A Working Theory of a Learned Model in a Partially Observable Environment for Cognitive Decision-Making

A Technical Report submitted to the Department of Engineering Systems and Environment

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Emma Graham

Spring, 2022

Technical Project Team Members

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Leidy Klotz, Department of Engineering Systems and Environment

Rider Foley, Department of Engineering and Society

A Working Theory of a Learned Model in a Partially Observable Environment for Cognitive Decision-Making

Emma Graham[1]

***Abstract*— To survive in our unpredictable, evolving world, cognitive beings learn to make decisions with the limited knowledge of the world they process. Reflective of an individual's view of the world, a cognitive decision-making model is explored in a partially observable, stochastic environment. The cognitive model uses the Partially Observable Markov Decision Process problem formulation, which is a framework for neurological models and considered implementable in neural circuitry [26] [16]. To structure a planning model comparable to that of DeepMind's MuZero in a partially observable environment, a belief function will translate the observations to a vector of belief states that will be discretized so as to be used as the observations of a MuZero-based machine learning algorithm [29]. The belief states are computed recursively from the previous belief state using Bayesian inference. Bayes rule is thought to capture the neurological and cognitive levels of reasoning [26]. Components of the planning, training, and action methods of the cognitive model will follow those of MuZero. The model could then be trained and act, in way parallel to that of MuZero, in a partially observable environment. Cognitive insights from a model structured in this form and additional considerations are discussed.**

I. INTRODUCTION

Decision-making is a basic cognitive process that selects a course of action, based on certain criteria, from a set of options [39]. As the only known form of general intelligence, cognitive decision-making is the inspiration for today's decision-making technologies. The current attempts to computationally capture this cognitive process is especially prevalent in the deep learning subset of artificial intelligence (AI).

Considered by AI experts to be a leading path to an artificial general intelligence, reinforcement learning is driven by maximizing rewards associated with objectives [29] [33]. Used as a foundation of reinforcement learning (RL), the Markov Decision Process (MDP) is a strong, ubiquitous mathematical framework that "models (an) optimal decision-making process in complex dynamic systems" for sequential decisions [32] [23] [2]. Though the MDP, problem formulation is representative of the sequential decision-making framework of our complex, stochastic world. However, the MDP does not adequately represent the perspective of a cognitive being. Individuals each have their own view of the world, literally, which is limited to their observations.

The Partially Observable Markov Decision Process (POMDP) inherently mimics an individual's incomplete knowledge of the world. This framework for decision processes is highly representative of our biological cognitive process, a *powerful analytical tool* itself, which strives to optimize an action-selection policy by making sequential decisions based on beliefs of the world [28] [1] [14]. When it comes to our beliefs of the world, experiments in areas of neurophysiology and psychophysiology indicate that the brain executes Bayesian inference probabilistic representations to gauge task-relevant quantities [19] [22] [11] [26]. Cognitive neuroscience uses POMDPs as the basis for human decision-making through neural circuitry and internal models [26] [18] [40].

Recent bounds in reinforcement learning and in finding solutions to MDPs and POMDPs, have enabled state-of-the-art learned models to outperform human experts in challenges in complex domains with unknown underlying dynamics [35] [34] [29] [33]. DeepMind's MuZero model is arguably the most advanced of these models. A working theory for cognitive modelling in a partially observable environment, that mirrors belief state updates with that of the brain's neural circuity, will use MuZero-based planning in the hopes to gain insight into cognitive decision-making.

II. PRIOR WORK

A. Neurological Models

The theory of POMDPs are used in neural models of action-selection and decision-making [26]. To survive in a constantly changing and uncertain environment, cognitive beings must solve problems while choosing actions based on noisy sensory information and incomplete knowledge of the world [8]. Neurophysiological and psychophysical experiments suggest that the brain relies on probabilistic representations of the world and performs Bayesian inference using these representations to estimate task-relevant quantities [19] [22] [11].

