Cognitive Banking Architecture - Human Centric AI framework for automated Customer Engagement in Banking

Sivaramakrishnan R Guruvayur
Bankbuddy
Dubai, UAE
shibu@bankbuddy.ai

***Abstract*—The purpose of this paper is to inculcate the concept of a Human centric in Cognitive banking framework in an attempt to deliver customer-centric AI-driven customer experience and engagement platforms. The modern-day banks have the compelling need to hyper personalize their customer experiences across all digital and physical channels. They need to leverage the power of data analytica and AI to bring back "personal" in banking, to create micro-moments based on dynamic customer journeys and context based streaming recommendations. Building an intelligent banking customer engagement model that can achieve human like intelligence and still a provide "personal Touch" is an uphill task due to the heterogeneous nature of banking systems, processes, data sources & more importantly, the prevailing regulatory requirements. Cognitive architectures have been designed to deal with the problem of complexity. But, when it comes to Banking systems, most of the existing cognitive architectures like SOAR, SiMA, LIDA, etc., have not solved the problem of complexity completely. The main objective of the proposed work is to develop a domain specific customizable Artificial Intelligence (AI) cognitive architecture to enable banks and financial institutions a human-centric approach to banking. The proposed framework employs a mix of AI faculties like domain and context specific Prebuilt Multi-lingual NLP chat bots (voice and text) across banking products, big data analytics, recommendation system to build dynamic customer persona & suggest next best action, customer journey hub, machine vision for biometric authentication & document ingestion, Task bots for with micro services to enable frictionless transactions , Channel manager to add conversational AI to bring ubiquitous presence across traditional (mobile banking, branch, contact center and new generation channels like WhatsApp, Facebook messenger etc. and an in-built middleware to connect to a variety of core systems ranging from core banking to CRM and contact centre solutions.Key performance indicators such =as Customer and Channel profitability, ease of customer navigations, customer engagement measures, and cross-sell abilities, Operational efficiencies can be derived out of metrics and compared with the current monolith banking systems.**

Keywords—Cognitive Banking architecture, Human Centric AI, Hyper Personalization, Recommender Systems, Ontology Engineering

I. Introduction

Financial institutions and the banking sector are currently being swept by a phase of transition and transpire the effective utilisation of personalisation across customer journeys. The demand for an AI based intelligent system that can provide personalized navigation, customized information content, recommendation for personalised products finally giving a ‘human face’ to the service has engrossed the banking sector[1]. Significant technological advancement in data science and computer vision is rapidly augmenting the pressure for banks to handle the explosion of data and find ways to mine actionable information in a jiffy will help them in customer advocacy and loyalty.. The recent demand for financial inclusion encompasses the challenges to cover differently abled customers [3], semi-literates and those with low level of financial literacy [4]. Hyper personalisation of the banking sector is expected by the customers [5] wherein rather than meeting the specific demands of the machine, the system should facilitate the machines to adopt human way of reasoning & decision making [6].The apparent solution to deal with such concerns involves the provision of user friendly platforms that enables everyone to communicate.

Inculcation of Artificial Intelligence (AI) can be seen in the banking sector with proposals involving Natural Language Processing (NLP) techniques to build intelligent banking assistants which not only respond to people’s queries and perform certain banking tasks, but also provide proactive customer service and customer delight [7]. Application of machine learning to understand and replicate human context leads to efficient machine learning tools and ways to enable learning computationally, facilitating the genesis of dynamic user focussed intelligent interfaces based on human behaviours [8]. These interfaces are infused with expressive gestural controllers with the understanding of particular feelings [9]. The employing of interactive machine learning for new real time interactive systems comprehends human actions sensed by various input devices with refined real time gestural control. These systems exhibits a natural “feel”, as they encode users’ embodied practices rather than just coding input/output mappings [9]. Security Pacific Bank's Corporate Cash Management group has developed an NLP interface for their Cash Management software system (SPACIFICS), Chase Manhattan Bank utilises cognitive products process wire traffic while The Carnegie Group has sourced a news categorization system for Reuters [10].Models are being proposed in which a particular financial customers can be proactively prompted to not only retrieve any type of information in real time but also guided in their financial journey and have a human like interaction with digital channels [11].

