ODDĚLENÍ ANALYTICKÉ FILOSOFIE | FILOSOFICKÝ ÚSTAV AV ČR
DEPARTMENT OF ANALYTIC PHILOSOPHY | INSTITUTE OF PHILOSOPHY CAS

PŘEDNÁŠKOVÁ SÉRIE
LECTURE SERIES

Tobias Schlicht (RUB)

Assessment or Attribution of Consciousness in AI Systems?

Butlin et al (2023) outline a research program for assessing consciousness in AI. They assume computational functionalism as a working hypothesis since it allows for conscious AI in principle and claim that well supported neuroscientific theories of consciousness can help us to assess consciousness in AI.

First, I scrutinize both computationalism and functionalism separately as different assumptions.Computationalism requires medium independence (Haugeland 1997) which is stronger than multiple realizability. Second, I argue that computationalism is not empirically supported and it is difficult to see how, for any neuroscientific theory, a computationalist interpretation could be better supported than its biological alternative. Neuroscientific theories determine neuronal signatures of consciousness which involve biochemical details that may be crucial for consciousness in human brains. Derivations of computational models which could be implemented artificially and used for the assessment of consciousness in AI require abstraction from such details and idealizations introducing falsehoods.

We face significant epistemic limitations regarding an evaluation whether computationalism is true of the biochemical details matter. While digital computation may be medium independent, neural computation may be not. Relying on prior work by Chirimuuta (2022), Cao (2022) and Block (2005), I make a case for the medium dependence of neural processing. The appeal to neuroscientific theories does not render their computationalistinterpretations empirically plausible. Since neurophysiological markers are absent in AI systems, and other markers of consciousness are unreliable (Bayne et al. 2024) – such as report (cf. the outputs of Large Language Models) – we can only attribute consciousness from the intentional stance (Dennett 1987) rather than assess it in AI systems.

19/06/2025 | 15.00 CET

Seminar room of Center for Medieval Studies, Jilská 1, Prague