events it experiences is a vital prerequisite of other characteristics important to agent believability such as social relationships, consistent identity, and the illusion of an independent life (see Loyall 1997). As such, there is a need for a new and less deterministic model of event knowledge, analogous to human autobiographical memory.

The ideal model of autobiographical memory would provide agents with a dynamic and extensible personal history of witnessed events, and could serve as the foundation for the more complicated systems necessary to build believable agents with simulated lives, social relationships, and consistent identities. We approached this problem from a psychological perspective by creating a conceptual model of autobiographical memory grounded in psychological research, then applying that model to the creation of complex, believable agents.

Overview of Long Term Memory

Declarative Memory

Declarative memory consists of the verbally communicable aspects of memory. Tulving (1972, 1987) distinguished between two types of declarative memory: semantic and episodic. Semantic memory includes knowledge of facts about the world, for example that pennies are round. We are interested in episodic memory, which pertains to knowledge of one's experienced life events and is used for the formation of autobiographical memories.

Autobiographical Memory Theory

McClelland, McNaughton and O'Reilley (1995) suggested that memories are first stored as changes in the hippocampal system. They argued that these changes in the hippocampus support reinstatement of recent memories in the neocortex, that these neocortical reinstatements elicit changes in the neocortex, and that the accumulation of these changes is what stores a memory. The authors concluded with the suggestion that the hippocampal system supports rapid learning of new items (because it does not require changes to neocortical structure), while the neocortex learns slowly, thereby encoding the structure of learned experiences. Meeter and Murre (2004) later

corroborated this conclusion by providing a neuropsychologically grounded account for the neocortical consolidation of hippocampally activated memories as described by McClelland, McNaughton, and O'Reilley (1995).

Burt et al. (1995) argued that there is evidence to support the hierarchical organization of autobiographical memory. In their view, at the top level of the hierarchy are representations of lifetime periods (e.g. time at university), at the second level of the hierarchy are representations of extended events within a lifetime period (e.g. a camping holiday), and at the lowest level are representations of specific events, or parts thereof (e.g. a specific picture). To conclude, the authors posited that the important factor for memory retrieval within the hierarchy they described is what cues are provided.

Conway and Pleydell-Pearce (2000), in a similar vein to Burt et al. (1995), argued that memories are transitory mental constructions within a self-memory system. The authors described the self-memory system as containing both an autobiographical knowledge base, and the current goals of the working self. In their model, to recall a memory a control system activates a specific part of the autobiographical knowledge base (this knowledge base is stored as a series of connected nodes in a hierarchical structure, organized by lifetime periods and themes).

Williams, Conway, and Cohen (2008), departing somewhat from previous theories on autobiographical memory argued for the script theory of memory. In their view, memories are stored as scripts of behaviors and outcomes, structured within the memory system. These scripts include roles, goals, subscripts, and relevant/irrelevant actions, and are constantly reorganized following new experiences. Scripts within the memory system are assembled as required and new experiences are built from existing memories; common elements among scripts are represented on a higher level. Agreeing with previous research (e.g. Burt et al. 1995), in their view memory retrieval is elicited by cues.

Key Features of Human Autobiographical Memory

In reviewing the aforementioned papers, several key features have emerged that our model combines to create a reasonable model for human autobiographical memory that is firmly grounded in psychological theory:

1. Strengthening over time of memories with repetition (McClelland, McNaughton, and O'Reilley 1995; Meeter and Murre 2004).

2. Cues provided are a dominating factor and elicit memory retrieval (Burt et al. 1995; Williams, Conway, and Cohen, 2008).

3. A memory is activated by a control system, which activates a specific part of the autobiographical knowledge base (Conway and Pleydell-Pearce 2000).

4. The memory system is dynamic – structures are assembled as required and new experiences are built from existing memories (Williams, Conway, and Cohen 2008).

Other Models of AI Memory

Before presenting our model of autobiographical memory for AI agents, a brief review of related work in the domain of AI agent memory systems is called for.

Cyril Brom's research has focused recently on technical issues surrounding the function of AI agent memory. Burkert et al. (2010) proposed a model that incorporates timing information into the storage of AI agents' memories. Using a neural network with context nodes (describing the agent's internal/external state) alongside Cartesian nodes (representing objective time), their episodic memory system was able to forget the details of tasks an agent had performed while still preserving high-level information about the task itself. More recently, Brom et al. (2012) presented a computational model for the representation of spatial memory for intelligent virtual agents. Using pointing data from psychological experiments, their model was able to demonstrate the so-called disorientation effect (increase in pointing errors following disorientation), as well as lend support for the psychological hypothesis that there exist both egocentric and allocentric spatial reference frames.

