Summary of Talks related to Consciousness, Metacognition and Representation Similarity Analysis

COGSCI 595 Cognitive Science Colloquium Final Paper, Fall 2022
Student: Sushma Anand Akoju, Professor: Prof. Mary Peterson

1 Context and motivation

As a first year PhD student, I am on a quest for how human brain learns to reason, consciously self-verifies. By learning such aspects, I would like to understand not only human brain but also the human brain-inspired Artificial Neural Networks. The line of reasoning, explainability of Deep Learning models has been debated while they are known to be "Black Box systems" for the lack of explainability, and thus lack of formal validation and verification for lines of reasoning. More specifically I am interested to learn what specific neuroscience evidences exists to provide that a reasoning is taking place, that we can express a line of reasoning and seek to verify/validate such line of reasoning. As a next natural step I am interested to extend my learnings towards my main research interest i.e. to find out how Artificial Neural Networks and present day Deep Learning models function. Then I want to know which of the facts or rephrased facts that are available in working memory and why certain wirings and factuals triggered the answer. However we do have information about inner workings of the Artificial Neural Networks and present day Deep Learning models. Yet we are confounded with a lack of explainability and verifiability of these complex, blackbox systems.

To explain this with an example, let us consider a small example for Natural Language Inference: for a given premise, i.e. an evidence, we can have a hypothesis. Whether such a hypothesis entails or not can be defined as a simplest form of Natural Language Inference. Premise: A football game is being played by males. Hypothesis: Some men are playing a sport.
Here hypothesis is an entailment of premise because hypothesis is a generalization of premise i.e. males are a subset of some men and football game is a subset of sport. We naturally recognize such subset from natural language sentences/utterances in day to day life. There are several ways to describe and explain a line of reasoning, such as by using Set Theory and containment, First Order Logic, using Meaning/Semantic representations, explanations and a combination of some or most of these methods. Despite such availability of many different approaches, it gets challenging to explain and extract a concrete line of reasoning from Deep Learning models. Since the Deep Learning models and Artificial Neural Networks are inspired from Human brain functioning, some questions :

1. What evidences exist that such line of reasoning occurs in neuronal mechanisms in Cognitive Science or Neuroscience research?
2. How to find clues towards reasoning and teach how to reason?
3. What computational neuroscience methods exist that can help towards reasoning?

From the talks during Cognitive Science COGSCI 595 Colloquium course, I wanted to learn and understand what some hints and clues towards reasoning. I have included summaries for 3 speakers' talks and the respective papers that I read through before the each of the talks and consolidate my learning in this report. The analysis and my learning from the 3 speakers' talks are as follows:

1. Prof. Liad Mudrik: Taking a neuroscientific-philosophical approach in studying free will and consciousness
2. Dr. Janet Metcalfe : A Metacognitive model of curiosity
3. Prof. Afra Alishahi: Getting closer to reality: Grounding and interaction in models of human language acquisition

Keywords: reasoning; higher order logic; deep learning; neural networks; cognitive science; neuroscience; formal verification and validation; consciousness; metacognition

2 Prof. Liad Mudrik: Taking a neuroscientific-philosophical approach in studying free will and consciousness

As per the talk from Prof. Liad Mudrik, the talk is about neuroscientific-philosophical approach in studying free will and consciousness. There seems to be various debates which are broadly classified into two categories that are: compatibilist and incompabilist groups. Each have been vastly debated with concrete criticism from empirical findings. This talk was based on (Mudrik et al., 2022) i.e. Free will without consciousness. I have summarized the talk and 2 papers of the speaker Prof. Liad Mudrik.

2.1 Free will without consciousness?(Mudrik et al., 2022)

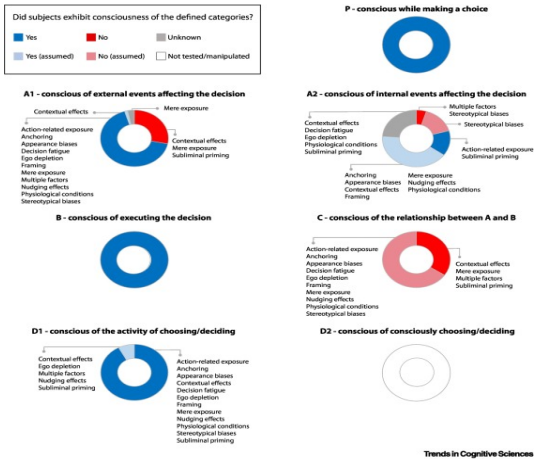
The paper (Mudrik et al., 2022) can be discussed based on 4 types of Cognitive architectures that include the theories of consciousness. They are as follows:

