****
********** HORA INFORMATICAE **********
****
**********(založeno v r. 1994)**********
****

V pondělí 18. 1. 2016 v 14:00 se bude
ve velké zasedačce (č. 318, 2. poschodí)
Ústavu informatiky AV ČR, v.v.i.,
Pod Vodárenskou věží 2, Praha 8 - Libeň
konat přednáška

Getting to Grips with Superintelligence and AGI through the Epistemic Theory of Computation

Jiri Wiedermann, UI AV CR
Jan van Leewuen, Center for Philosophy of Computer Science Utrecht University, NL

We study the nature of superintelligence from the epistemic point of view. To this end we consider AGI systems as computational systems generating knowledge over some epistemic domain. Rather than considering such systems based on a vaguely described idea of ``self-improving software", as it is customary in the literature about superintelligence, we will consider such systems operating with self-improving epistemic theories that will automatically increase their understanding of the world around them. This is quite a concrete idea for which it is possible to outline algorithmic principles by which the self-improving theories can be constructed. Then we concentrate on the problem of aligning the behavior of AGI systems with human values in order to make such systems safe. Necessarily, the behavior of any such system will depend on its own worldview and we show that there is no "universal" AGI system whose values will always agree with the values of any other system. Finally, based on the principles of interactive proof systems we design an architecture of AGI systems and an interactive scenario that will enable to detect in their behavior deviations from the prescribed goals. The conclusions from our epistemic analysis of superinelligent systems temper the over-optimistic expectations and over-pessimistic fears of singularity believers by grounding the ideas on superintelligent AGI systems in more realistic foundations.

Přednášet bude první autor.