Tamás Biró:

John and the hammer broke the vase

Generative AI and the innateness of counterintuitiveness

An early assumption in the cognitive science of religion (CSR) has been that ontological categories (including HUMAN, ARTIFACT etc.) are innate, and so is agency. Here, we show that large language models (LLMs) can learn these categories from the data out there, in the surrounding world.

Observe the following sentences:

(1) *John broke the vase*.
(2) *The hammer broke the vase*.
(3) # *John and the hammer broke the vase*.
(4) *John broke the vase with the hammer*.
(5) # *The hammer broke the vase with John*.

Sentences (3) and (5) are grammatical, but infelicitous (marked by the # sign), because the instrument of an action can only appear in the subject position, as in (2), if the agent is omitted. Sentences (3) and (5) would only be adequate in the fictive world of an animated cartoon, which bestows agency on counterintuitive hammers.

Prompting generative AI systems to draw sentences (1)-(5), we have observed they are interpreted by default according to folk theories: a hammer is an ARTIFACT, which cannot act as agent. Coercing a counterintuitive interpretation requires more complex prompting.

Although the training procedures of the LLMs are industrial secrets, it is unlikely they were trained to handle the above sentences. Consequently, they must have learnt the distinction between ontological categories purely from the training data. In turn, if the information can be acquired from the world without a prior bias, then the argument for innateness becomes weaker. The human mental system might but needs not include prior biases.

Keywords:
large language models; agency; counterintuitiveness; generative AI;
innateness; ontological categories.

Submitted to open panel "25. *Computing Religion: Computational Approaches, Research Databases, and Quantitative Methodologies in the Study of Religions*."