This paper introduces a Cognitive Banking Architecture that aims to enhance Hyper personalisation in banking orchestrated by customer centric AI approach by using AI enabled platforms adopting to Human Cognitive behaviour as discussed in the previous paragraph rather than humans adopting to machines & software intelligent agents.

The paper discusses the cognitive architectures available and their evident drawbacks in the second section. Section three describes the proposed architecture. Section four gives an explanation on the implementation and the working of the model. Section five present the Results and conclusion is given in the final section.

II. Cognitive Architecture

Cognitive model uses the principle that Homo sapiens share some important psychological invariants with computers. The mechanisms of human rationality are organized in the human brain to work in a coordinated manner. The architecture meant to achieve this co-ordination refers to characterization of the cognitive system at an abstract level without explaining the biology of neurons [12]. Cognitive architecture refers to the structure of the human brain in computational instantiation of such theories applied in Artificial Intelligence (AI) and computational cognitive science. It epitomizes the various results of cognitive ability in a comprehensive computer model [13]. The cognitive module is patterned to receive the specific sensory features and recognize a context based on the feature. On the basis of this, the module learns, creates, recalls a set of actions and evaluates the action plans over any previous action plans in a related context [14].

Over the last two decades, cognitive architectures are being proposed and infused in almost all sectors. Work by [15] proposed an architecture for general intelligence called SOAR that provides appropriation for intelligent actions. The decision making process of cognitive architecture SiMA (Simulation of Mental Apparatus & Applications), previously known as ARS [16], which is entrusted on a functional holistic model of the human mind based on psychoanalysis has been explained by [17] and [18] discusses a cognitive architecture (LIDA) that exhibits attention, action and human-like learning in controlling cognitive agents that depict human experiments in performing real-world tasks. Study by [19] presents a comprehensive review of the implemented cognitive architectures in the form of a table. Work by [20] gives past, present and future scope of grounded cognition and predicts that cognitive science will increasingly witness the integration of classic symbolic architectures, statistical systems, and grounded cognition. Research by [21] provides a comparison of important cognitive architectures. Study done by [22] review 40 years of development of cognitive architectures. Research by [23] introduces the Cognitive Systems Toolkit (CST) that provides flexibility in culling specific techniques and algorithms for setting up the CST architecture usually in robotic application. They proposed a Multipurpose Enhanced Cognitive Architecture (MECA), a hybridization of SOAR based on Dual Process Theory, Dynamic assumption, Conceptual Spaces and Grounded Cognition, and is constructed using CST. There are various issues which have to be considered while infusing Cognitive architectures in banking. Intelligent solutions are required to deal with users asking repeated questions and to provide appropriate feedbacks to customers in cases where user is not responding for a long time or when continuously asking questions [7]. Banking sector is constrained with static digital interfaces with non-existent personalisation [24] and no domain specific machine reasoning and recommendation systems. However, sectors which provide customised service rather than an auto reply to customer queries is the need of the hour [25].

Cognitive architectures like SOAR, LIDA and SiMA fail to provide personalised environment to users and lacks domain specific intelligence. SOAR architecture does not support decision and conflict resolution that has slight variation with the architectural decision process. The high level learning styles must be included within its basic learning paradigm such as to parse an input it requires predefined declaration of the type or channel of the message [26]. In LIDA framework, data exchange happens through codelets that process data in one memory and put it in other memory. A drawback thus seen is that it only allows loop-based systems rather event-based systems and hence lacks control over new occurrences [27]. Further SiMA is a theoretical model and not a technical application [28].

Thus in an attempt to meet the business challenges involved in staying hyper customer centric and in their need for ubiquitous access to financial services across legacy &newly emerging access channels, customized services and proactive and personalized service delivery, a cognitive banking architecture has been proposed which is inspired from the Cognitive architecture of the human brain specifically the Semantic Pointer Architecture Unified Network (Spaun)

