5-7-2019

Machine Awareness

Steven D. Harbour

Jeffery D. Clark

William D. Mitchell

Krishnamurthy V. Vemuru

MACHINE AWARENESS

Dr. Steven D. Harbour
Dr. Jeffery D. Clark
Mr. William D. Mitchell
Riverside Research
2640 Hibiscus Way, Beavercreek, Ohio 45431, USA
Dr. Krishnamurthy V. Vemuru
Riverside Research
2900 Crystal Dr., Arlington, Virginia 22202, USA

Current and future research that embodies a pathway to achieving machine common sense (MCS), including Capsule Neural Networks, Hebbian Plasticity Theory, Dual Process Theory, and machine awareness (MA). The final frontier may well involve a framework that is capable of machine curiosity, exploration, automatic self-direction and adaptation. The artificial intelligence (AI) system of the future will possess an innate curiosity and explore its own environment to gain knowledge, exhibiting a basic element of human cognition and awareness. The resulting MA system will possess inherent self-driven curiosity and related entropy in the decision space as it explores the environment in much the same manner as humans do.

Since the onset of AI research, MCS has been a crucial, yet missing and elusive component (Gunning, 2018). Common sense reasoning among machine systems has been unattainable. "Ultimate Machine Awareness" is a singularity that involves producing a human-made machine that is itself humanoid. Being self-aware involves knowing presence of self and consciously knowing one's own character, feelings, motives, and desires. Advances in AI will eventually lead to ultimate MA (consciousness). Providing AI systems human-like reasoning will enable human – humanoid symbiotic partnerships. Linking Hebbian Plasticity Theory (Brown, Zhao, & Leung, 2009; Widrow et al., 2019) and Capsule Neural Networks (Sabour, Frosst, & Hinton, 2017) provides a potential foundation for achieving MCS and MA by emulating the human mind. Similarly, Dual-process Theory (Evans & Stanovich, 2013) that is premised on System 1 autonomous processes and System 2 explicit processes is typically associated with consciousness that is emulating the human. A common framework for AI can be based in MA, constructed from third generation neural networks [Spiking Neural Network (SNN); Widrow et al., 2019] with the basic building blocks acting as Temporary Memory Neurons within the framework of the Leaky Integrate and Fire (LIF) Neuron Model learning from rules such as Hebbian Learning, capsule learning, or reinforcement learning. Capsule networks have a machine learning process that is similar to Hebbian Plasticity in neuroscience. SNNs demonstrate efficient learning with the integration of memory and information processing mechanisms. Like Hebbian learning, a higher-level capsule will receive its input from a lower-level capsule that shares affinity based on the largest scaler product of the activity vectors as a prediction coming from that lower-level capsule (Sabour, Frosst, & Hinton, 2017). A capsule can consist of a group of neurons whose outputs represent different properties of the same entity, e.g. a capsule layer built upon multiple capsule layers for perceptiveness. SNNs have the computational capability to continuously process spike trains to respond rapidly and accurately.

However, most SNNs do not retain the information of the spike train of a previous time step because the spike information is not retained once it causes a neuron to fire. One novel approach for machine consciousness could stem from a variation of the LIF neuron model that allows the spike train of the previous time step to be remembered. This modified LIF neuron approach has adjustable parameters that govern the 'remembering' time frame and the 'forgetting' time frame (Clark et al., 2019).

Capsule Neural Networks and Hebbian Plasticity Theory

We explore the connectivity between Hebbian Plasticity Theory and Capsule Neural Networks as a starting place for achieving MCS by emulating the human mind. Both of these learning frameworks are based on an adaptive neural computation in which manipulation of synaptic weights allow for strengthening or weakening of synaptic connections between neurons (W.D. Mitchell, personal communication, Feb 15, 2019). Hebbian Theory is biologically based and uses the concepts of long-term potentiation (LTP) and long-term depression (LTD) for the strengthening and weakening mechanisms, respectively (Brown, Zhao, & Leung, 2009). In Hebbian Theory, neurons connect when learning and self-organizing due to LTP and LTD. Consequently, by Hebbian principles: "units that fire together, wire together." (abstract, Widrow, et. al, 2019). Furthermore, neurons that fire out of sync weaken the link. Capsule Theory is based on artificial neural networks and uses a cluster of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part.

Capsules use non-linearity procedures to convert the set of activation probabilities to estimate the membership of a capsule to a post-capsule (Sabour, Frosst, & Hinton, 2017). Each capsule will learn and store discriminating biases potentially analogous to working memory. It iteratively adjusts the means, variances, and activation probabilities of the capsules and resulting outputs in potential similarity to Dual-process Theory (System 2 working memory).

