A cognitive architecture based on Deep Reinforcement Learning

Adrian Millea

September 19, 2018

1 Introduction

In recent years, the reinforcement learning community has been significantly enhanced by deep learning, with the advent of computational power easily accesible to the common researcher. Moreover, the seminal paper [Mnih et al., 2015] has shown that complex tasks can be solved, even to super-human performance. However, Deep Reinforcement Learning (DRL) agents still perform well on a limited number of tasks, and what's more, these tasks are generally treated separately. There are a few exceptions, where the multi-tasking is specifically considered, even though they are generally from the same domain, or have similar characteristics [Parisotto et al., 2015, Teh et al., 2017]. When looking at multi-modality we see that almost no research exists on integrating different input types into the same agent [Qureshi et al., 2016]. Many researchers see the DRL framework as paving the way to general AI, or strong AI, an agent that can match or surpass human intelligence in all its aspects. However, when looking at the human brain, we see that it has many components which are not at all incorporated in the current research directions. For example, long term memory is implemented in a DRL agent as a simple episodic buffer (we note that there exists smarter and more complex variants of it, which we will discuss in more detail later [Andrychowicz et al., 2017, Schaul et al., 2015], however the main principle is the same), which does not include semantic memory [Jacoby and Dallas, 1981, Tulving et al., 1972], temporal memory [DuBrow and Davachi, 2014], event memory [Burgess et al., 2001] etc, which we know exist in the brain and are surely definitory for the complex information processing we find in humans.

One other major component which we find in humans and we consider critical for complex abilities is the reward function, which can be arbitrarily constructed, as a function of the current task and the existing knowledge (we include here different types of memory) [Schultz, 2015]. However, no such implementation exists in the DRL literature. There are a few approaches which enhance DRL agents with carefully constructed reward functions [Jaderberg et al., 2016], or statistical properties of the environment [Machado et al., 2018, Tang et al., 2017], but this set is very limited and does not come close to what humans do.

We will discuss in more details all aspects which we believe are currently not treated in the DRL literature but we know from the cognitive science, neuroscience, psychology are critical to human reasoning. Thus, we base our work on the fact that many functional (we use this word with no special technical meaning, but how it is defined in the dictionary) aspects existing in the brain are not at all dealt with in the DRL literature. To support our approach we cite [Hassabis et al., 2017], but we note that this idea germinated before this publication.

1.1 Long Term Memory

It's obvious that a big part of what makes us human is our ability to remember complex things from our experiences, properties of objects, long sequences of events, their outcomes, other people's experiences, things we read in books, technical processes (how things work), etc. Obviously there is a significant problem when storing so much data, in whatever form: accessing it. The easier, faster, more accurately we access relevant (to the current context) data, the better we do at the current task. Integrating more data in a shorter amount of time has been long conjectured to be a signficant trait of consciousness and intelligence [Baars, 2005]. This is called memory retrieval and in the brain is manifested as a complex interplay between HippoCampus (HC) and PreFrontal Cortex (PFC) [Preston and Eichenbaum, 2013].

1.1.1 One single principle for all types of memory and all levels of hierarchy

It is well known in the literature that human have many levels of representations, for entities (objects, animals, persons, etc) and policies (sequences of actions with a specific goal). In the PFC these are generally referred to as basic, subordinate and superordinate categories. In the HC the different scales representation are generally associated with anterior (for coarse) and posterior (for fine-grained) HC. This is generally referred to as a hierarchical representation. There are quite a few new approaches which deal with hierarchy, however most of the time they talk about hierarchy but implement only two levels [Kulkarni et al., 2016, Bacon et al., 2017]. The lower is generally goal specific and the higher selects among the lower ones. However for arbitrary policies, concepts, etc we need an arbitrary number of levels. We propose here a very simple and natural general principle of representation which can be applied to all levels of the hierarchy and to any type of memory, be it semantical, temporal, event-based, etc. We call this *the neighbouring principle* and we describe it shortly here. The main idea is to look at an entity from the perspective of its neighbours. The definition of a neighbour is obviously dependent on the representation used, if it is temporal, then we look at neighbours in time, if it's semantic, we look at the relation between entities (an apple and a pear are related in a clear way, apple and juice are also related but in a different way, but an apple and a screwdriver are almost never related, maybe only as an exercise in imagination). So we see that we should look at the set of neighbours of a single entity but also at the nature of the neighbouring function. We will develop these concepts formally in a later section.