A. Spaun

The Spaun has been spearheaded by Chris Eliasmith of the University of Waterloo Centre for Theoretical Neuroscience. It is a spiking neuron model of 2.5 million simulated neurons organized into subsystems and is focussed to bridging the understanding of brain behaviour gap. The model is composed of three hierarchies, an action selection mechanism, and five subsystems. Components of the architecture communicate through spiking neurons engaged in implementing neural representations called semantic pointers, by various firing patterns [29].The visual system is the first hierarchy, which compresses the input image into an abstract representation of that input formulating a semantic architecture because the abstract representation maintains similarity with input image. It manifests itself as a pointer in the sense that the original information can be recovered from the compressed form. It also consists of a motor hierarchy that dereferences an output semantic pointer by decompressing firing patterns to drive a two degree-of-freedom arm [30]. The third internal hierarchy is a working memory that can bind and unbind semantic pointers, arranging for the compositionality required for complex cognition. It portrays intermediate task states, task sub goals, and context. Anatomically they represent prefrontal and parietal cortex. The five subsystems are engaged in mapping the visual hierarchy to a conceptual representation (information encoding), extracting associations among input units (transform calculation), evaluating the reward affiliated to the input (reward evaluation), mapping output to a motor semantic pointer (information decoding) and controlling motor timing (motor processing). These subsystems and hierarchies utilise multiple components to perform the specific functions [31]. The drawback seen in this model is that the tasks of Spaun are all pre specified. The model cannot be briefed to achieve a new task or master one on its own [32].

III. Proposed Architecture

The proposed architecture presents a cognitive banking platform that is compatible to multiple input channels (bots, Web, Touch pads, Mobile,, Social media, Kiosk banking, Voice interfaces), Customer Journey automation and so on. Central to the theme are the creation & availability of Banking Ontology engineered by ML algorithms, a multi-step reasoning engine which revolves around and respond to multisensory inputs from humans, AI driven Recommender Systems , machine vision and NLP to enable intelligent interactions in assisting banks and financial institutions with customer experience management, hyper-personalization, Cross selling, Risk & Compliance and operations automation. The four hierarchies in the model are the input modality, output

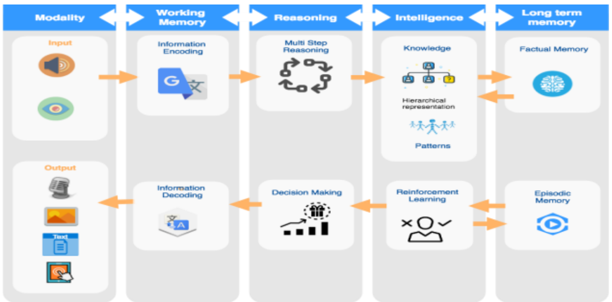

Fig 1: Cognitive Banking Architecture

modality, the working memory and long term memory. The action selection falls into the broader domain of reasoning and is supported by intelligence (ontology) and long term memory. The proposed model can be understood through Figure1.

A. Input Modality

The input modality of the proposed architecture is remarkably ubiquitous with the capability accepting inputs in any format (text, voice, touch, image) over a AI platform compatible with various communication channel managers (WhatsApp, Facebook, Google Assistant, Amazon Alexa, Mobile apps, Web chat, Twitter, Slack, etc.) much in the same way a human brain takes messages from the outside world through the sensory organs, process these inputs to normalize them to "make sense"& 'Process ready". It is a dynamic platform with immediate response functionalities making communication interactive, accurate and appeals to the human sensory inputs. This is accomplished by Customer journeys mapping (various touch points processes between customer & Bank's operational processes both Digital-in & Digital-Out) which are part of Banking Ontology discussed in Long term Memory section. It encompasses prebuilt Multi-lingual NLP chat & voice bots leveraging Deep Learning to handle slang, abbreviations, spelling errors, phonetics, dialects etc. with wide functional coverage across financial services products. It incorporates dynamic and nested AI based customer journeys e.g. origination flow with embedded computer vision to read data from uploaded documents like Passport, National Id, etc., ensuring face recognition for authentication, liveliness detection, emotion, gender, age analysis. This also helps feeding the context to the AI engine and to provide near perfect answer or recommendation than giving a generic preconfigured response.

