**Report on the discussion session - Session 5. Social Impact and Neuro-AI Ethics**

Ho-Ching Chiu, OIST

Chair and Discussant: Masamichi Sakagami

Speakers: Anne Churchland, Kenji Doya, Arisa Ema, Hiroaki Kitano, Stuart Russell

Masamichi Sakagami started the discussion by introducing some studies on pro-social behaviors: the extent of prosocial-ness and proself-ness reduced when thinking time increased; also proself people have higher prefrontal cortex activity while prosocial people have higher amygdala activity. Sakagami suggested that there might be a link between model-free/model-based mechanism with proself/prosocial-ness, but the detailed neurological mechanism is still unclear.

Sakagami asked Anne Churchland how we can utilize our uninstructed movement activities for brain inspired AI system. Churchland answered that the difficulty is that a majority of electronic signal measured from brain is just low level or movement related activities that we do not care. From an evolutionary perspective, brain is likely developed for movement, and other higher activities just hijacked the machineries and evolved from the movement-oriented brain. Social behavior is also one of basic functions of our brain, since large part of brain activities are devoted to it. Learning how those circuits are repurposed to do other activities will be relevant in developing artificial networks.

Kenji Doya commented that recent advancement of measurement technology might help the dynamic investigation of meaning of different brain activities.

Terry Sejnowski supplemented Churchland’s answer that language is one of the examples of subsequently evolved brain functions, due to the substantial reference of body and space relationships. Sejnowski also mentioned that we are just evolved in the ability to learn to do many things, and practices are needed. Stuart Russell argued, using the example of computer, that despite the first computer was used to decode German, but it turned out the implementation was just logic gates which is vastly general. Therefore, activities other than those biologically oriented do not necessarily have to be based on those evolutionarily designed.

Churchland raised that there are obviously fundamentally different types of motor system like that of ocular system compared to motor and pre-motor cortex. Which suggested there are multiple different motor systems. She also suggested a hypothesis that cognition is at least linked to movement to certain extent, given the difficult of eliminating movements.

Doya asked Kitano whether embodied experience is necessary for one to understand knowledges. Kitano’s opinion is that to develop cognition and thought process embodiment might be involved but not essential.

Arisa Ema suggested that it is also important to consider social, institutional and legal kind of influences when we create the brain inspired models.

Russell then stated that the difference between human and machines is whether it has purposes. The current design of AI systems is based on the notion that intelligence means to have a purpose and to plan how to achieve goals, and so that machines should also have purpose to be intelligent. But this is a mistake because we do not have a way of transferring our entire preference structure to the machine and we end up transferring fragments of it. It might end up we have a machine that is a psychopath, pursuing a distorted version of what we think our objective might be and doing it in a completely single minded way, completely oblivious of anything that is not specified in the objective. What is preferable is some system that is directed entirely towards human’s benefit and its extension, like the environment.

Also, it knows that it does not know the objective.

Another philosophical question raised by Russell is how a machine should make decisions on behalf of mankind. How to trade off the preferences of lots of people? One scenario is if someone derive pleasure from the suffering of others, the preference of those persons should be weighted zero since those people have put themselves outside the social contract. Then machine could be designed in a way that it observes humans and figure out their preferences.

Furthermore, Russell asked how many reward is enough? Where does the reward come from? Is it purely biological basis of dopamine which relates to pain, hunger and reproduction? Russell suggested that human may act in a way by observing the others to learn about other’s preference and possibly eventually to adopt other’s preference. So that the reward system is not just driven by biology but also by culture.

Angela Langdon asked Russell how about the case of addiction, in which there is mismatch of preference between long and short term – in long term the addicts want to integrate to society but short term their preference does not align with their ultimate goal. In this case what should be taken as preference?

Russell respond that there is imperfection in human, so that the mapping of actual behavior to the true underlying preference are very complicated. We need to be able to reverse engineer it for designing AI systems. Also, preferences are not stable and are modifiable since people are not born with those set of complex preferences. AI systems will take advantage of that, and an example is social media. Whether the notion of preference change is good or bad is nevertheless an open question, since it depends on what the change is.

Ema asked, from a philosophical perspective, what should be the starting question an AI researcher should ask? Sejnowski responded that early philosophy was actually oriented towards physics and science and we are heading in the same direction, but now with the brain. And he mentioned the emergence of a field called neuro-philosophy. Churchland suggested that to discuss this issue it needs to bring together a diverse group of people who is expert in dealing with people whose brains are fundamentally altered, like those with addictions.

Doya commented that there are mechanisms of human society, like democracy and peer reviewing to harness runaway of intelligences. Kitano added that any solutions could not be optimal for all time horizons, like the importance of scientific discovery depends on value systems. An example is that discovery of cure for C. elegans cannot get Nobel Prize. Maybe AI should develop a dynamical way for aligning to different value systems.

Russell emphasized that we are not uploading any value system to AI so that the AI should be constantly uncertain about what people want and how to trade off those preferences.