1. Global Neuron Workspace: This theory suggests that while the unconscious processes are encapsulated, they are not completely unavailable but are only limited to specified modules. Once the signal is strong enough, information is amplified by the attention. If this crosses a threshold, the state is said to enter Global Neuron workspace, which then allows interactions and integrations.
2. Integrated Information Theory: This theory is based on phenomenological axioms: existence : loosely based on Descartes "I exist therefore I am", composition: each experience is a various combination of multiple aspects, information : each experience differs from other possible experiences, integration: each experience is irreducible to non-interdependent components (red triangle versus triangle) and exclusion: each experience has boundaries that excludes all others i.e. certain things can be experienced but others cannot be. An interesting aspect that seems to intersect with higher-order thinking or consciousness is: the higher the number of different concepts and their Phi Max value, the higher the conceptual information. Existence is considered that of zeroth-order which seems closer to facts, we could say boolean concepts candidates set with some operators, can be a boolean concept. Compositionality seems to represent the higher order mechanisms. For example, let A, B and C sets hold some logical relations which is referred as first order mechanism. The higher order mechanisms seems to suggest some kind of inferential or formal logic based concepts. Information here seems to require selectivity. This was suggestive of causal nature of information which selective and constrained over past states. I think that this discussion of information had a great intersection to Information Theory.
3. Higher Order Thought: This theory, as it seems to be intuitive, is inspired from philosophy. David Rosenthal suggests that in order to say "I have consciousness, it is not enough to say I have activation in sensory areas, I should have had some higher order state to distinguish the state when such consciousness existed versus unconsciousness which would not have led to the awareness that came from consciousness.
4. Recurrent processing: is a theory that suggests higher order thought or global broadcast is not required. It is simply from local recurrency where feedback separates between two axes between consciousness and attention.

At the first look out, one could easily believe and be confused by the underlying philosophical intersection from popular philosophers from whom they were inspired. So it maybe easier to find out how each of the theories of consciousness chosen here for this talk would evaluate and support the theoretical constructs with that of empirical results. The talk extends discussion on findings from (Mudrik et al., 2022). The challenge defined as per the talk comes from the questions:

1. How to translate philosophical construct to psychological construct?
2. How to translate or formulate empirical definition from psychological constructs?
3. How to relate empirical findings with that of philosophical constructs?

Thus Free will seems to suggest a theoretical construct consisting of: lack of external constraints, indeterminism, emergence, consciousness, higher order thought and reasons response. While the paper (Mudrik et al., 2022) describes an example about last minute change in decision on an Election day, when the voter changes his/her decision after entering the voting booth. The scenario discussed was about how irrelevant factors are known to influence a voter's choice such as candidate's face width to height ratio or the location of their names of the ballet which were known to have shown a deliberate decisions. The question asked in this paper(also as discussed during the talk) is that when one would not have imagined such irrelevant factors influencing the voting decision would effect such an important decision. Neverthless, even before verifying the some information or hypothesis that a voter's decision is influenced by such irrelevant factors, the mere awareness of such a theory about influence can significantly constrain the thought process.



From the impact of consciousness on decision making, the paper discusses an interesting aspect about how lack of consciousness could threaten free will. Free will implies certain amount of control i.e. being conscious of motivations, reasons, decisions and actions. While various types of consciousness might be responsible for reason responsiveness. For example, reason here could be a consideration of for/against an action rather than simply considering why a person acted in a certain way. This aspect also considers if a person has ability to recognize and react to reasons for/against then it seems consciousness exists which indicates free will. If a person is unable to recognize or react to reasons which is suggestive of compulsions or delusions, then it can render an action unfree. The experiments conducted for consciouness generally included a default threshold of 200 with an upper bound of 210 which detects the presence of "consciousness" responses during each of the experiments conducted towards assessment of consciousness.

To some extent, on the hind sight to this talk and further reflecting on the consciousness theories, we could refer another talk by Dr. Peng Peng which suggests that Working Memory includes the attention. First (JENNINGS, 2015) suggests a skilled behaviour seems to suggest consciousness could exist without attention by means of consciousness entrainment, which seems to be an interesting proposition. Dr Peng's work on Working Memory suggests that attention is part of Working Memory (Peng et al., n.d.). But in case of an example such as solving a Homework, we tend to know which questions we easily know answers from recollection without Information retrieval since factuals are available to solve those questions. But for questions for which we may need to put the pen down to solve, we need significant amount of Information retrieval, consciousness, Working memory and attention to solve it.