Working Memory

The working memory is the context engine which builds context to the customer interaction using the on-going conversation based on user actions, emotion and continuous interaction analysis. The input received is successfully accepted by normalizing any form of data received into a machine processable format. The in-built middleware connects to a variety of core systems ranging from Core Banking Systems, Customer relationship management (CRM), CIF, Risk & Compliance systems, Contact centre solutions and so on.The system analyses the input to provide the best response/action possible using a multi-step reasoning paradigm.

The most appropriate action is determined by interaction between information retrieved from Big data analysis and customer persona generation and the nearest analytically close neighbour. Multiple iterations are performed to arrive at the best possible answer to the customer question when a direct answer is not available using Decision Tree techniques supported by Ontology. The system can further ask questions using Natural Language Generation to find out which product/service the customer is interested in and when the customer makes a selection, pulls the appropriate answer from memory through multiple iterations & past history/patterns. This will also have cognizance to keep an eye on the on-going interactions, understand and build journey maps which will be used to measure current interaction satisfaction.

B. Long Term Memory

The action hierarchy is bolstered by the use of Banking Ontology and in-memory analytics provided by ML algorithms. Artificial Intelligence and Big Data based Product recommendation engine executes data mining to recommend products/service//actions to customers based on persona, chat context and past history. The semantic memory or the long term memory manages general information and connects them to the present issue whenever in need. The episodic memory is associated to specific personal events and is usually active in Recommendation systems. Knowledge graphs helps in retaining context to ensure human like conversations. The system is made capable to even understand partial intents by inculcating the technology of utilising reverse contexts. Task automation bots for functions like ATM pin reset, lost card reporting, stop cheque, cheque book request, statement request, account balance, transfers and so on are handled by these ML models. Using Reinforcement learning, the underlying model is expected to make some sequence of decisions in an uncertain and potentially complex customer live environment. The ML models employ Supervised and Reinforcement learning to come up with the most appropriate solution to the problem. The challenge lies in accomplishing a learning environment that can also serve as reference in future, in other words, updating the underlying Banking Ontology. It identifies and learns from human sensory inputs & outputs such as sentiment handling and escalation matrix. It is this engine which decides at what point in time to introduce a Human agent into the process. Based on the sentiment analysis, the ML engine would decide whether the human should be included at the Brink-of-the-process (if needed) or at the escalation touch points.

C. Output Modality

Finally the most appropriate decision is arrived at and the motor processing output provides the “all channels” compatible best response. The output is compatible to the input channel through which the input was received. In addition to that the output is made compatible to all other channels so that it can be received by the customer through any channel at anytime from anywhere.

IV. IMPLEMENTATION

The implementation of the cognitive banking framework can be studied under the following heads.

A. Customer Journey Mapping

It indicates the journey from the customer approaching the framework, to understanding of what the customer intends fromthe service provider and providing the best response. They are designed from perspective of the customer intent rather than from the traditional perspective of bank’s operational processes. For example, if a customer wishes to open a Fixed Deposit account and the platform realizes that the customer just turned senior citizen and is yet to provide self-declaration to qualify for preferential rate, instead of terminating the session, or offering standard rate, the platform opens up another nested customer journey to complete the required documentation process. Following which it goes ahead and offers fixed deposit at the special rate to the customer. Such Hyper Personalization design makes the customer journey frictionless from human perspective when interacting with an AI enabled customer engagement platform. These nested journeys sit as a wrapper layer above the above bank's internal processes coupled with the knowledge representation provided by Banking domain Ontology which makes the whole personal banking really "personal", creates micro-moments based on dynamic customer journeys and streams appropriate recommendations.

B. Data Collation

The framework deals with both structured and unstructured data. Structured data remains pre specified in the architecture. However to achieve a human perspective, arises the requirement of certain data which has not been structured and must be collected from various sources available. This framework understands the customer intent and provides response and service by referring to both these structured and unstructured data.

C. AI Model Implementation

For the implementation of AI within the framework it is essential to perform customer and product attribute mapping. Knowing the attributes of the customer and product becomes pre requisite to understand customer intent as well as to recommend the appropriate product. It gives an idea of how a customer’s query can be linked to the output desired. Apart from these, pattern detection comes into account. This enables the framework to link customer to its closest neighbour and respond or recommend on the basis of the group the customer has been mapped into. The past customer behaviour and interaction patterns triggers relevant nested journeys for the recommender system. For example, if the customer’s attributes include organic food and indicate an inclination towards sustainable living, then he can be patterned into group of customers who might be looking for an electronic vehicle loan. Further, the framework is tuned to learn from experiences and keep expanding its boundaries.