In the coginitive science literature they talk about and use *schemas* from an early work of Piaget [Piaget, 1923]. Interestingly work in ML has used the concept of schemas to develop a very successful architecture [Kansky et al., 2017] which outperformed all others on some specific tasks (e.g. captcha). Schemas are a type of prototype policies (e.g. at a restaurant one ought to order, eat and pay, in this order). We see that using the neighbouring principle we can easily describe such sequence, by defining the neighbouring set of entering a restaurant as a one element set, ordering, then its neighbour is a one element set, eating and so on. Moreover, we know, still from cognitive science that humans access different representations of the same entity as a function of the context, or the task at hand [Friston and Price, 2001]. This can be easily accomodated by changing the neighbouring function depending on the context. A specific context enables a small subset of neighbouring functions.

1.2 Arbitrary Reward

When it comes to rewards, humans can define for themselves arbitrary rewards as a function of the context and goals they have. Besides the physical needs, which don't apply to machines, as humans we even discover reward functions as we learn about the world. Thus we can safely state that the reward function is built on top of semantic memory, and uses it in arbitrary ways. Sometimes we play a game and we are just interested in the high score, however sometimes we maybe don't care about the high score but just to defeat one specific oponent, and we can change this on the fly. There is no such mechanism in current approaches to DRL. We strongly believe that this should be a defining characteristic of strong AI, the ability to construct its own reward functions depending on information in different types of memory and specific goals.

1.3 Constructive and prospective memory

It has been recently discovered [Schacter and Addis, 2007] that remembering things is not at all a retrieve only process, but a more constructive process, where different pieces are put together to form a coherent, plausible experience. That's why remembering is prone to so many alterations. Prospective memory is the process of planning ahead, of evaluating possible scenarios, and choosing the most desirable ones, based on we already know about the world (existing experiences). One major difference between remembering and prospecting from memory is quantitative. When prospecting, the amount of activity in the respective brain areas (the same brain regions are activated in both remembering and prospecting) is greater than when remembering [Addis et al., 2007]. We put forward a framework which obeys these constraints, and possibly explaining also the increased activity for prospecting memory. In the DRL literature the process of planning and evaluating possible outcomes is known as model-based (MB) reinforcement learning [Kuvayev and Sutton, 1997]. This means one should first construct a model of the world (similar to a simulation) and then based on that evaluate possible scenarios and outcomes. However, constructing a model is quite expensive and there are many choices of models [Atkeson, 1998, Doya et al., 2002, Ormoneit and Sen, 2002]. The choice of model for

each individual task is non-trivial and special expertise is needed to choose the right one, in terms of complexity, modeling power, predictive ability, etc. In our framework model-based mechanisms come out naturally and the problem with choosing a model dissappears, we choose a simple non-parametric model, based on existing memories. We leverage iterative memory retrieval for this process. Our approach is somehow consistent with [Kumaran et al., 2016].

1.4 Central executive and working memory

In cognitive science, the central executive (CE) [**?**] plays a major role in decision making, planning, retrieving memories and problem solving in general. It is considered to be part of the working memory system. In the DRL literature the general approach for its functionality is the last pool of neurons, just before the decision for action is made. However, in humans the CE can allocate different resources as a function of the task or goal, can replan if the current plan is not satisfactory, can retrieve certain memories relevant to the current setting and many other cognitive management tasks. It is a sort of conductor of the cognitive orchestra. Working memory is also a critical component in human reasoning which is mostly left out in the DRL literature, not to mention the differentation of working memory which is mostly agreed upon, i.e. visuo-spatial sketchpad, phonological loop, and episodic buffer (page 219).

We see that even at a high-level, the general DRL approach is lacking many critical components that make us humans, extensive and specialized memory systems, simple prospective abilities which use the same mechanisms as the remembering processes, working memory and a complex central executive that manages all resources efficiently and can change task, goals and approach on the fly. We strive in this work to bring these components to the DRL literature, even if we do it as simple functional variants, leaving performance for later. We strongly believe that what is lacking is an integrative framework which includes these complex components that enable humans to have their high degree of intelligence. There are many simple tasks on which neural networks and DRL agents have been outperforming humans for a while now, however when it comes to complex tasks and especially dealing with multiple tasks and multiple modalities, there are many gaps which we try to fill with this work.

