PREDNÁŠKA SPOJENÁ S DISKUSIOU NA TÉMU

UMELÁ INTEL!GENCIA
DOBRÝ SLUHA A ZLÝ PÁN?

... ľudia, stroje, spoločnosť, etika, hodnoty ...

... riziká a problémy súčasných systémov ...

... superinteligencia ako (nielen) etická výzva ...

... vplyv nástrojov na báze AI vo vzdelávaní na VŠ ...

Peter Šantavý

Uvedenie do problematiky umelej inteligencie (AI)

Limity a riziká súčasných systémov AI

Návrh riešenia etických problémov slabej AI

Vízia silnej a všeobecnej umelej inteligencie

Umelá inteligencia je ako Colombova žena,

nikto ju nevidel a všetci o nej hovoria...

Uvedenie do problematiky AI (artificial intelligence)

Technológie umelej inteligencie (AI)

***Buzzword*, *hype* alebo realita?**
Čo môžeme považovať za AI?

- **technológie sa prelínajú – o AI sa hovorí i tam, kde ide len o IA***

- **technológie AI skutočne fungujú**
- **technológie AI sú súčasťou nášho života**

*** AI – artificial intelligence**
IA – intelligent assistance

Technológie umelej inteligencie (AI)

Klimatická zmena v oblasti AI nastáva...

- **striedanie ročných období** (**AI spring vs. AI winter**)
  striedanie aktívneho rozvoja a útlmu v oblasti umelej inteligencie
- **klimatická zmena nastáva** – schopnosti súčasných systémov umelej inteligencie i miera ich akceptácie a adaptácie v modernej spoločnosti **presiahli hranicu**, za ktorou je ťažké si predstaviť návrat späť...

Technológie umelej inteligencie (AI)

Definícia

- **inteligencia**
  Súbor zručností, medzi ktoré patrí abstraktné a logické myslenie, ďalej predstavivosť, a teda aj schopnosť uvažovať o hypotetických možnostiach, a tiež aj jazykový cit (James Flynn).

- **umelá inteligencia**
  Umelá inteligencia je o vytváraní systému, ktorý vykazuje také správanie, o ktorom si myslíme, že vyžaduje inteligenciu (AI100).

- na prasknutie naplnený kufor (Marvin Minsky)...

Technológie umelej inteligencie (AI)

Základné vlastnosti

- **autonómnosť – schopnosť samostatne konať**
  Schopnosť systému vykonávať úlohy v komplexnom prostredí bez neustáleho vedenia používateľom.
- **adaptabilita – schopnosť sa prispôsobovať**
  Schopnosť zlepšovať svoj výkon (a schopnosti) učením sa (nielen) zo skúseností.

Technológie umelej inteligencie (AI)

Základné delenie – ANI a AGI

- **slabá umelá inteligencia (ANI)** – úzko špecializované systémy umelej inteligencie (narrow AI), ktoré sú **optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh**.
  Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú **inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát**.

  Hovoríme teda o systémoch, ktoré sú zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.

Technológie umelej inteligencie (AI)

Základné delenie – ANI a AGI

- **všeobecná umelá inteligencia (AGI)** – tzv. silná (strong) a všeobecná (general) AI.
  Všeobecná, lebo **dokáže zvládnuť akúkoľvek intelektuálnu úlohu a má schopnosť zovšeobecňovať** a prenášať, či adaptovať naučené schopnosti na iné úlohy.
  Silná, pretože aj **skutočne rozumie tomu, čo rieši a vykonáva**.

Technológie umelej inteligencie (AI)

Základné delenie – symbolická a subsymbolická AI

- **symbolická AI** – napodobňuje ľudské myslenie na úrovni pojmov, slov, fráz (= symboly) a vzťahov medzi nimi.
  Na základe definovaných pravidiel a postupov („ak niečo, tak potom toto“) sú jednotlivé symboly spracovávané a vykonávajú sa priradené úlohy.

  Symbolická AI – veľmi zjednodušene povedané – sa pomocou matematickej logiky snaží emulovať procesy myslenia.

  Čisto symbolické systémy zlyhávali a zlyhávajú pri riešení problémov, ktoré sa nedajú exaktne popísať a v reálnych prostrediach, ktoré nie je možné deterministicky uchopiť a sú plné nejasných informácií.