Focusing on a different aspect of agent memory, Wan Ching Ho has explored the creation and structure of memories for believable agents. For example, Ho and Dautenhahn (2008) described an approach to create a coherent life story for intelligent virtual agents which they implemented in anti-bullying software. In their model, memories are stored in an XML database as a combination of abstract (summary of the event), narrative (detailed description of the event), and evaluative (the agent's psychological evaluation of the effects) components. Intelligent virtual agents using this model were able to 'describe' events they had witnessed using a language engine to convert the raw facts of an episodic memory into a narrative of the event. In another paper, Ho, Dautenhahn, and Nehaniv (2008) presented a computational memory architecture allowing complex virtual agents to retrieve meaningful information from their dynamic memories, increasing their survival rate in complex virtual environments. This architecture combined short term memory and long term memory, and outperformed purely reactive, short-tem only, and long-tem only memory systems in terms of AI agent survival time.

Finally, the work of Mei Yii Lim provides a psychologically-inspired computational memory architecture that uses spreading activation in a network alongside compound cues to recall memories. In the model

presented by Lim et al. (2011), an AI agent's episodic memory events are assumed to be nodes that are associated with one another via attributes forming a network; the higher the number of commonly shared attributes between two nodes, the more related those two nodes are. In the spreading activation mechanism, when an attribute in the memory network is activated, activation spreads along associative pathways to related attributes and nodes according to the frequency of their co-occurrence. In this way, agents can answer questions regarding events with which they have had no experience based on the rate with which other events had occurred in the past. Of the AI agent memory architectures reviewed here, this model is the most relevant precursor of our model; the connectionist network model we propose, however, extends Lim et al.'s (2011) work by adding a hierarchical structure to represent time, handling the integration of new events, and situating the memory model within an overall AI agent model.

Our Model

To create a model of autobiographical memory possessing all of the key features as described previously, we used a multi-layer connectionist network. Each layer in the model represents one of three hierarchically arranged memory pools; immediate memory, short term memory, and long term memory. Within each pool, a given memory pattern is represented as a collection of interconnected nodes; each node represents a key feature of the memory and has a unique level of activation.

In addition to the three networks, our model also employs a controller which we refer to as the memory manager. This manager performs several tasks: handling the integration of new events, activating nodes within the networks, and simulating the decay of memories over time (these tasks are explained in detail below).

Memory patterns themselves are comprised of six components:

ID: A unique identifier for the memory.

Keywords: The collection of objects, places, and characters referenced by the memory.

Type: The general content of the memory, for example social interaction or combat.

Emotional Valence: Whether the memory is happy, sad, scary, etc., determined by event appraisal.

Weight: How important the memory is. This value is determined by the importance of the NPC, and by the emotional valence of the memory. This value controls whether a newly created memory will move from one store to the next in the memory pool hierarchy.

Timestamps: Identifies when the memory was first created, last accessed, and last updated.

Our overall agent model, shown in Figure 1, has been adapted from our earlier work in Bailey et al. (2012). Agent behaviour (actions, dialogue, etc.) is determined by reactive and rational processing units that depend upon psychosocial state (such as personality traits, emotions/mood, and social ties to other agents) and memories retrieved through interactions with the memory manager. Reactive processing provides quick emotion, instinct, or need-driven reactions in response to appraised events and changes in state; rational processing conversely provides deliberated, goal-oriented behaviour achieved through a planning process influenced by emotion and other psychosocial factors as suggested in Damasio (1994). After actions are selected through these processes, they are injected back into the event system for distribution in the game. Further details can be found in Bailey et al. (2012).

