Analyzing internal world models of humans, animals and AI

July 18 2024

Credit: Pixabay/CC0 Public Domain

A team of scientists led by Prof. Dr. Ilka Diester, Professor of Optophysiology and spokesperson of the BrainLinks-BrainTools research center at the University of Freiburg, has developed a formal description of internal world models and published it in the journal *Neuron*.

The formalized view helps scientists to better understand the development and functioning of internal world models. It makes it possible to systematically compare world models of humans, animals and artificial intelligence (AI). This makes it clearer, for example, where AI still has deficits compared to human intelligence and how it could be further developed in the future. Eleven Freiburg researchers from four faculties were involved in the interdisciplinary publication.

Humans and animals abstract general laws from everyday experiences. They develop internal models of the world that help them to find their way in unfamiliar contexts. Based on the abstracted models, they can make predictions in new situations and behave accordingly.

For example, knowing comparable cities that also have a city center, pedestrian zones and public transport can help them find their way around a foreign city. Even in social contexts such as dinner in a restaurant, comparable experiences help you to behave appropriately.

In order to formalize internal world models across species, the researchers distinguish between three abstract spaces that are intertwined: the task space, the neural space and the conceptual space. The task space encompasses everything that an individual experiences.

The neuronal space describes the various measurable states of the brain, from the molecular level to the activity of individual neurons through to the activity of entire brain areas.

The latter is visualized with the help of a functional magnetic resonance imaging (fMRI) scanner, for example, or measured using techniques such as high-density electrodes or calcium imaging. The equivalent of neural space in AI is the activity of the nodes within the corresponding artificial neural network.

The conceptual space consists of pairs of states of the task space and the neural space. These pairs thus represent the status of an individual, which links internal processes with external influences.

The current state is constantly changing by transitioning to the next state with a certain probability. These combinations of an individual's experiences on the one hand and the corresponding brain activity on the other, as well as the dynamic transitions, make the individual internal world models scientifically tangible.

With the help of the formalized view, scientists can now analyze internal world models across disciplinary boundaries and discuss how they arise and evolve. Findings from research on humans and animals, for example, should help to improve AI. For example, current AI systems are not yet able to check the plausibility of their predictions.

Even large language models such as ChatGPT have so far only functioned as pattern recognition machines without the ability to actually plan. However, planning is important in order to play through and correct strategies in unknown situations before they are implemented and possibly cause damage.

Researchers also suspect that deficits in internal world models could be the cause of some mental illnesses such as depression or schizophrenia. A deeper understanding of world models could help to use medication and therapy in a more targeted way.

More information: Ilka Diester et al, Internal world models in humans, animals, and AI, *Neuron* (2024). DOI: 10.1016/j.neuron.2024.06.019

Provided by University of Freiburg