Dual Process Theory

Dual-process Theory is premised on the idea that human behavior and decision-making involves System 1 autonomous processes that produce default reflexive responses. These reflexive responses involve an implicit process unless interceded upon by a distinctive higher order reasoning processes. System 2 involves an explicit process and burdens working memory which is typically associated with consciousness. Studies have shown that System 1 or 2 decisions are influenced by neurocognitive capabilities, knowledge, and working memory (Harbour & Christensen, 2015).

Qualia Exploitation of Sensory Technology (QuEST) is an innovative approach to improve human-machine team decision quality over a wide range of stimuli (handling unexpected queries) by providing computer-based decision aids (software and hardware) that are engineered to provide both intuitive reasoning and "conscious" context sensitive thinking (Rogers, 2019). QuEST provides a mathematical framework to understand what can be known about situations to facilitate prediction of the future state to make a particular decision. In so doing, QuEST is additionally utilizing this emerging theory Dual-process or Dual-system theory (Evans & Stanovich, 2013). It is premised on the idea that human behavior and decision-making involves autonomous processes (Type 1) that produce default reflexive responses involving an implicit process unless interceded by distinctive higher order reasoning processes (Type 2). Type 2, on the other hand, involves an explicit process and burdens working memory. Type 2 is

typically associated with: controlled, conscious, and complex. The Harbour and Christensen (2015) study compared Type 1 and Type 2 decisions made by pilots in actual flight, and assessed the impact of these decision types on cognitive workload and situation awareness (SA) under the enhanced-Theoretical Model of Situtation Awarness (TMSA) (Harbour & Christensen, 2015).

The query (Harbour & Christensen, 2015; Rogers, 2019) as the act of a stimulus being provided to an agent has characteristics that completely capture the salient axes (keep in mind what is salient in a stimulus is agent-centric) of the stimuli. Some of those axes are captured by an agent in its conversion of that stimuli into data (agent-centric internal representation of the stimuli). The term query is used instead of stimuli to capture the idea that a given agent must injest the stimuli and appropriately respond (thus an action). That response may be to just update its representation or not, or may be to take an action through an agent's effectors.

The unexpected query (UQ) (Harbour & Christiansen, 2015; Rogers, 2019) is an unexpected stimulus being provided to the agent and with the uncontrolled nature of in-flight events, pilots encounter unexpected queries and have to engage in both types of processing on any given flight. The enhanced-TMSA predicted that pilots with stronger perceptual and attentive capabilities needed to engage the arduous Type 2 system less, thus preserving spare capacity for maintaining SA. During 24 flights, there were UQ encountered by the pilot as well as expected queries (EQ) based on mission events and environmental stimuli. Results indicated that differences in workload and SA assessed both subjectively and through neurocognitive means existed. As UQ are encountered cognitive workload increases and SA decreases. During UQ working memory can become burdened leading to deficits in SA, which can be moderated by individual differences in perceptual and cognitive ability. Moreover, results from the research accomplished by Harbour and Christensen (2015) support Dual-process theory and assists in the development of the Theory of Consciousness (Fig. 1).

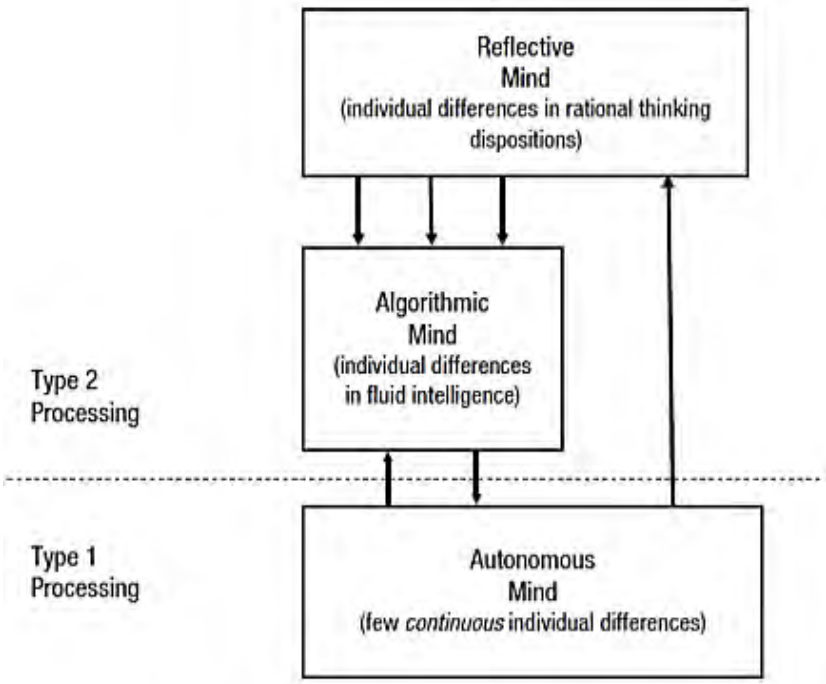

Figure 1. Conceptual Diagram of the Dual-Process Theory (Source: Evans & Stanovich, 2013).