B. MuZero

In 2019, DeepMind published its MuZero model. Without the structure of rules, MuZero attains superhuman performance in dynamic and intricate strategy games, and is infamously the first model to master Atari games [29].

MuZero is a learned model that uses model-based planning and reinforcement learning to predict the quantities "most directly relevant to planning", the action-selection policy, the value function, and the reward; it then selects the policy that maximizes reward [29]. These relevant quantities are the policy $p_t^k \approx \pi(a_{t+k+1}|o_1,\ldots,o_t,a_{t+1},\ldots,a_{t+k})$, value function $v_t^k \approx \mathbb{E}[u_{t+k+1}|o_1,\ldots,o_t,a_{t+1},\ldots,a_{t+k}]$, and immediate

[1]Emma Graham is a mathematics and systems engineering student in the School of Engineering and Applied Science at The University of Virginia, Charlottesville, Virginia USA emmagraham@virginia.edu

reward function $r_t^k \approx u_{t+k}$, where u is the true, observed reward [29].

The model assumes a real MDP, $< \mathscr{S}, \mathscr{A}, \mathscr{T}, \mathscr{R}, \gamma >$, from which an embedding function transitions the model's observations to a hidden state in an abstract MDP, $< \widetilde{\mathscr{S}}, \mathscr{A}, \widetilde{\mathscr{T}}, \mathscr{R}, \gamma >$, which is the input for the recurrent neural network [8].

To predict the policy, the value function, and the immediate reward, the model connects three functions, $\mu_\theta(h_\theta, g_\theta, f_\theta)$:

1) representative function, $h_\theta(o_0, \ldots, o_t) = \widetilde{s_t^0}$,
2) dynamic function, $g_\theta(\widetilde{s^{k-1}}, a^k) = r^t, \widetilde{s^k}$,
3) prediction function, $f_\theta(\widetilde{s^k}) = p^k, v^k$.

The model is trained end-to-end to estimate the three relevant quantities in the effort to find a policy, $\pi : \mathscr{S} \rightarrow p(\mathscr{A})$, that maximizes the expected infinite sum of returns [29]. The optimal policy is the policy maximizing the infinite sum:

$$\pi^* = argmax_\pi \, \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t] \tag{1}$$

Given hypothetical future trajectories $a^1, \ldots, a^k$, the model selects the action from the recommended policy, $a_{t+1} \; \pi_t$ by the applying a Monte Carlo Tree Search algorithm to the latent states and internal rewards generated by the dynamic function [29].

The parameters are then trained by minimizing the error between the predicted policy p_t^k and the search policy π_{t+k}, then the predicted value v_t^k and the value target z_{t+k}, and finally the predicted reward r_t^k and the observed reward u_{t+k}, where $z_t = u_{t+1} + \gamma u_{t+2} + \cdots + \gamma^{n-1} u_{t+n} + \gamma^n v_{t+n}$ for n steps and $u_t \in \{win, draw, lose\}$ such that $win = 1, draw = 0, lose = -1$ [29].

An L2 regularization term is also added, giving the overall loss:

$$l_t(\theta) = \sum_{k=0}^{K} l^p(\pi_{t+k}, p_t^k) + l^v(z_{t+k}, v_t^k) + l^r(u_{t+k}, r_t^k) + c||\theta||^2 \tag{2}$$

where l^p is the loss function for the policy, l^v is the loss function for the value, and l^r is the loss function for the reward [29].

III. Cognitive Decision-Making Model Theory

The cognitive decision-making model uses the Partially Observable Markov Decision Process problem formulation. POMDPs are used in neurological models and considered implementable in neural circuitry [26].