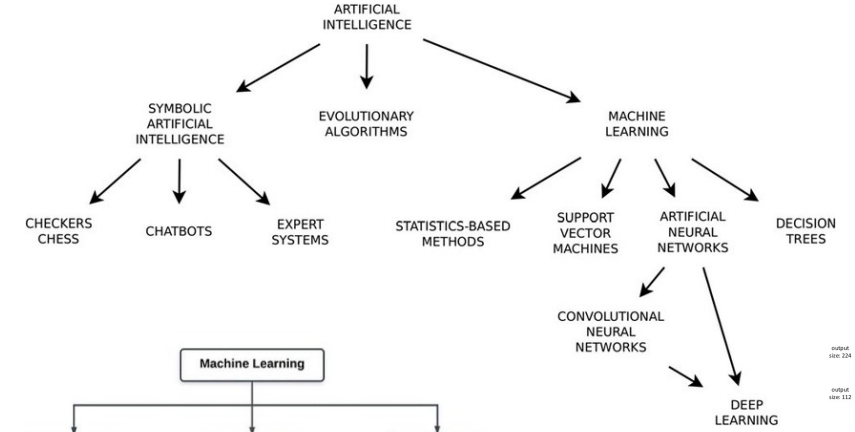
Technológie umelej inteligencie (AI)

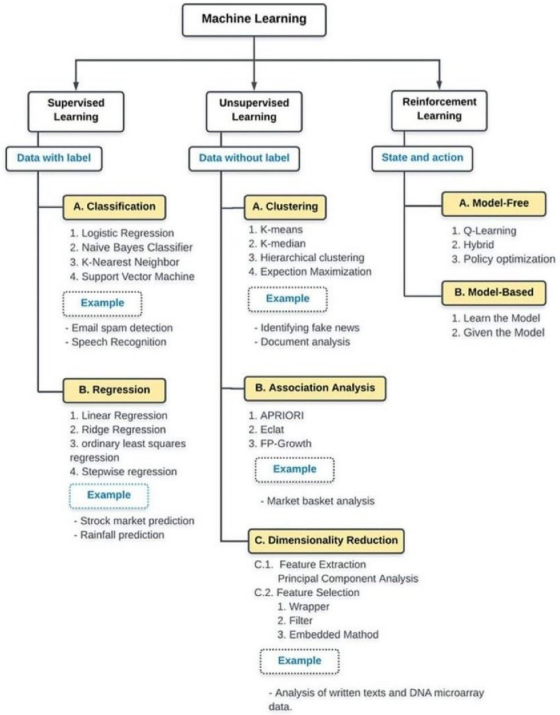
Základné delenie – symbolická a subsymbolická AI

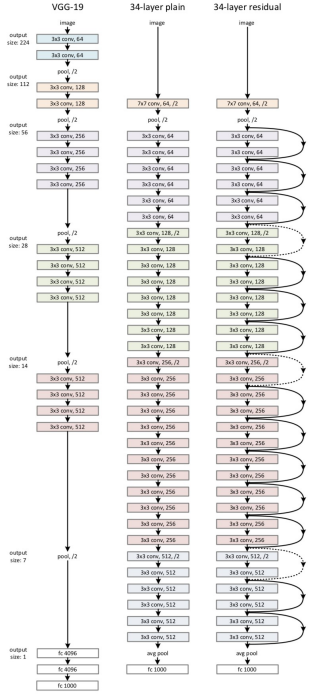
- **subsymbolická AI** – napodobňuje myšlienkové procesy, ktoré by sme mohli nazvať niekedy nevedomými, či automatickými, a ktoré sú základom tzv. rýchleho vnímania (fast perception), čo využívame napr. pri rozpoznávaní tvárí alebo identifikácii hovorených slov.

  Subsymbolická AI – veľmi zjednodušene povedané – sa snaží emulovať činnosť mozgu na úrovni neurónov.

  Väčšina moderných implementácií systémov AI, medzi ktoré patria systémy strojového učenia a neurónové siete, vychádza zo subsymbolického prístupu, ktorý sa snaží reálne problémy uchopiť a v určitej miere ich aj úspešne riešiť.