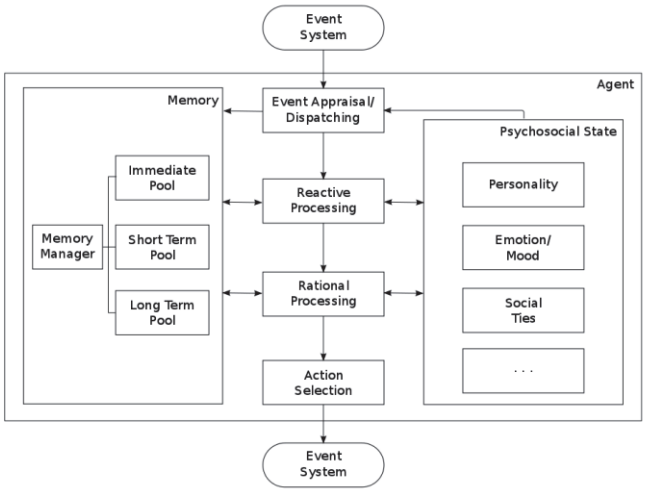

Figure 1. Overall Agent Model

Assumptions

Our model is designed to provide autobiographical memory for a specific type of agent, referred to as non-player or non-playable characters (NPCs). In the creation of our model, we have made two basic assumptions about these NPCs. We first assume that the NPCs differ in importance from one another. That is, some NPCs are more important than others to the storyline of the game, and therefore warrant the allocation of more resources in building their autobiographical memory stores (for a discussion of NPC importance, see Bailey et al. 2012).

More importantly, we assume that the game containing the NPCs will possess an event system capable of distributing in-game events (which can be likened to the NPCs' experiences) to each affected agent; agents can then appraise and dispatch events to the memory manager.

Adding New Memories

The memory model for each NPC is built iteratively by the memory manager, one memory pattern at a time. Memory patterns can come either from a preexisting database (for NPC creation) or from the in-game event system (for NPC maintenance). When the control system receives a new memory pattern the keywords from the pattern are first parsed into the immediate memory pool. If one of the pattern's keywords is not already represented as a node in the immediate memory pool network, a node for that keyword is created (duplicates are not allowed). Each node is connected unidirectionally to every other node in its memory pattern, so any two connected nodes are connected bidirectionally. For example, Figure 2 shows the immediate memory pool if it contained only two memories, *fishing at the lake by the cottage on vacation* and *going on vacation with Dad to Scotland.*

As the control system parses a pattern, if a connection between two nodes already exists, its strength is increased (key feature one); this allows a pair of nodes to become more strongly connected if many memories relate them. This increase in connection strength with multiple presentations is consistent with current views on cortical rewiring and information storage in the memory system (e.g. Frankland and Bontempi 2005; Chklovskii, Mel, and Svoboda 2010).

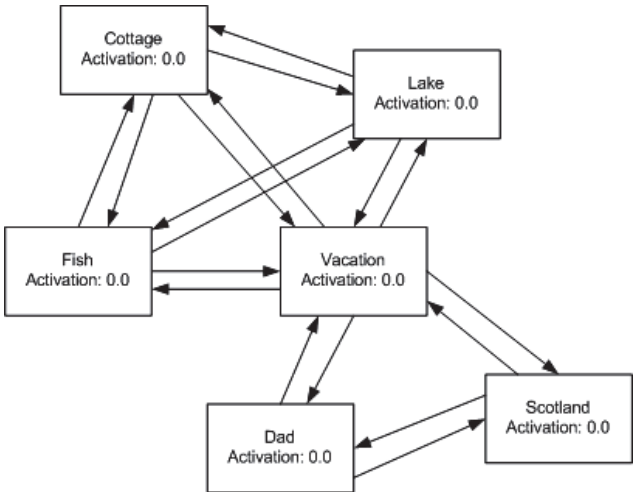

Figure 2. The Immediate Pool with Two Memories Stored.

This same algorithm is used to add memories to each pool of the memory model. Over time, the memory manager moves memories from the immediate memory pool to the short term pool and then to the long term pool. This allows the system to differentiate between memories which were recently recalled (in the immediate memory pool) and memories which have not been recalled for a long time (in the long term memory pool). The time before a memory is moved from one pool to the next in the hierarchy is flexible, and is tunable for each specific implementation of our system. Importantly, nodes within each pool are persistent, and remain even after the memory that was used to form them has been moved to another pool.

Retrieving Existing Memories

Memories are retrieved by prompting the NPC with a cue. For example, a user in dialogue with the NPC could provide a keyword as a cue to initiate memory retrieval. Such cues arrive as events to the agent and are appraised then dispatched to the memory manager. When provided a cue the memory manager iteratively checks if the nodes corresponding to the cued keywords exist in the immediate, short term, and long term pools; if found, the manager activates the matching node (key feature two and three) from the highest level of the hierarchy. In theory more than one keyword could be used to cue the memory manager, however at present our model is limited to one keyword.

When a given node is activated by either the memory manager or another node, if the incoming activation is greater than the node's current activation the node's activation increases as a weighted average of the incoming activation and its current activation value. Following this increase in activation, the node then passes an outgoing activation to all of the nodes it is connected to; this outgoing activation is a fraction of its own new activation value. Importantly, nodes do not 'backtalk' and spread activation back to the nodes that activated them, nor can a node's activation value be decreased by an incoming activation. The specific ratios used to handle incoming and outgoing activation can vary and are specific to a given implementation of the model.