A. Neural Model Structure Framework

The Partially Observable Markov Decision Process, POMDP, is generally defined as a tuple $< \mathscr{S}, \mathscr{A}, \mathscr{O}, \mathscr{T}, \mathscr{R}, O, \gamma >$, where

- $\mathscr{S}$ represents the unobserved state space,
- $\mathscr{A}$ the action space,
- $\mathscr{O}$ the observation space,
- $\mathscr{T}$ the state transition function, $\mathscr{T}^a_{s\,s'} = Pr(s_{t+1} = s' | s_t = s, a_t = a)$,
- $\mathscr{R}$ is the reward function, $\mathscr{R}^a_{s\,s'} = \mathbb{E}[r_{t+1} | s_t = s, a_t = a]$, where $r(s,a)$ specifies the immediate reward obtained when in state s and performing action a,
- O the observation function, $O^a_{s'\,o} = Pr(o_{t+1} = o | s_{t+1} = s', a_t = a)$, and
- γ is the discount function.

The objective of the POMDP model is to find at least one optimal policy, π^* where the policies map the belief state vector to an action, $\pi(b) = a$. The optimal policy is the policy that maximizes the value function, $V^\pi(b) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | b_t = b]$, the infinite-horizon sum of future rewards:

$$\pi^* = argmax_\pi \, V^\pi(b) \tag{3}$$

such that $V^\pi(b)$ is the value function, where $\gamma \in [0,1]$ is a discount factor specified by the environment and $t \in \mathbb{N}$ is a discrete time representation.

A discrete POMDP problem formulation is proposed as the framework for this cognitive model.

B. Belief States

The belief states are computed recursively from the previous belief state using Bayes inference. Bayes rules are thought to capture the neurological and cognitive levels of reasoning [26] [18]. "A prerequisite for a neural POMDP model is being able to compute the belief state b_t in neural circuitry" [26]. It is generally assumed that cognitive beings use the intuitive Bayesian method to update their belief states [16].

Computed recursively, Bayes theorem gives the current belief from the previous belief state:

$$b'_{s'} = \eta \; O^a_{s'\,o} \sum_{s \in S} \mathscr{T}^a_{s\,s'} b_s \tag{4}$$

where η is the normalization constant.

Notice that the Markov property is upheld in this process because belief states are only dependent on the previous belief state. Additionally, notice that belief states are probability distributions over the states.

C. How to Decide

Recall that in MuZero, the observed states in the real MDP are transformed by the representative function to a hidden state in an abstract MDP [34] [8]. This latent state is iteratively updated, through a recurrent process, in conjuncture with a hypothetical next action. This hidden state space has no constraints, facilitating flexibility and ingenuity [8].

For planning, the relevant quantities are similar to that of the MuZero model: the policy, $p^k_{\omega,t} \approx \pi_\omega(a_{\omega,t+k+1} | o_{\omega,1}, \ldots, o_{\omega t}, a_{\omega,t+1}, \ldots, a_{\omega,t+k})$, value function $v^k_{\omega,t} \approx \mathbb{E}[(u_{\omega,t+k+1} | o_{\omega,1}, \ldots, o_{\omega,t}, a_{\omega,t+1}, \ldots, a_{\omega,t+k}]$, and immediate reward function $r^k_{\omega,t} \approx u_{\omega,t+k}$ [29].

Recent advances in RL have provided novel extensions of Monte-Carlo Tree Searches (MCTSs), an algorithm for MDPs, to partially observable MCTS (PO-MCTS) [34]. Before these advances, and due to the intractability of

POMDP problems, a canonical POMDP solution method applied MDP "solvers" to a continuous *belief* MDP. In the POMDP problem formulation, the beliefs for each time t are probability vectors for the states at time t . So, using the n-dimensional probability vectors, that are the beliefs, as states in a fully-observable MDP would result in a continuous MDP with an n-dimensional state space. To use the beliefs as discrete states in a corresponding belief MDP, the state space of the belief MDP will need to discretized. The discretization of the belief MDP's state space allows the belief MDP to act as a *real* MDP assumed in MuZero. The transition from the cognitive model's observations to the discrete, belief state space will be performed by the belief function, represented by σ_ω.

