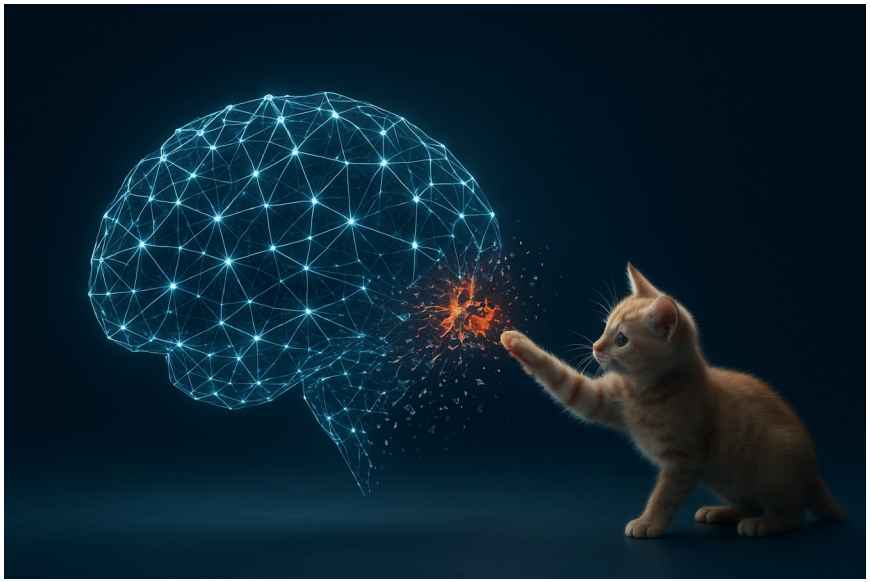
Adversariální útoky na neuronové sítě: skrytá křehkost v srdci AI

Stanislav Fort (AISLE)

Deset let zpátky byl prvně popsán neobvyklý jev, kterému dnes říkáme adversariální útoky. Malé perturbace vstupu do neuronové sítě vedou k obrovským rozdílům ve výstupu a to i přes to, že pro člověka je vstup (například obrázek kočky) v podstatě nezměněn. Tento efekt se vyskytuje v neuronových sítích od těch nejmenších modelů až po ty největší AI systémy dneška, jako je například GPT-5, Claude 4.5 Sonnet, nebo Gemini 2.5 Pro. Za dekádu od jejich objevu bylo napsáno již 10 000 vědeckých článků, ale jejich vysvětlení nebo řešení nám pořád unikají.

V této přednášce si popíšeme historii tohoto problému, jeho propojení s vysokorozměrnou geometrií, tradiční přístupy k jeho řešení a vizi do budoucnosti.



Stanislav Fort, absolvent Standford University a University of Cambridge v oborech umělé inteligence a teoretické a matematické fyziky, je zkušeným odborníkem s praxí ve společnostech jako Google, DeepMind a Anthropic. Nedávno spoluzaložil startup AISLE, který se zaměřuje na vyhledávání a odstraňování kritických zranitelností pro zvýšení bezpečnosti softwaru.

středa 26. listopadu

17:30 v posluchárně K1

MFF UK, Sokolovská 49/83

nebo live stream na YouTube

MATEMATICKÉ PROBLÉMY NEMATEMATIKŮ

www.karlin.mff.cuni.cz/mffseminar najdete nás i na FB