Consciousness beyond the human case

Chris Frith

In: LeDoux, J., Birch, J., Andrews, K., Clayton, N.S., Daw, N.D., Frith, C., Lau, H., Peters, M.A.K., Schneider, S., Seth, A., et al. (2023). Consciousness beyond the human case. Curr Biol *33*, R832-R840.

Two levels of consciousness

When a creature is conscious it is having subjective experiences. The mechanisms that underpin such experiences have evolved. Early in evolution there were creatures in which control of behaviour occurred without any subjective experience. Later, sentient creatures appeared. Such creatures create a model of the world to guide their interactions, which frees them being slaves to external stimuli. The subjective experience associated with sentience became richer as nervous systems became more complex. Later still creatures appeared, best exemplified by humans, who can reflect in their subjective experiences and discuss these experiences with others. Such discussions provide an additional means for developing our models of the world. In humans, these three levels of control operate in parallel.

I believe that many animals are sentient, having the bottom two levels of control (Ginsburg and Jablonka, 2019). A few may reach the third level, but this third level is highly developed only in humans. Large language models (LLMs), in contrast, having been trained on words, have, in some sense, the highest level without the two lower levels. In this case they are not conscious. They have no sentient level on which to reflect (Lau and Rosenthal, 2011).

But what does it mean to have only this highest level? And what would be needed for AIs to become conscious? To answer this question, we need to study the interactions between the different levels.

The importance of metacognition

Signals from lower levels provide information about the working of the brain/mind (metacognitive signals). For example, we have experiences of fluency: how quickly and easily we perceive an object, how quickly and easily we choose an appropriate action. These feelings of fluency can be interpreted as markers of confidence. By sharing our degree of confidence with others we can improve our decision-making (Fusaroli et al., 2012).

These signals from the lower levels have a vital role in keeping our models of the world in check. Without them we would not have the deference to the world that we need to distinguish appearance from reality. LLMs lack these signals. LLMs have a model, but it is not a model of anything. It is not grounded in sentience. AIs do not have the constraints provided by metacognitive signals. They have no clue when they are deviating from reality.

Learning from instructions

Signals from the highest level modify the functioning of the lower levels. And these signals often come from others in the form of verbal instructions (Heyes et al., 2020). But how does this work?

We can contrast learning from instructions with learning from direct experience. For example, in threat conditioning we learn that when we see a cue, such as a blue square, it is likely to be followed by a painful shock. This association is built up slowly by trial and error. The unexpected shock elicits a prediction error, but, once the cue elicits a prior expectation of shock, the prediction error is no longer elicited. In contrast, if the instruction '*from now on the blue square will be followed by a shock'* is given, then a prior value is immediately attached to the cue and prediction errors are suppressed. No association needs to be learned. We just need the prior (Lindström et al., 2019).

LLMs learn solely by verbal instruction. They build up an impressively complex network of prior expectations. It would be as if you had knowledge of an unknown country from the fanciful tales of explorers, with no experience if your own country to relate to it. There wouldn't be anything to be 'like'. If AIs are to achieve consciousness, lower-level systems of learning through direct experience will have to be incorporated.

references

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., and Tylen, K. (2012). Coming to terms: quantifying the benefits of linguistic coordination. Psychol Sci *23*, 931-939.

Ginsburg, S., and Jablonka, E. (2019). The evolution of the sensitive soul: learning and the origins of consciousness (MIT Press).

Heyes, C., Bang, D., Shea, N., Frith, C.D., and Fleming, S.M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. Trends in Cognitive Sciences *24*, 349-362.

Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences *15*, 365-373.

Lindström, B., Golkar, A., Jangard, S., Tobler, P.N., and Olsson, A. (2019). Social threat learning transfers to decision making in humans. Proceedings of the National Academy of Sciences *116*, 4732-4737.