The *belief* MDP will play the same role as the *real* MDP assumed by MuZero in the cognitive model. The belief MDP will be defined $< \mathscr{S}_\omega, \mathscr{A}, \mathscr{T}_\omega, \mathscr{R}_\omega, \gamma >$.

In MuZero, the real MDP states at time t, which are the observations at time t, are embedded as a latent state of a abstract MDP. In the cognitive model, the belief MDP states space, from the acquired belief vector at time t, will be embedded as an abstract state in an abstract belief MDP, defined $< \widetilde{\mathscr{S}_\omega}, \mathscr{A}, \widetilde{\mathscr{T}_\omega}, \mathscr{R}_\omega, \gamma >$. The embedding is carried out by, essentially, the embedding function of μ_θ, h_θ. The representative function of the cognitive model will be represented by h_ω.

Now in the hidden neural network, the three quantities deemed "most relevant to planning" are extracted. The dynamic function, g_ω, takes the latent state and a hypothetical action and determines the reward and generates the next latent state. The policies and the value functions are determined by the prediction function, f_ω, from the latent states. These functions on the abstract belief MDP framework most closely resemble that of MuZero's dynamic and prediction functions, g_θ and f_θ, respectively [29].

To predict the policy, value function, and immediate reward, the model connects the belief function with the representative, dynamic, and prediction functions, $\mu_\omega(\sigma_\omega, h_\omega, g_\omega, f_\omega)$. These functions described above are given again in the order they are connected in the enumeration below:

1) belief function, $\sigma_\omega(o_1, \ldots, o_t) = b$,
2) representative function, $h_\omega(b_0, \ldots, b_t) = s_t^0$,
3) dynamic function, $g_\omega(s_t^{k-1}, a^k) = r^t, s^k$,
4) prediction function, $f_\omega(s^k) = p^k, v^k$.

The way the model plans, explicitly the connection of the functions, can be visualized in time step t in Fig. 1.

D. Choosing

The value, policy, and reward are learned using the hidden, deep neural network [29]. To choose given the planning of the cognitive model's neural networks, the model will use a Monte-Carlo Tree Search (MCTS) [29]. The consideration of potential future action sequences by the hypothetical actions used by g_ω, enables choice of the best action, $\pi^*(s_t^0) = a_{t+1}$, using a MCTS by the cognitive model [29]. The recurrent

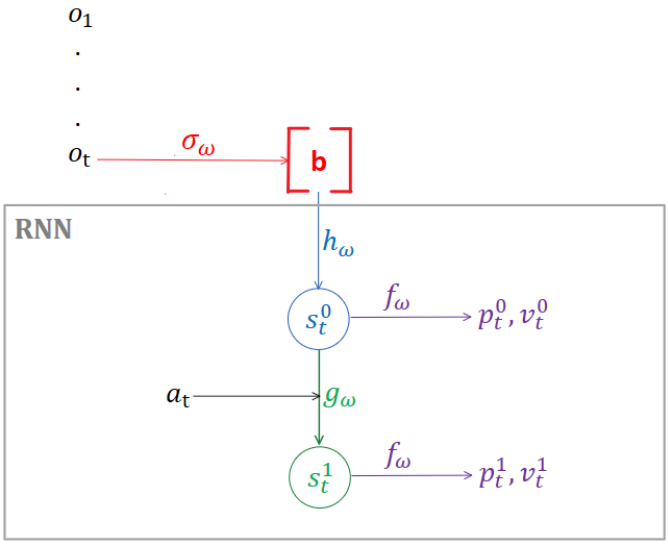

Fig. 1. How the model plans: the connection, $\mu_\omega((\sigma_\omega, h_\omega, g_\omega, f_\omega)$ of the belief function, $\sigma_\omega(o_1, \ldots, o_t) = b$, the representative function, $h_\omega(s_0, \ldots, s_t) = s_t^0$, the dynamic function, $g_\omega(s_t^{k-1}, a_t^k) = r_t^t, s_t^k$, and the prediction function, $f_\omega(s_t^k) = p_t^k, v^k$. [8] [29]

neural network unfolding during a MCTS is illustrated in Fig. 2 [29].