2 Multi-modality, multi-tasking and training regime

The multimodal treatment in the machine learning community and especially in the deep reinforcement community is quite sparse. The multi-task literature is a bit richer but still in its infancy [**?**, **?**, Teh et al., 2017]. There is no research (to our knowledge) which deals with both multi-modality and multi-tasking, we thus want to address this problem. Moreover, we propose an agent which deals with more modalities at once and and in a continuous fashion, not by dealing with well crafted datasets as is usually the case (e.g. Imagenet, MNIST, Youtube-8M). There is no approach (to our current knowledge) in the deep reinforcement community which has this human-like methodology for training.

There exists however a few cognitive architectures where this methodology is emphasized (cite). As in [?] our perspective is that the pieces of the great general AI puzzle are already there, they are just not put together. And the majority of the community is focusing on improving the individual performance results (e.g. object recognition, game playing) and not on integrating the pieces into a coherent whole.

As opposed to the human environment, a digital agent has different modalities available to it. For input humans have visual, audio, tactile and olfactory, whereas a (basic) digital agent has only visual, audio. For output, humans have verbal and movement, whereas a digital agent has verbal as well as visual (a digital agent can learn to display visual data just just as it receives it, e.g. by means of an autoencoder, whereas humans have no such functionality, but they can describe visual data in terms of verbal cues) and can also have movement dynamics available to it, assuming we give it also access to a robot's actuators, but we'll assume we don't for the time being.

3 Background - Relation between episodic and semantic memory: the Hippocampus and Prefrontal Cortex

It has been concluded in the cognitive psychology literature and adjacent fields that semantic memory is a type of memory which is acquired from episodic memory by extracting things (concepts, entities) which do not change between episodes, in other words semantic memories are invariant to episodic change, or another formulation is that semantic memory is a form of integration of information from different episodes. The interplay between HC and PFC plays a major role in the encoding and retrieval of semantic and episodic memories. Reactivation of similar memories is mediated by hippocampal pattern completion processes [?]. It has been shown that hippocampal reactivation patterns are similar for spatial memories that share position information, spatial context [?] or even relationships [?]. This similarity reactivation plays a role in transfer learning when dealing with novel environment but which generate similar neural codes due to the similarity of the experience [?]. The same similarity of hippocampal activations with respect to temporal and conceptual proximity have been observed [?, ?, ?, ?, ?]

-place cells in HC, they support rapid synaptic modification, minutes, + neurogenesis = this supports episodic memory, minutes to hours + or transfer to neocortex -HC = autoassociative network, can retrieve the memory from partial inputs - HC enables short term memory (e.g. a song you remember a few minutes) = working memories (brief, transient memories constructed from the holding and manipulation of multiple pieces of information). - PFC supports semantic memory extracted from HC, which in turn is last that processes the input stream - more activity top-down, predictions from PFC, than bottom-up, the input stream - HC processes novelty and integrates it with existing memories, by combining existing smaller elements from a lower level of processing

Consolidation - transformation of episodic memories into semantic ones based on invariance between episodes. Try to explain the world as much as you can (reconstruction of different episodes, maybe a latent space where you model the distribution of episodes)

but look at things that do not change between episodes. or change little? (or on/off).

3.1 Flexible retrieval

3.1.1 Cues: rewards, goals, episodes, semantic entities

The cues for retrieving memories and associated desired or optimal behaviour for the current situation can be of many types. Sometimes just semantic memories suffice, however, often we need contributions for different types of memory system in order to infer correctly the current situation we are in and the appropriate action to select. Thus we need a flexible mechanism for retrieval which can accomodate the heterogeneity of experience and memory. However, we have to keep in mind that the multimodality of experience is essential for the human-like information processing, thus we propose to have separate memory and retrieval systems for each modality but also a unified memory model for the integration of all modalities.

4 Background: planning, decision making and working memory

Planning is an essential human trait, it is based on prediction of the sensory streams of data. Prediction has been concluded to be one of the most important characteristics of human reasoning and many architectures (either cognitive architectures or machine learning based) embed prediction fundamentally in their models [**?**, **?**, **?**]. However an important aspect of planning and prediction in humans is that it happens at many levels of abstraction, temporal and spatial resolution.