To quantitatively assess which memory pattern a given pattern of activation in the network represents, the N most active nodes in the network are compared to a database of memory patterns to determine which pattern has the most keywords (by percentage of the total number of keywords in the pattern) in common with those N most active nodes. The pattern with the most keywords in common is deemed to be the recalled memory; if two patterns have the same number of keywords in common, one of them is selected at random as the recalled memory. Continuing the example above, with N = 4 we have a 50% match with the memory *fishing at the lake by the cottage on vacation*, and a 100% match with the memory *going on vacation with Dad to Scotland*, so the network would report that it recalls the latter memory. Figure 3 shows the four most active nodes after one time tick.

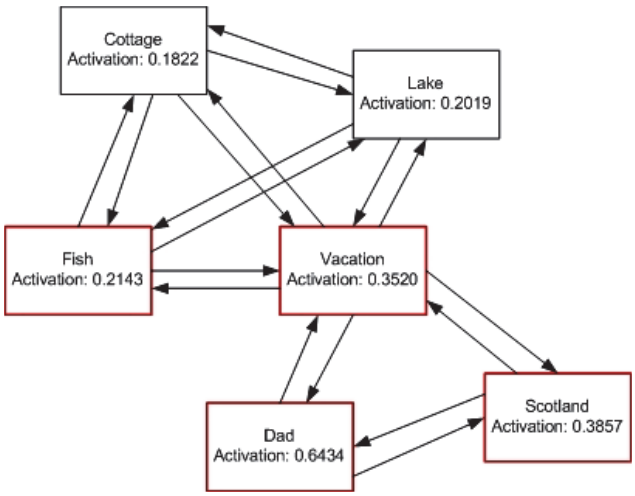

Figure 3. The N Most Active Nodes with N = 4.

Memory Decay

The passage of time affects NPCs' memories in two ways. After a memory has been activated by the memory manager based on a cue from the user, that activation will decrease over time. Each node's activation is decreased according to a bounded exponential function on a set time interval (this decrease is called a 'tick'). The function used to decrease a given node's activation is based on the forgetting curves described in Sikström (2000).

Secondly, our model includes a simulation of memory corruption. Over time, nodes within the long term memory pool are randomly changed to other members of the same conceptual category. In addition, low weight memories are randomly removed from the model; this both reduces the computational load and simulates the gradual forgetting of mundane events.

Prototype Minecraft Implementation

In order to validate our model, we created a prototype implementation in Java. Our implementation is flexible and portable; to assess its performance in a realistic setting however, we linked our model into the computer game Minecraft as a mod (Mojang 2009). We chose this platform because Minecraft has an avid modding community and natively supports Java code (MCP Team 2013).

Minecraft is a multi-platform multiplayer online first-person adventure/sandbox game. The game revolves around the player breaking and placing blocks. In addition to other players, users can fight hostile creatures and interact with friendly villagers. Villagers are very simple, exhibiting simple behavioural patterns governed by day/night cycles, and by default have almost no interactivity with the player. Given this obvious gap in player interactivity, we integrated our model of agent autobiographical memory into the villagers to enhance this aspect of the Minecraft player experience.

To begin, we used an automated script to create 50 NPC villagers with between 50-150 long term memories each. Using ten different memory templates corresponding to common events in the Minecraft game, we filled in the keywords for each memory template randomly in a 'mad lib' style. This allowed us to create an entire 'town' of NPCs each with a unique history and a unique long term memory pool. Once the NPCs were created, we used the in game event notification system to continue passing the NPCs new events to be integrated into their autobiographical memories.

Player interaction with the NPC villagers is done through a text chat window. The player cues the NPC with topics in the form of a singular keyword. This keyword is passed to the memory manager, which searches through the three memory pools for a matching node. If there is a match, the memory manager moves that memory from its current pool to the immediate memory pool, and activates the nodes in the immediate memory pool corresponding to that memory's keywords (as described above). If the player prompts the NPC with a keyword that is not in its memory model, the NPC replies that it does not know anything about that topic. Figure 4 shows the user cueing an NPC with *Creeper* (a type of hostile creature in the game), to which the NPC replies "*Salena Banks told me they walked through the Hills biome and ran into a Creeper*".