E. Taking Action

After the best action is selected, as outlined in the *Choosing* section, this action a_{t+1} is taken and the entire process started over with the new observations gleaned at this new timestep, $t = t + 1$. This process is visualized in Fig 3 [29].

IV. Discussion

A. Cognitive Insights

1) Field Contributions: Insights from this model structure, or its implications, would contribute to neuroscience. The lack of knowledge and understanding of the decisions the brain makes under uncertainty is a key problem facing the field of study [16].

By using the Bayesian approach for belief propagation that has been determined to "contribute to an understanding of the brain on multiple levels, by giving normative predictions about how an ideal sensory system should combine prior knowledge and observation, by providing mechanistic interpretation of the dynamic functioning of the brain circuit, and by suggesting optimal ways of deciphering experimental data", the theory for a cognitive model explored above gives a new perspective on the process of determining our next response from our interpretation of our current situation [11]. A way of interpreting the influence of possessing an incomplete understanding of the world, having beliefs, as opposed to having the complete knowledge of the environment could be viewed in how this model's results compare to that of MuZero, in a fully-observable environment. The results of this cognitive model could provide insights into the effects of decision-making with varied degrees of knowledge. These

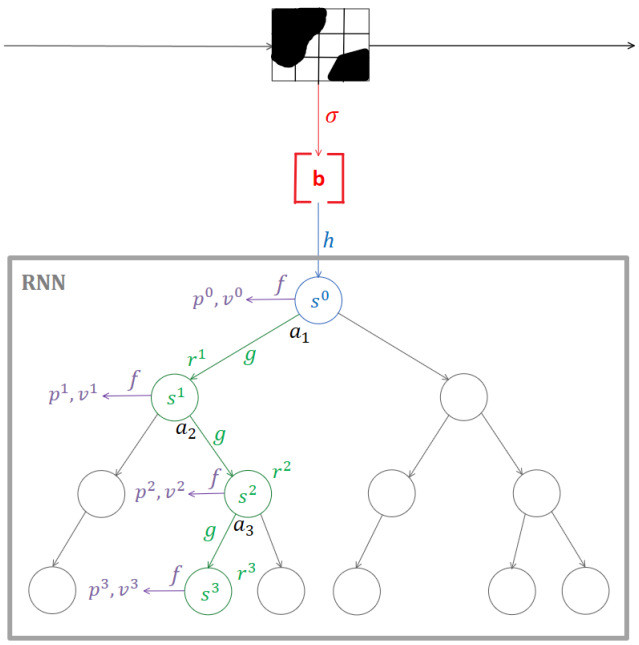

Fig. 2. How the model chooses the best action: The MCTS considering future actions in the recurrent neural network [RNN] [29].

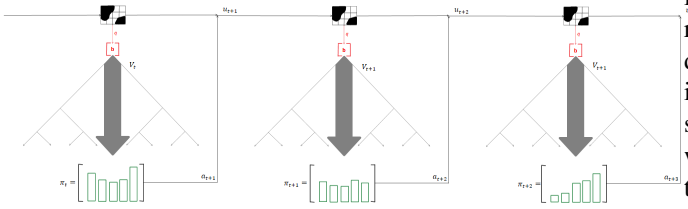

Fig. 3. How the model acts in the environment: At each timestep, the cognitive model goes through the decision-making process outlined in the *Planning* section. At this timestep, it then selections an action as outlined in the *Choosing* section with the MCTS (run on the recurrent neural network represented by the grey arrow). Finally, that action a_{t+1} is taken, bringing the model to the next timestep.

results could computationally back the importance of, or lack thereof, intelligence collection and dissemination.