For humans, planning is done based on experience, and is an iterative process, a plan can take from seconds to hours and the plan itself can embed behaviour spanning hours, days or even years. How does this remarkable feat get accomplished? It's easy to see that once the spatio-temporal environment can be structured hierarchically such that the highest level representations spans long spatio-tempo-behavioural slices, planning is still an essentially computational process, where different trajectories are evaluated (e.g. like in MCTS) and then selected according to the desired outcomes. However planning is not always the same, sometimes interests or needs change and thus we have to have more value functions which evaluate plans. Evaluations should be made based on the specific policy chosen and on the available memories with respect to the current context. However after an initial evaluation, we might not be satisfied with the evaluation result and thus we would like to make another one, this time with a different set of memories, and some additional information we got from previous evaluation(s). Thus we might need to store some additional information in a short-term memory, or working memory, which we can later use to enhance planning, or can use it to integrate multiple plans. Thus, we see how the need for such a memory might arise, which can contain arbitrary information, either with respect to previously retrieved memories or evaluated plans. Semantic memory can be used here to narrow the set of retrieved memories based on

the current context. As has been argued in the previous section relating episodic and semantic memory..

Having a variable size short-lived working memory which can contain arbitrary information in an iterative planning and decision making-module enables a wide range of complex behaviour as seen in humans.

5 Background: Human Learning

5.1 Structure learning or statistical learning

5.2 Behaviour learning or policy learning

5.3 Learning to learn or meta-learning

6 Background: Affect

As humans we are driven by our physiological, emotional or psychological needs. But how can we describe these needs into a computer? Obviously we leave the physiological apart because firstly, it is not something which we want to pass to our digital progenies (one of the main reasons we strive to develop general artificial intelligence is exactly to free intellectual existence from the physiological aspects of existence) and secondly the principles of need and drive are the same, even in the emotional or psychological realms. Following the cognitive psychology literature we devise a simple 2 component model of the affect of drive and need in terms of valence (how positive or negative a certain state is, this is equivalent to pleasure or pain) and arousal (which is related to how calm or exciting a state is). These two components vary almost continuously with time, and are generally dependent on the environment and the states an agent encounters or discovers, however, it can also be completely independent of the environment at times when the environment is uninformative, thus enabling different behaviours even in an unchanging environment.

7 Overview

In short, we first start with **a new methodology for training** which looks and processes continuous data streams from all modalities in the same time, while also integrating the representation into a unified multi-modal experiential representation. Individual memory pools exist (for each modality) but also a unified memory pool exists which stores the experiental stream into a combined representation of all modalities. The individual and unified stores are related in meaningful ways, such that if data from only one modality is received, this can still be enough to identify the current context in the unified pool. The incomplete data retrieval scenario is often met in the machine learning community and is also known as pattern completion, but is generally dealt with considering a single modality, thus the missing data are for example some missing pixels in an image or missing frequencies in a soundsample. Pattern completion with respect to a unified

pool of memories integrating multiple modalities is extremely sparse in the literature [Barsalou, 2017]. The respective pools of memories are then structured into semantic memories, during a first learning phase (which we can loosely equate with the learning observed in sleep for humans). Semantic memories are then used to enhance the memory retrieval process to narrow down the possible behaviours allowed and desired in the current context. Then the planning process begins, which is an iterative process that retrieves sets of memories based on different cues, which can be rewards, goals, ...? In addition, the current context is critical for retrieving relevant memories. After retrieval, the evaluations process takes place, i.e. evaluating the different policies proposed by the planning module. After this, more planning can take place and so on. We see how similar to the Value Iteration algorithm [] our approach is from a high level perspective. We see how our architecture resembles a so called cognitive architecture (MANIC, SOAR, ACT-R, etc.) however we replace modules (memory, planning, etc.) which are normally functional models (often symbolic) carefully engineered and plugged in the bigger architecture, with different flavours of deep neural networks, thus we approach the architectural problem from a machine learning point of view instead of artificial intellligence, which is a greater set which encompasses many symbolic approaches to intelligence. We base our approach on a few critical observations: 1. many lower-level tasks denoting some sort of intelligence, or information structuring, have been solved (some to even super-human performance) by current approaches 2. one of the most important traits of intelligence is to store and retrieve data, the more data one can store and retrieve and the more fast and efficient it can do that, the more intelligent one is, assuming the appropriate information processing in place. What I mean by this is that obviously just storing and retrieving raw or random data is futile, one needs to extract relevant features (to compress and represent meaningfully) from it.