Another approach that provides a very concrete discussion in detail is (Tsuchiya & Koch, 2014). This work suggests "...attentional amplification of the neural representation of an event or object always necessary to experience it." and "...attentional amplification is necessary to experience an object only when it needs to be "selected". In a situation without any competition (e.g.,an isolated object or a uniform texture), selective attention may not play any significant role. Accordingly, we argue that the neuronal mechanisms that give rise to consciousness need to be carefully disentangled from the neuronal mechanisms that resolve competition....". This discussion seemed very profound, although does not conclude but separates the neuronal mechanisms that rise consciousness do not imply neuronal mechanisms that resolve competition and thus consciousness can be tested for top-down attentional amplification or not. (Tsuchiya & Koch, 2014) argues that consciousness and attention covary. Thus there seems to be more scope of learning about consciousness.

2.2 The ConTraSt database for analysing and comparing empirical studies of consciousness theories (Yaron, Melloni, Pitts, & Mudrik, 2022)

The second paper I studied was the prior to (Mudrik et al., 2022) is "The ConTraSt database for analysing and comparing empirical studies of consciousness theories" (Yaron et al., 2022). This paper verifies about the various theories on consciousness that exist and evaluating the methods applied for theoretical verification. This work collects articles from various theories, suggests that some experiments were conducted posthoc while most others were explicitly testing theory predictions. Theory driven experiments challenged the theories most frequently than expected. The post-hoc experiments, the research question and hypothesis reported in the introduction did not related to any prediction by the theories. Thus (Mudrik et al., 2022) and (Yaron et al., 2022) paper's discussions suggests why defining the challenge in terms of questions, as discussed above, is helpful to see how difficult it is to verify the theories of consciousness in Cognitive Sciences and Neuroscience.

3 Dr. Janet Metcalfe : A Metacognitive model of curiosity

As per the talk by Prof. Janet Metcalfe, the experiments were conducted towards curiosity as a metacognitive model. I have recently started reading Prof. Metcalfe's books on this subject written by Prof. Metcalfe i.e. Metacognition (Dunlosky & Metcalfe, 2008) and another book on Metacognition: knowing about knowing (Metcalfe, Shimamura, et al., 1994).

3.1 Curiosity: The Effects of Feedback and Confidence on the Desire to Know (Metcalfe, Vuorre, Towner, & Eich, 2022)

The talk was primarily based on Prof. Metcalfe's work (Metcalfe et al., 2022). It is proposed that one kind of curiosity — which we call Curiosity 1. This can be understood in terms of the Region of Proximal Learning (RPL) framework which is a Metacognitive framework that underpins motivation to learn. This framework proposes that people feel most curious when they feel they are on the verge of knowing or understanding. The processes, conditions, and outcomes specified by the RPL view of curiosity were reviewed along with several lines of relevant evidence including

1. differences in the conditions under which experts and novices seek information or, alternatively from mind wandering approach
2. experiments investigating people's choices of whether to study materials for which they have high versus low feelings of knowing
3. results related to people's engagement with corrections to errors made with high confidence
4. curiosity, attention, EEG, and learning data related to the tip-of-the-tongue state

In addition to Curiosity 1, however, we also propose that there is a second kind of curiosity, which we call Curiosity 2. As was discussed, Curiosity 2 is based on exploration as well as from opposing cognitive/motivational principles from those in evidence for Curiosity 1. The conflation of Curiosity 1 and Curiosity 2 has resulted in considerable confusion in the literature.

Curiosity and its effects on people's confidence was discussed in the (Metcalfe et al., 2022). It was noted in this work, curiosity increased when there was a high-confidence error. As it was noted to be a serendipitous finding that when no feedback was given, people were more curious about high-confidence errors than they were about equally high-confidence correct answers.

My learnings from this talk as well as Lunch time discussion with Prof. Metcalfe:

1. How to gain clarity for experimental design and example demonstration ? Example of this: how the example about Tip of the Tongue questions was chosen towards this experiment was very interesting. The motivation to include such an example among many examples suggests the clarity and simplicity of the problem.
2. How to formulate research question closely towards experiment idea description? The research question the authors chose to discuss about seemed to have contributed towards hypothesis test and experimental design. Although this seems like an extension of previous work by Prof. Metcalfe i.e. Region of Proximal Learning model. The experiment 1 statement: "Effects of Yes/No Feedback on Curiosity About Errors." suggests the hypothesis studies. To validate the effect size and confidence intervals and significance intervals, they used Bayesian Regression model. This choice of model seems reasonable since Priors are a domain-specific choice. It seemed that authors wanted to find out significance from a posterior distribution of model parameters. The Bayesian Regression model used is an extension of uncertainty prediction for domain-specific problem. Thus this makes for a case of how answers can impact with or without feedback (Bürkner, 2018).
3. What is the motivation to research on this research question? The why, what and how. The experiment 2's (Metcalfe et al., 2022) design seems to suggest that the research question was very clear just from description: "Experiment 2 investigated curiosity for both correct and incorrect answers when yes/no feedback was provided following people's answers, as well as when no feedback was given. The specific design of Experiment 1: curiosity following yes/no feedback as a function of confidence in errors. This suggests that modelling such functions when no feedback was given, people were more curious about high-confidence errors than they were about equally high-confidence correct answers.