2) Theory of Mind: As we increase our understanding of our own cognitive process of decision-making, we increase our understanding of how others make decisions. In cognitive science and psychology, the theory that describes our inferences about the psychological state of others is termed Theory of Mind (ToM) [24]. Deemed a cognitive mechanism, ToM combines intellectual abilities to enable cognitive beings to understand and interpret the beliefs, goals, plans, and intentions of others [20]. A significantly strong ToM, relative to another's, could enable the prediction of their future decisions. This could open the opportunity for manipulation as well as prove a central concept of the cognitive science in 1960s, called the Computational Theory of Mind (CTM). The CTM proposes that the mind is, itself, a computational system [27].

3) Social Interaction: A fundamental problem in cognitive neuroscience is the limited understanding of human-human interactions in making decisions. The POMDP framework has already been successfully employed in modeling human decision-making in games of a social context [18]. The decisions computed with POMDP problem formulation explain both the neurological and behavioral response of the subjects [18]. Giving cognitive models different or contrasting political, social, economic, and/or philosophical ideologies could simulate the interactions and decisions of individuals from societies with different or contrasting values. For instance, agents representing each member of the strategic triangle, the United States, China, and Russia, interacting in a simulated nonproliferation negotiation, could give insight into the underlying values, rewards, and optimal policies of each country in addition to unseen potential outcomes.

B. Limitations

1) Discretization: The MuZero model takes in discrete input states. Partially defined in the *Belief States* section, the proposed function σ_ω updates the belief state from the observations and then discretizes the belief space. The belief states were proposed to be updated via Bayesian inference but the method of discretizing the belief space is not discussed. Using the belief space as the state space results in a continuous MDP (coMDP). The reason for discretizing this space is to enable the use of a process that is essentially the MuZero model's groundbreaking action-selection process. Dicretization can be achieved in many ways. However, common methods of discretization result in the optimal policy and value functions for the discrete state space MDP which are not always, or not even typically, optimal for the original coMDP. This discrepancy creates a hesitation toward discretizing. It also elicits the question: Will the discretization of the belief space significantly effect the optimal policy and value functions of the cognitive model that is trained to predict the relevant policy and value function? And if so, how?

2) Training: This paper's working theory for a cognitive model focused on the *plan* for making decisions and not the action or training of the model. Games similar to first-person games that take place in a partially observable, stochastic environment would need to be created in order to train the model.

3) Solvers in the Partially Observable Environment: Recent contributions to the reinforcement learning field have given new convergence algorithms for "solving" POMDPs. Methods of extending Monte Carlo Tree Search (MCTS), including the upper confidence bounds for trees (UCT), from their MDP application to that of POMDPs have been published [34]. Solving a POMDP without the transformation to a MDP is another consideration for computationally exploring the cognitive process of decision-making.

C. Extensions

1) Multi-agent Interactions: Both MDPs and POMDPs are single-agent systems. The extension to multi-agent sys-

tems can be modeled with Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) [35]. Novel transformations of Dec-POMDPs to continuous-state MDPs have been recently published and could give further insight into our decision-making process [10]. Multi-agent systems would be a fascinating next step to explore, and one toward an artificial general intelligence.

ACKNOWLEDGMENT

Thank you to Group 011 of the Engineering Systems and Environment 2022 CAPSTONE. Professor Leidy Klotz, Cat Dunn, Alex Partridge, and Ryan Fruehwirth allowed each person in this team to dive fulling into their chosen topic of interest, *subtract*ing busy work and anything that did not inspire.

REFERENCES

[1] Alagoz, O., Hsu, H., Schaefer, A. J., Roberts, M. S. (2010). Markov Decision Processes: A Tool for Sequential Decision Making under Uncertainty. Medical Decision Making : an international journal of the Society for Medical Decision Making, 30(4), 474-483. HHS Public Access. 10.1177/0272989X09353194

[2] Banerjee, S. (2021). Real World Applications of Markov Decision Process. Towards Data Science. Retrieved 2021, from https://towardsdatascience.com/real-world-applications-of-markov-decision-process-mdp-a39685546026