Humanizing Digital User Experience (UX)

Instead of a rigid format which most of the current customer engagement platforms uses, it understands the way human interacts with machines. Human often use incomplete sentences, questions, slangs, abbreviations, mix of languages in a sentence with multi lingual format. This framework provides built-in mechanism wherein if someone asks an incomplete question, the framework can complete the query and give response making the interactions more active & accurate based on the human characteristics. This framework facilitates the machines to adopt human way of reasoning & decision making. For example, if a customer queries about interest rate but don't specifically mention a product such as Deposit, Home/Car/Education/Personal loan etc., the framework fetch out all the products associated with "Interest rates" and present it to the users to choose from the product list much in the same way as a human's cognitive reasoning framework would function. It provides an active and dynamic interactive environment, understands the customer intent, gives an appropriate output on the basis of its attributes and recommends on the direction of human perspective thus creating a humanized experience.

This frame work can also help in differently abled people to interact with the banking systems confidently using unified multi-sensory inputs and outputs.

V. RESULTS

Time and Motion study was conducted to identify the whole process gap on how much time the physical channel took in handling customer. The implementation findings were collected from a large bank in Africa and a Tier 1 Insurance company in Middle East. The findings reveal results like On-boarding process took 28-30 minutes whereas the pilot platform using this framework 3-4 minutes. The results have been included in Table I. In addition to that significant increase in lead conversion rate was observed.

TABLE I. IMPROVEMENT IN SPECIFIC AREAS AGAINST THE LEGACY

Areas	*Improvement (%)*
Call deflection for general information	9.2%
Revenue upliftment via cross sell and up-sell	5.7%
Churn prevention for liability products	4.8%
Reduction on boarding dropouts	20 – 30%
Reduction in customer acquisition cost	30- 40%
Instant resolution	27%
Retention of customers	10-20%

VI. CONCLUSION

The proposed model is found to be of considerable significance in its broad domain of service offered in the advantage of the banking and financial sector. Unlike existing frameworks, the proposed model has a laser sharp focus on customer experience in automating financial services using Deep learning, NLP, computer vision, big data, recommendation system & speech analytics fused with BFSI domain knowledge. It has the ability to transform the traditional service paradigm to an effective and interactive platform with the inculcation of Embedded Financial services that has experienced the metamorphosis from legacy to latest channels and from transactions to goal based journeys with apt emphasis on personalized customer experience. The model has achieved the status of a proactive advisor with its infused AI based advisory from a reactive service provide. The BFSI sector appears to benefit significantly from the Open banking ecosystem offered by the model in comparison to the silos of unstructured information of the traditional approach. This framework works on the ground of a human-centric perspective and includes workflows inspired by human working practices. This coadaptation of human and the system establishes a human centric environment for the customers enabling customization and hyper personalization in the proceedings of bank related services.

REFERENCES

[1] Murugesan, S., & Ramanathan, A. (2001, December). Web personalisation-An overview. In *International Computer Science Conference on Active Media Technology* (pp. 65-76). Springer, Berlin, Heidelberg.

[2] Angelakopoulos, G., & Mihiotis, A. (2011). E-banking: challenges and opportunities in the Greek banking sector. *Electronic Commerce Research*, *11*(3), 297-319.

[3] Muhammad, A., Gatawa, N. M., & Kebbi, H. S. B. (2013). Impact of information and communication technology on bank performance: A study of selected commercial banks in Nigeria (2001–2011). *European Scientific Journal*, *9*(7).

[4] Menon, A. D. (2018). Future of Banking. TallyXinc

[5] Goyal, M. (2019, April). Artificial intelligence: a tool for hyper personalization.International Journal of 360 Management Review, Vol. 07, Issue 01.

[6] Shah, K. B., Shetty, M. S., Shah, D. P., & Pamnani, R. (2017). Approaches towards building a banking assistant. *International Journal of Computer Applications*, *166*(11).