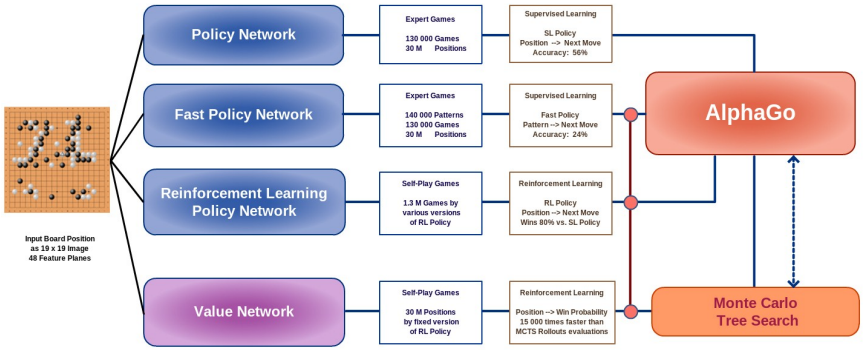


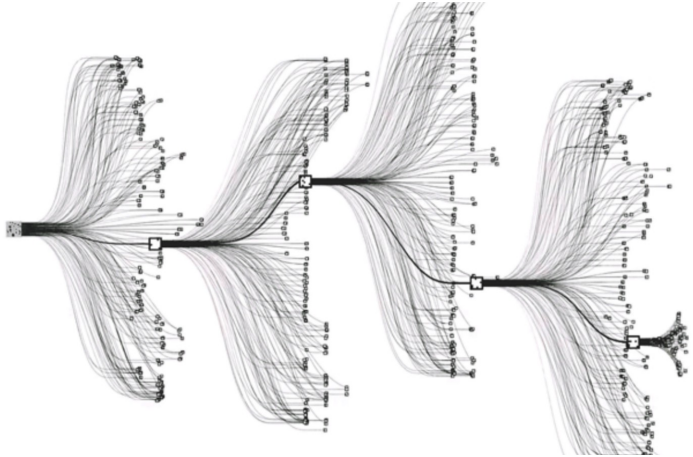




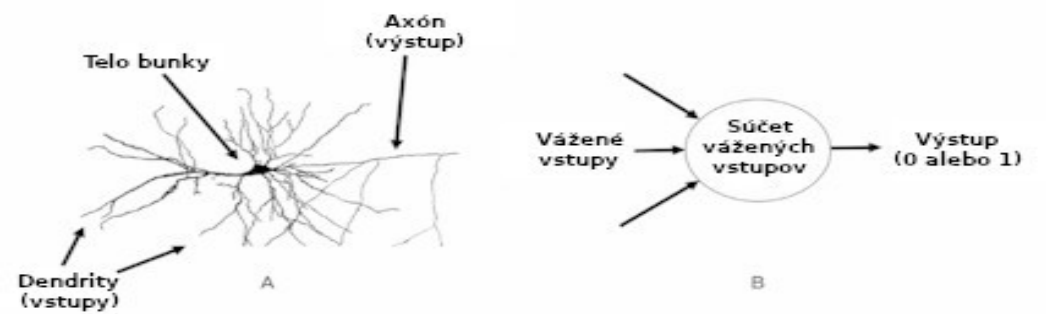
AlphaGo Overview

based on: Silver, D. et al. Nature Vol 529, 2016
copyright: Bob van den Hoek, 2016