[3] Bostrom, N. (2016). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

[4] Chades, I., Pascal, L. V., Nicol, S., Fletcher, C. S., Ferrer-Mestres, J. (2021, August 2). A primer on partially observable Markov decision processes (POMDPs) (S. Ramula, Ed.). Methods in Ecology and Evolution, 12(11), 2058-2072. https://doi.org/10.1111/2041-210X.13692

[5] Cherry, K. (2020, June 3). What Is Cognition? Verywell Mind. Retrieved December 3, 2021, from https://www.verywellmind.com/what-is-cognition-2794982

[6] Cowan, N. (2014). Working Memory Underpins Cognitive Development, Learning, and Education. Education Psychology Review, 26(2), 197-223. HHS Public Access. 10.1007/s10648-013-9246-y

[7] Dabney, W., Kurth-Nelson, Z. (2020, January 15). Dopamine and temporal difference learning: A fruitful relationship between neuroscience and AI. DeepMind. Retrieved March 1, 2022, from https://www.deepmind.com/blog/article/Dopamine-and-temporal-difference-learning-A-fruitful-relationship-between-neuroscience-and-AI

[8] de Vries, J. A., Voskuil, K. S., Moerland, T. M., Plaat, A. (unpublished). Visualizing MuZero Models. https://arxiv.org/pdf/2102.12924v2.pdf

[9] Diana, F. (2021, June 16). The Road To Artificial General Intelligence. Reimagining the Future. Retrieved April 5, 2022, from https://frankdiana.net/2021/06/16/the-road-to-artificial-general-intelligence/

[10] Dibangoye, J. S., Amato, C., Buffet, O., Charpillet, F. (2015). Exploiting Separability in Multiagent Planning with Continuous-State MDPs (Extended Abstract). International Joint Conference on Artificial Intelligence, 24.

[11] Doya, K., Pouget, A., Rao, R. P. N., Ishii, S. (Eds.). (2007). Bayesian Brain: Probabilistic Approaches to Neural Coding. MIT Press.

[12] Fard, M. M., Pineau, J. (2009). MDPs with Non-Deterministic Policies. Advances in neural information processing systems, 21, 1065-1073. US National Library of Medicine National Institutes of Health. Retrieved 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103230/

[13] Glen, S. (2017, August 25). Memoryless Property: Definition and Examples in Statistics. Statistics How To. Retrieved October 29, 2021, from https://www.statisticshowto.com/memoryless-property/

[14] Hahsler, M., Kamalzadeh, H. (2021, May 20). POMDP: Introduction to Partially Observable Markov Decision Processes. CRAN. Retrieved October 27, 2021, from https://cran.r-project.org/web/packages/pomdp/vignettes/POMDP.html

[15] Hollinger, G. (2007). Partially Observable Markov Decision Processes (POMDPs). Carnegie Mellon Computer Science. Retrieved October 27, 2021

[16] Huang, Y., Rao, R. P. N. (2013, January 22). Reward Optimization in the Primate Brain: A Probabilistic Model of Decision Making under Uncertainty. PLoS One, 8(1). National Library of Medicine. 10.1371/journal.pone.0053344

[17] Jonsson, A., Barto, A. (2007). Active Learning of Dynamic Bayesian Networks in Markov Decision Processes. University of Massachusetts Publications. Retrieved 2021.

[18] Khalvati, K., Park, S. A., Dreher, J.-C., Rao, R. P.N. (2016). A Probabilistic Model of Social Decision Making based on Reward Maximization. Advances in Neural Information Processing Systems 29 (NIPS 2016). Retrieved 2022.

[19] Knill, D. C., Richards, W. (Eds.). (1996). Perception as Bayesian Inference. Cambridge University Press.

[20] Korkmaz, B. (2011, January 5). Theory of Mind and Neurodevelopmental Disorders of Childhood69. Pediatric Research, 69, 101-108. https://doi.org/10.1203/PDR.0b013e318212c177

[21] Lawer, G. F. (2006). Introduction to Stochastic Processes (2nd ed.). Chapman and Hall/CRC.