During my lunch discussion with Prof. Metcalfe, I have discussed about my research topic and essentially a problem that I was keen to solve. Prof. Metcalfe suggested I and rest of the participants from lunch group participate and email her. The document which has full summary of the experiments we are keen to conduct once Prof. Metcalfe likes to start early January 2023 is: https://tinyurl.com/252xnryu.

3.2 Epistemic curiosity and the region of proximal learning (Metcalfe, Schwartz, & Eich, 2020)

The second paper (Metcalfe et al., 2020) is about the designing Region of Proximal Learning framework from epitesmic curiosity as a metacognitive feeling state that is related to individual's Region of Proximal state, which is an adaptive mental space which represents the feeling when we are on the verge of knowing or understanding. The underlying motivation for this is the curiosity which is Explore versus Exploit. First the paper discusses prior works on curiosity by first clarifying what is not epitesmic curiosity. The rat searching for food is driven hunger and thus is a perceptual curiosity.

4 Prof. Afra Alishahi: Getting closer to reality: Grounding and interaction in models of human language acquisition

The talk was very helpful. I was specifically more interested towards comparing neural representations between image and language embeddings in deep learning models that are trained for multi-modal tasks.

4.1 Correlating neural and symbolic representations of language (Chrupała & Alishahi, 2019)

This paper attempts to show that by using Representational similarity Analysis and by comparing Tree kernels which are essentially the substructures that are present in neural activations. The paper compares the substructures between true symbolic structure (which is what we know as parts of speech constituency tree) and the neural representations of Language model. The paper systematically compares first over arithmetic expressions with clearly defined syntax and semantics. And they extend the experiments and evaluation to correlate neural representations of English sentences with their constituency parse trees. The paper considers semantic value, tree depth and tree kernel. Such analysis and evaluation methodology can be a key to understanding the black-box "Deep Learning models", which may hold clues towards formal verification and validation of Deep Learning models.

Acknowledgements

This course was supervised by Prof. Mary Peterson and advised by my advisor Prof. Mihai Surdeanu. I am grateful to the valuable discussions with Prof. Liad Mudrik, Prof. Janet Metcalfe and Prof. Afra Alishahi and continued email conversations. I am very grateful for the guidance from Prof. Mary Peterson, Kirsten Leigh Cloutier Grabo and my classmates for the discussions and opportunities to learn.

References

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. Retrieved from https://doi.org/10.32614/RJ-2018-017 DOI: 10.32614/RJ-2018-017

Chrupała, G., & Alishahi, A. (2019). Correlating neural and symbolic representations of language. *arXiv preprint arXiv:1905.06401*.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

JENNINGS, C. D. (2015). Consciousness without attention. *Journal of the American Philosophical Association*, *1*(2), 276–295. DOI: 10.1017/apa.2014.14

Metcalfe, J., Schwartz, B. L., & Eich, T. S. (2020). Epistemic curiosity and the region of proximal learning. *Current Opinion in Behavioral Sciences*, *35*, 40–47.

Metcalfe, J., Shimamura, A. P., et al. (1994). *Metacognition: Knowing about knowing*. MIT press.

Metcalfe, J., Vuorre, M., Towner, E., & Eich, T. S. (2022). Curiosity: The effects of feedback and confidence on the desire to know. *Journal of Experimental Psychology: General*.

Mudrik, L., Arie, I. G., Amir, Y., Shir, Y., Hieronymi, P., Maoz, U., ... Roskies, A. (2022). Free will without consciousness? *Trends in Cognitive Sciences*, *26*(7), 555-566. Retrieved from https://www.sciencedirect.com/science/article/pii/S1364661322000638 DOI: https://doi.org/10.1016/j.tics.2022.03.005

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., ... Tao, S. (n.d.). A meta-analysis on the relation between reading and working memory.

Tsuchiya, N., & Koch, C. (2014). *On the relationship between consciousness and attention*. MIT Press.

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022, 04). The contrast database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour*, *6*, 1-12. DOI: 10.1038/s41562-021-01284-5