7.1 Main hypothesis

We conjecture that intellligence is a process which involves only two high-level operations: process and storage. By process we mean any type of transformation of the input data, for example feature extraction, or policy learning or memory retrieval. By storage we mean any type of set of similar items (high-dimensional vectors) which can then be queried or associated (processed) for a desired outcome.

concepts and objects are related through a relational nn, there are a finite amount of relations possible and from what other similar objects or concepts are related to one can extrapolate on unseen ones.

8 Cognitive maps

We conjecture that all types of concepts or data types are processed similarly in the brain as found in quite a few works in the cognitive psychology literature []. This means we don't differentiate in the way we process or store different types of encodings of the data, be it semantical, spatial or temporal. However because the nature of the data

itself is different we most account for these differences in some way such that the final way we obtain the representation is invariant to the data type. We propose to do this by specific functions over the data which enables us to process it in a similar manner for all types of data. We call this neighbouring functions and they enable us to create maps of the data we are dealing with using the same principle independent of the type of data. Moreover, constraining the locality of the data (assume that only a finite amount of states are accessible from any single state - we call these neighbours) we can even get comparable representations between different data types. The principle makes one critical observations: because we are dealing with more data points, by definition of the problem formulation, we look at the relation between data points and not at individual ones. Of course we know in this case we are in danger of being cursed by the dimensionality, but we just said that we assume a finite set neighbours for any point, in practice this set is actually quite small, even in real-world environments. Strong evidence from the cognitive science literature shows that similar

8.1 Semantic: What

As we said in the background section, semantic memory is a type of memory which is invariant to the changes in episodes. Looking for structure in the data stream, we consider semantic memory as the features that change but not with episodes. Obviously if something does not change is not interesting, is not worth remembering. However if something does change but just sometimes, that could be interesting. Then we structure the semantic entities based on proximity in time or space to each other, or the color similarity between them, or shape similarity. This is where the neighbouring function comes in. We see that by applying different neighbouring functions we get completely different semantic structures of the data, or semantic maps. Putting all these together we get a comprehensive representation of the semantic entities, giving rise to an analogue of the *what* pathway in the brain []. Something is meaningful if it gives some information, if it is useful for some sort of prediction. If something is completely random it is by definition not meaningful. Thus semantic entities should enable prediction of other entities. The process of extracting these entities from episodic memories is as follows. By looking at multiple episodes we consider two entities related if the presence of one increases the probability of the other. Hierarchically chaining this type of probability increasing entities should give us causality. Object-vector cells (Høydal et al., 2018)

8.2 Spatial: Where

Similarly, the analogue of the *where* pathway can be constructed by devising a spatial map of the environment as above. We look at spatial neighbours of actual observed states this time and structure the states into a spatial map. We make an important observation here. As the states or observations can be high-dimensional and there can be quite a significant number of different ones, since our representation is relational anyway, we can identify or denote the states with some simple indexing scheme and then the neighbouring representation will be based on this scheme, which will enable comparison

and further hierarchical structuring. Slowly we can get a hierarchical structure which has a larger spatial scale at each level.

8.3 Temporal: When

The temporal neighbouring process is simpler, since time can be represented by just one dimension. It is analogue to the *when* pathway in the brain [Battelli et al., 2007]. Structuring events based on time hierarchically is much simpler than for the spatial or semantic case. We can just increase the time window until the size is irrelevant, meaning there is no differentiation between states anymore.

8.4 Unified: what, where and when

A unified experience and associated memory pool is critical for integrating information, associating information from different modalities and pattern completion. Thus, we encode the three streams by concatenating them and then feeding them as inputs to an autoencoder which will compress them to be efficiently stored. By having the unified model and memory pool connected to all three individ

8.5 Reward pathway: what for

As we discussed earlier there are many possible rewards that can be constructed in an environment based on semantic data. Rewards are something which should be under the control of the agent, like intrinsic motivation, but until now, there is no research letting the reward definition function under the control of the agent (to our knowledge). We explicitly construct a reward network which can receive input from the planning module (to evaluate a possible plan, how rewarding it is). Moreover, the agent can construct reward functions on the fly based on semantic memory

References

[Addis et al., 2007] Addis, D. R., Wong, A. T., and Schacter, D. L. (2007). Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45(7):1363–1377.

[Andrychowicz et al., 2017] Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, O. P., and Zaremba, W. (2017). Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5048–5058.