[22] Lewicki, M. S., Rao, R. P.N., Olshausen, B. A. (Eds.). (2002). Probabilistic Models of the Brain: Perception and Neural Function. MIT Press.

[23] Littman, M. L. (2001). Markov Decision Processes. Elsevier. https://doi.org/10.1016/B0-08-043076-7/00614-8

[24] McCarthy-Jones, S. (2019). The Autonomous Mind: The Right to Freedom of Thought in the Twenty-First Century. Frontiers in Artificial Intelligence, Technology and Law.

[25] Prezenski, S., Brechmann, A., Wolff, S., Russwinkel, N. (2017, August 4). A Cognitive Modeling Approach to Strategy Formation in Dynamic Decision Making. frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01335

[26] Rao, R. P.N. (2010, November). Decision making under uncertainty: a neural model based on partially observable Markov decision processes. Frontiers in Computational Neurosceince, 4(146). 10.3389/fncom.2010.00146

[27] Rescorla, M. (2015, October 16). The Computational Theory of Mind (Stanford Encyclopedia of Philosophy). Stanford Encyclopedia of Philosophy. Retrieved April 7, 2022, from https://plato.stanford.edu/entries/computational-mind/

[28] Santos, L. (2020). Markov Decision process - Artificial Intelligence. GitBook. Retrieved October 29, 2021.

[29] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., Silver, D. (2019, November 19). Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. DeepMind. Retrieved April 6, 2022, from https://www.deepmind.com/publications/mastering-atari-go-chess-and-shogi-by-planning-with-a-learned-model-2

[30] The Senses — Biology for Majors II. (n.d.). Lumen Learning. Retrieved December 1, 2021, from https://courses.lumenlearning.com/suny-wmopen-biology2/chapter/the-senses/

[31] Shani, G. (2007). Learning and Solving Partially Observable Markov Decision Processes [Dissertation]. Ben-Gurion University of the Negev. Retrieved 2021, from https://www.bgu.ac.il/ shanigu/Publications/Dissertation.4.pdf

[32] Si, N., Zhang, F. (2017, December 2). MSE 310 Course Project II: Markov Decision Process. Stanford Class Material. Retrieved October 27, 2021.

[33] Silver, D., Baveja, S., Precup, D., Sutton, R. (2021, May 12). Reward is Enough. DeepMind. Retrieved April 5, 2022, from https://www.deepmind.com/publications/reward-is-enough

[34] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D. (2017, October 19). Mastering the game of Go without human knowledge. Nature, 550, 354-359. https://doi.org/10.1038/nature24270

[35] Silver, D., Veness, J. (2010, December). Monte-Carlo Planning in Large POMDPs. Twenty-fourth Conference on Neural Information Processing Systems. http://jveness.info/publications/nips2010

[36] Suchow, J. W., Griffiths, T. L. (2016). Deciding to Remember: Memory Maintenance as a Markov Decision Process. Department

of Psychology, University of California, Berkeley, Berkeley, USA. Retrieved 2021.

[37] Thomas, P. S., Okal, B. (2016, September 8). A Notation for Markov Decision Processes. Retrieved October 29, 2021, from https://arxiv.org/pdf/1512.09075.pdf

[38] Walch, K. (2020, October 23). General AI vs. narrow AI comes down to adaptability. SearchEnterpriseAI. Retrieved December 3, 2021, from https://searchenterpriseai.techtarget.com/feature/General-AI-vs-narrow-AI-comes-down-to-adaptability

[39] Wang, Y., Ruhe, G. (2007). The Cognitive Process of Decision Making. International Journal of Cognitive Informatics and Natural Intelligence, 1(2), 73-85. 10.4018/jcini.2007040105

[40] Wu, Z., Schrater, P., Pitkow, X. (2022). Inverse POMDP: Inferring What You Think from What You Do. Retrieved 2022, from http://xaqlab.com/wp-content/uploads/2018/05/InversePOMDPs.pdf