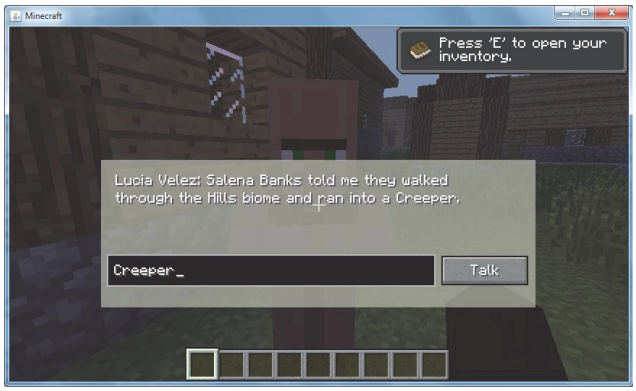

Figure 4. The In-Game User Interface.

Context-Dependent Recall

In addition to simply recalling a relevant memory when prompted by a cue from the player, our memory model allows NPCs to recall different memories depending on the conversational context of the cue. Consider an NPC with two different memories containing the keyword *wood*; *I found some wood in the hills biome* and *I traded some wood to Duncan McGrath*. One of these memories also contains the keyword *hills* while the other also contains the keyword *traded*.

If the player cues the NPC with the keyword *hills*, the memory manager will activate the hills node in the immediate memory pool, and this activation will spread to all of its connected nodes, and then out through the remainder of the network. If the player then cues the NPC with *wood*, because the hills node was activated on the previous time tick the NPC will recall the memory *I found some wood in the hills biome.*

If the player had instead first cued the NPC with *traded*, the traded node in the immediate memory pool would have been activated by the memory manager, and that activation would have spread first to all of its connected nodes and then out through the remainder of the network. If the player then cued the NPC with *wood*, in this case it would instead recall *I traded some wood to Duncan McGrath*.

Recency Effects

Our system is also able to model the difference between memories that have been recalled recently and those which have not been recalled for a long time. When a memory is accessed by the memory manager, it is moved to the immediate memory pool, and when new memories are created, they begin in the immediate memory pool. When the user provides an NPC with a cue, the memory manager first searches for a keyword match in the immediate memory pool, then the short term pool, and finally the long term pool. As such new memories and those which have been recalled recently, because they will be in the immediate pool, are the first to be recalled by the NPC.

Performance Considerations

We made several design decisions in our implementation to ensure that the Minecraft client would perform well even with the addition of our model. First, all NPCs' long term memory pools are serialized to disk. This is to avoid out of memory errors when many NPCs each with many memories are all simulated simultaneously. For simplicity, NPC memory databases and the lists of keywords corresponding to objects, places, and other NPC names are also stored as Java serialized objects.

Despite the increased computational load of running the autobiographical memory model, during testing we observed no decrease in Minecraft's frame rate on a dual core 1.7Ghz laptop with 2GB of RAM. This system satisfies the minimum requirements for the game, however we will be conducting more rigorous performance testing as the number of agents including our model increases. If performance becomes an issue under higher loads, techniques for simulating many complicated believable agents such as those presented in Rankin, Acton, and Katchabaw (2010) can be employed to ensure that scarce computational resources are allocated to the NPCs that deserve and need them while maintaining the performance of the game at a level acceptable to the player.

Other Applications

Further to improving the realism and performance of traditional NPCs in games, our model could be used to facilitate a gameplay dynamic whereby recent events and conversational context are relevant considerations for the player when influencing or predicting agent behavior. Importantly, this gameplay dynamic is not tractable with more naïve representations of agent event knowledge, which do not allow agents to respond dynamically and unpredictably based on recent events and conversational context.

Similarly, our model could be applied to chat bot design. A chat bot built based on this model could recall and discuss different memories depending on the conversational context. This would provide a much more complex, less predictable, and likely more realistic flow of ideas in the chat bot's responses than a naïve user input keyword-matching algorithm.

In both of these cases, a robust memory system has been recognized as a critical support system for enabling believable agent behaviour (Wheeler 2013).

Conclusions

Our model captures several characteristics of human autobiographical memory described in psychological literature, namely: nodal connections which strengthen with repetition, cues are the dominant factor modulating memory retrieval, and dynamic memory structure construction.

To demonstrate our model's validity, we implemented a prototype of our model in Minecraft. This prototype possessed several important features of human memory: rapid online integration of new memories, context dependent recall, and the effect of recency on recall. In sum, our model demonstrates the capability to simulate autobiographical memory at a level beyond traditional methods, and has the potential to serve as the foundation for more complicated believable agent systems.