Machine Awareness

The final frontier of MA will involve a framework that is capable of curiosity, exploration, automatic self-directed real-time learning and adaptation, learning with little data, learning with no labels, and learning rapidly. This framework will provide the structure, theory, and method for artificial reasoning, intuition, and human-like cognition. The artificially intelligent system proposed will possess an innate curiosity and explore its own environment to gain knowledge, exhibiting a basic element of human cognition and awareness.

The concept and model will demonstrate machine awareness and will be the building block for artificial general intelligence. Artificial general intelligence (AGI) is the ability for a machine to possess curiosity, investigate, think, reason, learn and be as dynamic, cognitively flexible and skillful as a human. It has been theorized that the ultimate MA singularity event is the moment true AGI will be achieved. This paradigm has brought about what is being called the Third Wave of AI. The Third Wave of AI boasts of the ability to develop machine cognition in the sense that it will be able to have general learning ability, self-directed learning, learn on the order (speed and accuracy) of the human biological brain, know when it needs to learn, and possess abstract reasoning, self-awareness, and cognition. Likely, combining the Law of Conservation of Energy and 3rd Generation Machine Learning, e.g. a SNN, may be a solution to this paradigm. One method is to use computational models of spiking neurons and synapses implementing the biologically plausible Leaky Integrate-and-Fire (LIF) model (Dayan, Abbott, & Abbott, 2001). This simulates the dynamics of a spiking neuron that is driven by the input spikes through plastic synapses. The LIF neuron processes the input spikes modulated by the inter-connecting synaptic weights, leading to a change in its membrane potential. The Spike Response Model zero order (SRM0), equation below, is a generalized LIF model where the model parameters are temporally based from the last output spike (Gerstner & Kistler, 2002). The membrane potential of the LIF model is given by:

$$\boldsymbol{u_i}(\boldsymbol{t}) = (\boldsymbol{t}-\hat{\boldsymbol{t_i}}) + \Sigma_j \boldsymbol{w_{ij}} \Sigma_{\boldsymbol{f}}\, \boldsymbol{\epsilon_i}(\boldsymbol{t}-\boldsymbol{t_j^{(f)}}) + \boldsymbol{u_{rest}}\ ,\ \text{(Gerstner \& Kistler, 2002)}$$

where $u_i(t)$ is the membrane potential of neuron i at time t, and $\eta(t-\hat{t_i})$ is the model 'form' of the spike at some time t after the last spike of neuron i ($\hat{t_i}$), $\Sigma_j w_{ij}$ is the synaptic efficacy (the sum of the synaptic weights of the presynaptic neurons j exciting the postsynaptic neuron i), $\Sigma_{\boldsymbol{f}}\, \boldsymbol{\epsilon_{ij}}(\boldsymbol{t}-\boldsymbol{t_j^{(f)}})$ is the sum of the postsynaptic potential (ϵ_{ij}) based on the current time and its relation to the presynaptic spikes of presynaptic neurons j at time f, and the membrane resting potential of neuron i (u_{rest}).

Recent neuroscience views curiosity as being internally motivated and absolutely intrinsic (Sharpee, Calhoun, & Chalasani, 2014; Loewenstein, 1994). In neuroscience, according to Incongruity Theory, when the human is presented with something that it has not seen before and therefore does not understand, human curiosity becomes stimulated, excited, and motivated (Kidd & Hayden, 2015; Loewenstein, 1994). Typically, the environment is viewed by the human as being predictable and orderly; this paper refers to that as being in a Low Entropy state: - H. Incongruity Theory states that when this order is challenged and the environment is poorly understood and not perceived to be predictable, something like curiosity is aroused (Friston & Buzsáki, 2016); this paper considers that as being in a High Entropy state: + H. According to Friston (2018), a want to reduce it occurs, (curiosity is aroused then exploration is stimulated according to Harbour, 2019), when a concept referred to as variational Free Energy (ΔF) is high.

