Seminar Hora Informaticae **Institute of Computer Science,Prague**

Tuesday, June 11, 2024, 14.00 – 15.30 (2 – 3:30 PM) CEST

Meeting Room 318, Address: Pod Vodárenskou věží 2, Prague 8

ZOOM Meeting ID: 954 7823 4977 ,   Passcode: 712564

ZOOM: **https://cesnet.zoom.us/j/95478234977?pwd=dXoyekFHbDJ0MkNrTjVVS3F2STZqUT09**

David Herel, Department of Cybernetics, Faculty of Electrical Engineering, CTU Prague:

Language models and Artificial Inteligence.

In this talk, I will delve into the evolution and advancements in AI chatbots, from scripted dialogues to sophisticated language models of today, as exemplified by our award-winning Alexa Prize chatbot. However, with evolution comes new challenges. One such challenge is the susceptibility of these systems to adversarial attacks. My research has delved into the pressing issue of adversarial attacks and the need for robust defenses in AI systems. As we navigate this landscape, the question of achieving human-level AI remains. I'll share insights from my research into evolutionary algorithms and sugar search, an innovative approach favoring divergence over convergence, offering a potential solution. I will also show how current language models could be improved by "thinking" tokens.

References:

(1) Emergence of Novelty in Evolutionary Algorithms **David Herel** and Dominika Zogatova and Matej Kripner and Tomas Mikolov In *Artificial Life Conference Proceedings*, 2022. ALIFE 2022

(2) Thinking Tokens for Language Modeling **David Herel** and Tomas Mikolov In *Artificial Intelligence and Theorem Proving Proceedings*, 2023. AITP 2023

David Herel (https://davidherel.com/) is a PhD student in the Computer Science doctoral study program at the Department of Cybernetics, Faculty of Electrical Engineering (FEE), Czech Technical University (CTU) in Prague. His research is focused on Novel Methods in Natural Language Processing under the supervision of Tomas Mikolov. David is the third winner of prestigious Alexa Prize Socialbot Grand Challenge 5. He regularly contribute to top-tier conferences in the AI&NLP field (ICLR, ECAI, Amazon Science).

HORA INFORMATICAE (meaning: TIME FOR INFORMATICS) is a broad-spectrum scientific seminar devoted to all core areas of computer science and its interdisciplinary interfaces with other sciences and applied domains. Original contributions addressing classical and emerging topics are welcome. Founded by Jiří Wiedermann, the seminar is running since 1994 at the Institute of Computer Science of the Czech Academy of Sciences in Prague.

https://www.cs.cas.cz/horainf