IV130 Přínosy a rizika inteligentních systémů

Jiří Zlatuška

Úvod

23. února 2024

Zaměření předmětu

- Motivace, metodologie, paradigmata otevírající uvažování směrem k éře masivního využívání umělé inteligence (AI) ve všech ohledech;
- *nikoli* zvládnutí vytváření/programování AI, její zvládnutí v technickém smyslu nebo dostatečná průprava pro pracovníky v AI.
- Vazba inteligence v obecném smyslu a konstrukce obecně užitelných strojů zpracovávajících informace (počítačů) s dopady náhrady lidských kognitivních činností inteligentně jednajícími stroji (spolupracujícími s člověkem, i jako jeho protivníci).
- Dopady na kyberbezpečnost nasazovaných systémů.
- Bezpečné systémy stavějící na AI a vyhýbající se konfliktu cílů sledovaných strojem a zájmů člověka.
- Zásadní etické otázky včetně dopadu do bezpečnostních politik (od lokálních po geopolitiky).
- Rozšíření technicko-operativního pohledu informatika–bezpečnostního experta o pohled dopředu a vědomí hrozeb i příležitostí, které AI přináší

Témata

- Inteligence v lidech a strojích, inteligence, paměť, výpočty, učení
- Možnosti a limity strojů, biologie a evoluce
- Aktéři, prostředí, záměry, tvorba a uskutečňování plánů
- Určitost prostředí, nejistota, pravděpodobnost, teorie her
- Vyvozování, logické systémy, pravděpodobnostní jazyky, Bayesovské sítě
- Vývoj inteligentních systémů, metody učení, učení s učitelem, zpětnovazební systémy, hluboké učení, učení na základě vysvětlování, GPT
- Výhled superinteligence, dopady, fikce, nebo nebezpečí
- Rizika AI, průlomy, chyby vs. robustní systémy, dohled, duševní bezpečnost, autonomní zbraně, substituce lidské práce, ekonomické dopady
- Dopady superinteligentní AI, problém gorily a problém krále Midase, inteligenční exploze, problém ovládání
- Koexistence inteligentních systémů a člověka, princip prospěšných strojů, přátelská AI
- Matematické principy formulace prospěšných strojů, asistenční hry, problém vypnutí stroje
- Preference a hodnoty, jejich neurčitost, zjišťování a přizpůsobování, psychologie a technologie
- Důsledky pro budoucí vývoj, politiky velkých dat, ochrana soukromí, bezpečný vývoj inteligentních aplikací

Doporučená literatura

- Stuart J. Russell. *Jako člověk: umělá inteligence a problém jejího ovládání*. Praha: Argo/Dokořán, 2021.
- Stuart J. Russell a Peter Norvig: *Artificial intelligence: a modern approach*. Fourth edition, global edition. Harlow: Pearson, 2022.
- Max Tegmark: *Život 3.0, Člověkem v éře umělé inteligence*. Argo/Dokořán, 2020.
- Nick Bostrom. *Superinteligence: až budou stroje chytřejší než lidé*. Praha: Prostor, 2017.
- Erik Brynjolfsson a Andrew McAfee: *Druhý věk strojů: práce, pokrok a prosperita v éře špičkových technologií*. Brno: Jan Melvil, 2015.

- Doporučený souběžný předmět jako doplněk k netechnickým aspektům chování: PV226 Psychologie (přednášející doc. Jiří Dan, předmět je realizován v rámci semináře LaSArIS, ale máte možnost se do něj zapsat bez dalšího omezení nebo návazností)

Požadavky k zakončení

Eseje v rozsahu cca 3-4 tisíce slov (6-10 stran textu) *s využitím a citováním další literatury a zdrojů* kromě přednášek – ve stylu odborného „článku“ vycházející z nějakého výchozího pohledu, rozvíjející téma a formulující závěr

Zkouška: Dva samostatné eseje, **první z nich** na vámi zvolené konkrétní téma s jeho výkladem a hodnocením možných dopadů pro kyberbezpečnost (resp. bezpečnost obecně), s využitím literatury a uvedením odkazů na ni mimo rozsah přednášek, **druhý z nich** na některé z témat uvedených výše v nástinu obsahu, opět s využitím literatury přesahující obsah přednášeného