**Session 3. Report on the discussion session.**

Carlos Gutierrez, OIST

After the introduction of the discussion by Keiji Tanaka, Matt Botvinick highlighted the importance of embedding hierarchy for certain mechanisms. In one of his papers, a simple implementation of hierarchy on the architecture leaded to the emergence of functional hierarchy. That factor can be important in meta-learning.

Xiao-Jing Wang, mentioned 2 important points. The 1st one, while conventional AI implements conventional learning by changing connection weights, an interesting idea is to realize adaptive processes without connection weight updates, where the system still behaving from trial to trial, by connecting neuronal circuits in meaningful weights. The 2nd point is, how to realize gating mechanisms, which still unknown in neurosciences, however, they are starting to have some insights from simple problems by 2 ways: a) trajectories and looking the dynamics, b) changing connection weights in subspaces, then looking the vector field changes. The last one is an interesting angle that would need to be tested experimentally.

About ideas on the interaction of dopamine and the pre-frontal cortex (PFC), Angela Langdon highlighted the question: what constitutes a reward? Perhaps ultimate goals and subgoals? however, looking at the brain, what is considered a reward? This is easy to define in artificial systems but at the brain? Meta-learning involves generalization and transfer from system to system, thus generalization is not what we think as a “reward”. On the other hand, dopamine is though as a reward for goal fulfillment. About interactions between PFC and dopamine, PFC is more related to setting the goal, whereas dopamine is related to goal fulfillment, they have different windows of time, this relation seems complex in biological systems.

Ryota Kanai raised a question about what people think of RL and it functions, considering that goal-directness seems not related to intelligence. How to distinguish between reflecting systems and goal directed systems?

Angela mentioned devaluation in animas, goal changing animal behaviors. The motivational stage in an evolving rodent when they chase water, dictates if they consider that as a goal.

Xiao-Jing mentioned the need of mapping to discrete localized structures. Is it computed locally, or a decision process is taking place within a hierarchy?

Ryota mentioned that, about confidence itself, it seems easy to compute. About metacognition or meta-representation, it would be interesting to know what architecture capture meta-representation and how computationally define it.

Matt said, in model base, the decision-making system retrieves the representation, however it still not a representation of itself. Adding another level of metacognition, such as social cognitions may be important.

Hiroyuki Nakahara raised concerns about the perspective of goal directedness, whether it can be related to model, or simulation, or inference of others behavior? … it is a broad question. David Silver mentioned that reward may be good enough to generate intelligent behavior. Nakahara continued by considering 2 ways: the only thing to be concerned about is the goal; or a simple reward is the basis for very complex behavior.

Also, he mentioned that meta-“something” seems to include something “without”-meta. The question is, what are we trying to uncover?

Kenji Doya mentioned that, traditionally people assumed modularity and hierarchical structure in the brain, however from Matt and Xiao-Jing work, meta-learning can be achieved without structure. Ryota also presented distributed predictive models, where instead of having metacognition, some coherent space with an overlapping function may be a form of metacognition.

More details were provided by Ryota about his work, where 10 NN are specialized for some tasks, for a given problem. However, he was wondering what area of the brain represents meta-representation of NN or the function of other regions?

Angela mentioned the control over the configuration. In recurrent NN, the configuration is not implicit but embedded in the weights. Metacognition is something like thinking about thinking. Humans can track their thinking, but not necessarily control the configuration of the system.

Xiao-Jing mentioned that he would like to look the learning-to-learn for different kind of problems, involving different blocks of cognition. Having some progress in that front, we can start to question about how we can control.

Matt mentioned that, deep RL will arrive to model-base RL (as shown in his 2018 paper) on explicit knowledge, procedural knowledge so far. But, to achieve reasoning about reasoning, agents may need to include language, or models of themselves (he introduced the driving drunk example). That still an open question.

Angela added, in biological agents, part of the structure that declares knowledge is social. In human, reward can be the approval from others, and it only exists in the interaction between agents. The social context is part of the meta learning or cognition.

Nakahara mentioned, once we uncover those things, they could be embedded as a meta part or part of the basic systems.

Matt thinks that actually, there is nothing special about meta-learning, it involves any system with large memory, training on many tasks, and with some structure.

Wang mentioned that social cognition is quite interesting. Nakahara said, given the requirement about what we need to learn for survival, (I missed the last part here, just last words before the end of the discussion session).

Below is my personal comments on the topics at this session.

Current AI outperformed the learning of specific tasks using mainly connection weight update and deep layered structures, however, still not able to flexible solve multiple and novel tasks. Meta-learning seems to be a key component on achieving artificial general intelligence. The AI symposium was an unique opportunity to better understand the different, but related, meanings and approaches towards meta-learning and meta-cognition.

One point of view assumes that meta-learning tries to realize adaptive processes in the system (the brain, an AI system) by connecting neuronal circuits in meaningful ways. That adaptation might be realized by gating mechanisms, weight update in subspaces, functional modularity and hierarchies; thought all of those remain poorly understood.

Some comments raised about the need of mapping meta-learning to discrete and localized structures, to understand whether it is computed locally, or a decision process is taking place within a hierarchy. Related to that, mapping what area of the brain represents the function of other regions, also known as meta-representation, could provide a cue on how “meta” computations are taken place in the brain.

Another view focus on the neuromodulators and their role in the regulation of learning. Particularly, the dopamine is thought as reward for goal fulfillment, however, it is not easy to define in neurosciences what really constitutes a reward. The devaluation of goals is a recent view on this. For example, the motivational stage in evolving rodents when chasing water, dictates if they consider that as a goal. A redefinition of what constitutes goals/rewards and their relationship with neuromodulators on different windows of time seem new research directions. Indeed, the perspective of goal directness still subject of concerns, whether reward may be good enough or not to generate intelligent behavior.

An emergent perspective is the addition of another level of meta-cognition, such as social cognition, for example, specie’s survival as learning or inference of other’s behavior.

Another arising topic considers meta-cognition as something like thinking about thinking. Humans can track their thinking, but not necessarily control the configuration of the system, while in recurrent artificial NN the configuration may be embedded in the weights. Model base RL showed that some form of representation is retrieved by the decision-making system, however it still not a representation of itself.

In conclusion, traditionally meta-learning assumed modularity and hierarchical structure in the brain, however, the results presented in the symposium showed that it may be achieved without structure. In addition, preliminary results indicate that some form of metacognition can be reached by distributed predictive models in a coherent space with an overlapping function.

The different views on this topic demonstrate its potential and importance on the understanding of distributed learning systems in the brain. The next practical steps, perhaps, should look more to the dynamics, the understanding of trajectories and vector field changes.