There are many possible directions for continued work. As discussed earlier, our work could benefit from more rigorous performance testing of our prototype implementation under a variety of workloads. Integration of a complex and robust dialogue system alongside our prototype, such as the one in Wheeler (2013), also has interesting potential for creating more dynamic and compelling interactions. Lastly, to assess our model in the context of improving agent believability, extensive user testing with the Minecraft community is called for. While we have added an interesting missing dimension to villager NPCs in the game, we need to confirm and measure the extent to which this ultimately improves agent believability and user experience.

References

Bailey, C., You, J., Acton, G., Rankin, A., Katchabaw, M. 2012. Believability through psychosocial behaviour: Creating bots that are more engaging and entertaining. *Believable Bots: Can Computers Play Like People?* Springer.

Brom, C., Vyhnánek, J., Lukavský, J., Waller, D., & Kadlec, R. 2012. A computational model of the allocentric and egocentric spatial memory by means of virtual agents, or how simple virtual agents can help build complex computational models. *Cognitive Systems Research* 17(18):1-24. doi: 10.1016/j.cogsys.2011.09.001

Burkert, O., Brom, C., Kadlec, R., & Lukavský, J. 2010. *Timing in episodic memory: Virtual characters in action*. Proceedings of Remembering Who We Are - Human Memory for Artificial Agents Symposium Artificial intelligence and simulation of behaviour convention, Leicester, UK.

Burt, C. D., Mitchell, D. A., Raggatt, P. T., Jones, C. A., & Cowan, T. M. 1995. A snapshot of autobiographical memory retrieval characteristics. *Applied Cognitive Psychology* 9:61-74.

Chklovskii, D. B., Mel, B. W., & Svoboda, K. 2004. Cortical rewiring and information storage. *Nature* 431:782-788.

Conway, M. A., & Pleydell-Pearce, C. W. 2000. The construction of autobiographical memories in the self-memory system. *Psychological Review* 107(2):261-268.

Damasio, A. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain.* Putnam.

Frankland, P. W., & Bontempi, B. 2005. The organization of recent and remote memories. *Nature Reviews* 6:119-130.

Ho, W. C., & Dautenhahn, K. 2008. *Towards a narrative mind: The creation of coherent life stories for believable virtual agents*. International conference on intelligent virtual agents, Reykjavik, Iceland.

Ho, W. C., Dautenhahn, K., & Nehaniv, C. L. (2008). Computational memory architectures for autobiographic agents interacting in a complex virtual environment: A working model. *Connection Science* 20(1):21-65. doi: 10.1080/09540090801889469

King, K. 2002. A dynamic reputation system based on event knowledge. In S. Rabin (Ed.), *AI Game Programming Wisdom* (pp. 426-436). Hingham, Massachusetts: Charles River Media Inc.

Lim, M. Y., Aylett, R., Ho, W. C., & Dias, J. 2011. Human-like memory retrieval mechanisms for social companions. In Proceedings of the 10th international conference on autonomous agents and multiagent systems.

Loyall, A. 1997. Believable Agents: Building Interactive Personalities. Ph.D. diss., Stanford University, Stanford, CA.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102(3):419-457.

MCP Team. 2013. Main page: Minecraft coder pack. Retrieved from: http://mcp.ocean-labs.de/.

Meeter, M., & Murre, J. M. 2004. Consolidation of long-term memory: Evidence and alternatives. *Psychological Bulletin* 130(6):843-857.

Mojang. 2009. Minecraft. Retrieved from https://minecraft.net.

Rankin A., Acton G., & Katchabaw, M. 2010. A scalable approach to believable non player characters in modern video games. In Proceedings of GameOn 2010. Leicester, United Kingdom.

Sikström, S. 2002. Forgetting curves: Implications for connectionist models. *Cognitive Psychology* 45:95-152.

Tulving, E. 1972. Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381-402). Retrieved from http://web.media.mit.edu/~jorkin/generals/papers/Tulving_memory.pdf

Tulving, E. 1987. Multiple memory systems and consciousness. *Human Neurobiology* 6:67-80.

Williams, H.L., Conway, M.A. & Cohen, G. 2008. Autobiographical Memory. In G. Cohen and M.A. Conway (Eds.) *Memory in the Real World* (pp. 21-81). New York: Psychology Press.

Wheeler, K. 2013. Representing Game Dialogue As Expressions in First-Order Logic. M.Sc. Thesis., The University of Western Ontario, London, ON.