[7] Gillies, M., Fiebrink, R., Tanaka, A., Garcia, J., Bevilacqua, F., Heloir, A., ... & d'Alessandro, N. (2016, May). Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3558-3565). ACM.

[8] Rebecca Fiebrink. 2011. Real-time human interaction with supervised learning algorithms for music composition and performance. Ph.D. Dissertation. Princeton University, Princeton, NJ, USA.

[9] Harris, M. D. (1992). Natural language in banking. *Intelligent Systems in Accounting, Finance and Management*, *1*(1), 65-73.

[10] Warnapura, A. K., Rajapaksha, D. S., Ranawaka, H. P., Fernando, P. S. S. J., Kasthuriarachchi, K. T. S., & Wijendra, D. (2014). Automated Customer Care Service System for Finance Companies. In *Research and Publication of Sri Lanka Institute of Information Technology (SLIIT)'* (p. 08). NCTM.

[11] Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1-20.

[12] Lieto, A., Bhatt, M., Oltramari, A., & Vernon, D. (2018). The role of cognitive architectures in general artificial intelligence.

[13] Srinivasa, N., & Khosia, D. (2008). U.S. Patent Application No. 11/801,377.

[14] Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence*, *33*(1), 1-64.

[15] S. Schaat, K. Doblhammer, A. Wendt, F. Gelbard, L. Herret, 0 Bruckner. A Psychoanalytically-Inspired Motivational and Emotional System for Autonomous Agents, in proceedings of IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society, 2013, pp. 6647-6651.

[16] Wendt, A., Gelbard, F., Fittner, M., Schaat, S., Jakubec, M., Brandstätter, C., & Kollmann, S. (2015, November). Decision-making in the cognitive architecture SiMA. In *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)* (pp. 330-335). IEEE.

[17] Franklin, S., Madl, T., D'mello, S., & Snaider, J. (2013). LIDA: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, *6*(1), 19-41.

[18] Samsonovich, A. V. (2010). Toward a unified catalog of implemented cognitive architectures. *BICA*, *221*(2010), 195-244.

[19] Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in cognitive science*, *2*(4), 716-724.

[20] Danilo Fernando Lucentini and Ricardo Ribeiro Gudwin. A comparison among cognitive architectures: A theo-retical analysis. Procedia Computer Science, 71:56–61, 2015.

[21] Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. *arXiv preprint arXiv:1610.08602*, 1-74.

[22] Paraense, A. L., Raizer, K., de Paula, S. M., Rohmer, E., & Gudwin, R. R. (2016). The cognitive systems toolkit and the CST reference cognitive architecture. *Biologically Inspired Cognitive Architectures*, *17*, 32-48.

[23] Loonam, M., & O'loughlin, D. (2008). Exploring e-service quality: a study of Irish online banking. *Marketing Intelligence & Planning*, *26*(7), 759-780.

[24] Madu, C. N., & Madu, A. A. (2002). Dimensions of e-quality. *International Journal of Quality & reliability management*, *19*(3), 246-258.

[25] Wray, R. E., & Jones, R. M. (2006). Considering Soar as an agent architecture. *Cognition and multi-agent interaction: From cognitive modeling to social simulation*, *33*, 53-78.

[26] Wendt, A., Wilker, S., Meisel, M., & Sauter, T. (2018, June). A multi-agent-based middleware for the development of complex architectures. In *2018 IEEE 27th International Symposium on Industrial Electronics (ISIE)* (pp. 723-728). IEEE.

[27] Zucker, G., Wendt, A., Siafara, L., & Schaat, S. (2016, October). A cognitive architecture for building automation. In *IECON 2016-*

42nd Annual Conference of the IEEE Industrial Electronics Society (pp. 6919-6924). IEEE.

[28] Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *science*, *338*(6111), 1202-1205.

[29] DeWolf, T. (2010). NOCH: A framework for biologically plausible models of neural motor control (Master's thesis, University of Waterloo).

[30] Stewart, T., Choo, F. X., & Eliasmith, C. (2012). Spaun: A perception-cognition-action model using spiking neurons. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, No. 34).

[31] Rasmussen, D., & Eliasmith, C. (2013). Modeling brain function: current developments and future prospects. *JAMA neurology*, *70*(10), 1325-1329.