[Atkeson, 1998] Atkeson, C. G. (1998). Nonparametric model-based reinforcement learning. In *Advances in neural information processing systems*, pages 1008–1014.

[Baars, 2005] Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, 150:45–53.

[Bacon et al., 2017] Bacon, P.-L., Harb, J., and Precup, D. (2017). The option-critic architecture. In *AAAI*, pages 1726–1734.

[Barsalou, 2017] Barsalou, L. (2017). Situated conceptualization: a framework for multimodal interaction (keynote). In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 3–3. ACM.

[Battelli et al., 2007] Battelli, L., Pascual-Leone, A., and Cavanagh, P. (2007). The 'when'pathway of the right parietal lobe. *Trends in cognitive sciences*, 11(5):204–210.

[Burgess et al., 2001] Burgess, N., Becker, S., King, J. A., and O'Keefe, J. (2001). Memory for events and their spatial context: models and experiments. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1413):1493–1503.

[Cuayáhuitl et al., 2016] Cuayáhuitl, H., Couly, G., and Olalainty, C. (2016). Training an interactive humanoid robot using multimodal deep reinforcement learning. *CoRR*, abs/1611.08666.

[Doya et al., 2002] Doya, K., Samejima, K., Katagiri, K.-i., and Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural computation*, 14(6):1347–1369.

[DuBrow and Davachi, 2014] DuBrow, S. and Davachi, L. (2014). Temporal memory is shaped by encoding stability and intervening item reactivation. *Journal of Neuroscience*, 34(42):13998–14005.

[Friston and Price, 2001] Friston, K. J. and Price, C. J. (2001). Dynamic representations and generative models of brain function. *Brain Research Bulletin*, 54(3):275–285.

[Hassabis et al., 2017] Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258.

[Jacoby and Dallas, 1981] Jacoby, L. L. and Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 110(3):306.

[Jaderberg et al., 2016] Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*.

[Kansky et al., 2017] Kansky, K., Silver, T., Mély, D. A., Eldawy, M., Lázaro-Gredilla, M., Lou, X., Dorfman, N., Sidor, S., Phoenix, S., and George, D. (2017). Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. *arXiv preprint arXiv:1706.04317*.

[Kulkarni et al., 2016] Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, pages 3675–3683.

[Kumaran et al., 2016] Kumaran, D., Hassabis, D., and McClelland, J. L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20 7:512–534.

[Kuvayev and Sutton, 1997] Kuvayev, D. and Sutton, R. S. (1997). Model-based reinforcement learning. Technical report, Citeseer.

[Liu et al.,] Liu, G.-H., Siravuru, A., Prabhakar, S., Veloso, M., and Kantor, G. Multimodal deep reinforcement learning with a novel sensor-based dropout.

[Machado et al., 2018] Machado, M. C., Bellemare, M. G., and Bowling, M. (2018). Count-based exploration with the successor representation. *arXiv preprint arXiv:1807.11622*.

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

[Ormoneit and Sen, 2002] Ormoneit, D. and Sen, S. (2002). Kernel-based reinforcement learning. *Machine Learning*, 49:161–178.

[Parisotto et al., 2015] Parisotto, E., Ba, J. L., and Salakhutdinov, R. (2015). Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.

[Piaget, 1923] Piaget, J. (1923). The symbolic thought and the thought of the child. *Psychology Archives*.

[Preston and Eichenbaum, 2013] Preston, A. R. and Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773.

[Qureshi et al., 2016] Qureshi, A. H., Nakamura, Y., Yoshikawa, Y., and Ishiguro, H. (2016). Robot gains social intelligence through multimodal deep reinforcement learning. In *Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on*, pages 745–751. IEEE.

[Schacter and Addis, 2007] Schacter, D. L. and Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1481):773–786.

[Schaul et al., 2015] Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

[Schultz, 2015] Schultz, W. (2015). Neuronal reward and decision signals: from theories to data. *Physiological reviews*, 95(3):853–951.

[Tang et al., 2017] Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, O. X., Duan, Y., Schulman, J., DeTurck, F., and Abbeel, P. (2017). # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2753–2762.

[Teh et al., 2017] Teh, Y., Bapst, V., Czarnecki, W. M., Quan, J., Kirkpatrick, J., Hadsell, R., Heess, N., and Pascanu, R. (2017). Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4496–4506.

[Tulving et al., 1972] Tulving, E. et al. (1972). Episodic and semantic memory. *Organization of memory*, 1:381–403.