To date, these theories and concepts have not been demonstrated in a machine; however, they have shown support in studies involving humans. The relationship of FE for Entropy (H) is:

$$(\boldsymbol{q}) = \langle(\boldsymbol{x})\rangle\boldsymbol{q} - \boldsymbol{H}[\boldsymbol{q}(\boldsymbol{x})],$$ (Friston & Buzsáki, 2016)

Where q is the probability density of belief and ⟨⟩ is the expectation of the total energy (E) of the environment with respect to q and H is the entropy of the density of belief of the environment. Entropy (H) and Free Energy (FE) can be viewed as the absolute difference between what is known and what is not known, or stated in another way, what is believed to be known and what is actually known. When (FE) is high then intrinsically curiosity is stimulated, resulting in exploration and learning occurring until (FE) reaches a minimum and (H) also has reached a minimum. The novel artificially intelligent system will need to possess both inherent self-driven curiosity and related changes in Free Energy as it explores the environment in much the same manner as humans do, thus demonstrating machine awareness.

Conclusion

It will take a team, bringing together multiple disciplines: electrical and computer engineering, neuroscience, neuroergonomics, cognitive engineering, psychology, physiology, biology, physics, philosophy, and mathematics in order to solve this arduous and nearly insurmountable quest of artificial general intelligence. Currently, the exact path toward MCS is not known, however, Capsule Neural Networks, Hebbian Plasticity Theory, Dual Process Theory, and MA, have one major characteristic in common, and that is they all use a biomimetic approach to some degree. As has been discussed it may require a blend of all four to achieve MCS. The one standout is MA, that additionally utilizes a psychophysiologically-inspired, -plausible, and topological approach which is showing to be centrally vital to solving the singularity to AGI. Consequently, while pursuing MA the researcher may be able to achieve MCS. Nevertheless, the researcher is encouraged to consider all of these pathways that embody a potential way to achieving MCS.

At some point ultimate MA and likely MCS will include the machine possessing the attribute of consciousness, which is a result of deliberation with a representation that is distinct from the sensory data representation and is situated, simulated, and structurally coherent about the past, present, and future. Consciousness can be cognitively decoupled and contain a cohesive narrative to represent reality that is easy to understand, to reason about, make decisions on, and take appropriate actions as a result of deliberations over and blended with information based on agent experiences (Harbour, Rogers, Christensen, & Szathmary, 2015, 2019). According to Rogers (2019) the ultimate goal of a theory of consciousness is a straightforward set of fundamental laws, analogous to the fundamental laws of physics.

References

Brown, T. H., Zhao, Y., & Leung, V. (2009). Hebbian plasticity.

Clark, J., Mitchell, W., Vemuru, K., & Harbour, S., (2019), *in print*. Temporary memory neuron for the leaky integrate and fire neuron model. ISAP 2019, Dayton, Ohio

Dayan, P., Abbott, L. F., & Abbott, L. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems.

Gerstner, W., & Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press.

Gunning, D. (2018). Machine Common Sense Concept Paper. *arXiv preprint arXiv:1810.07528*.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223-241.

Friston, K., & Buzsáki, G. (2016). The functional anatomy of time: what and when in the brain. *Trends in cognitive sciences*, *20*(7), 500-511.

Friston K. (2018). Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?). *Frontiers in psychology*, *9*, 579. doi:10.3389/fpsyg.2018.00579

Harbour, S. D., & Christensen, J. C. (2015, May). A neuroergonomic quasi-experiment: Predictors of situation awareness. In Display Technologies and Applications for Defense, Security, and Avionics IX; and Head-and Helmet-Mounted Displays XX (Vol. 9470, p. 94700G). SPIE.

Harbour, S.D., Rogers, S.K., Christensen, J.C., & Szathmary, K.J. (2015, 2019). Theory: Solutions toward autonomy and the connection to situation awareness. Presentation at the 4th Annual Ohio UAS Conference. Convention Center, Dayton, Ohio. USAF.

Kidd, C., & Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. *Neuron*, *88*(3), 449-60.

Loewenstein G. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological Bulletin*. 1994; 116(1):75–98.

Rogers, S., (2019). QuEST – Cognitive Exoskeleton. Kabrisky Memorial Lecture 2019. For U.S. Government and DOD Contractors ONLY. USAF. WPAFB, Ohio

Sabour, S., Frosst, N., & Hinton, G.E. (2017) "Dynamic routing between capsules." *Advances in neural information processing systems*.

Sharpee, T. O., Calhoun, A. J., & Chalasani, S. H. (2014). Information theory of adaptation in neurons, behavior, and mood. *Current opinion in neurobiology*, 25, 47-53.

Widrow, B., Kim, Y., Park, D., & Perin, J. K. (2019). Nature's Learning Rule: The Hebbian-LMS Algorithm. In *Artificial Intelligence in the Age of Neural Networks and Brain Computing* (pp. 1-30). Academic Press.