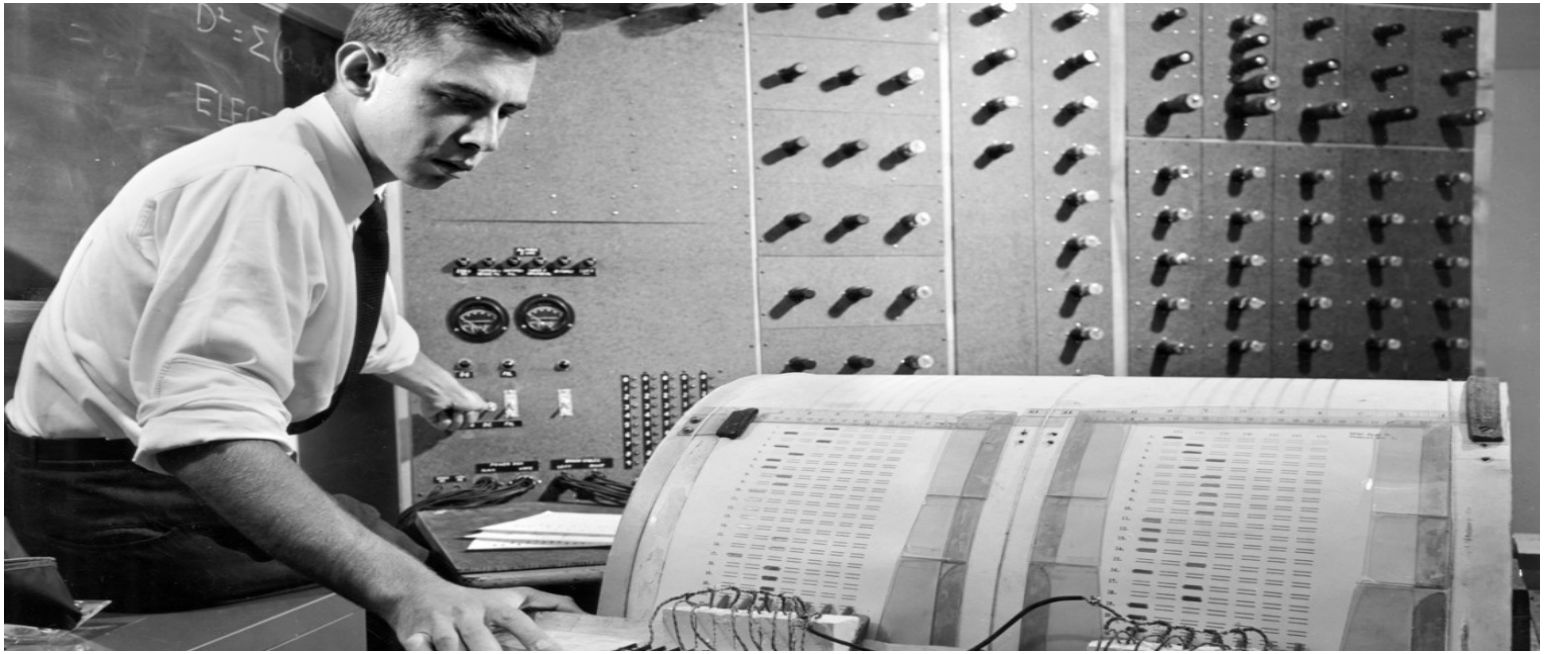




Kredit: Fabio Galbusera, Eric Gaubert, Bob van den Hoek, Seth Weidman

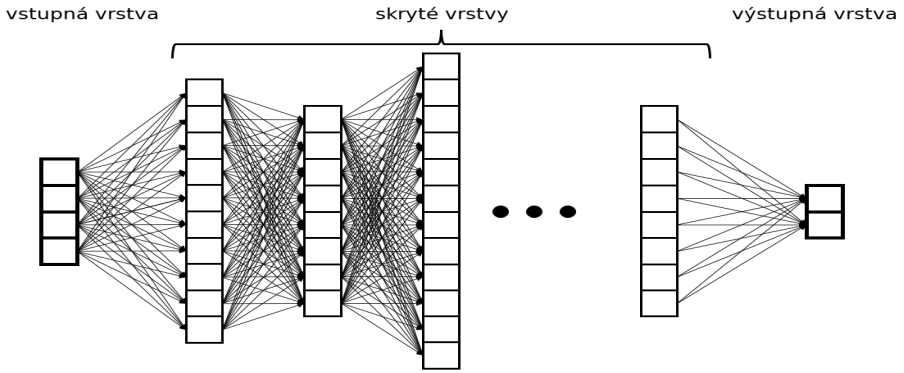


1957...

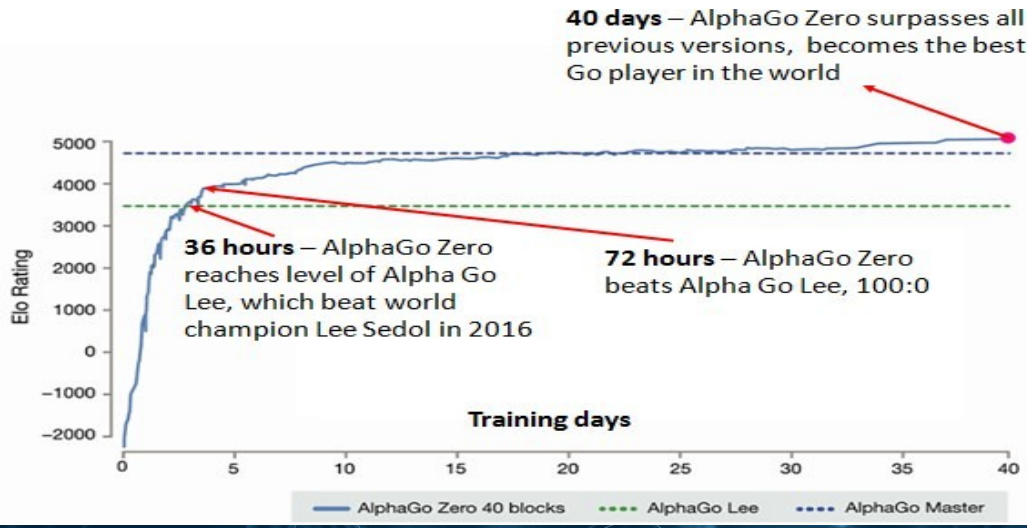


Perceptron vyvinutý psychológom Frankom Rosenblattom

Kredit: Ars Technica (arstechnica.com)



...súčasnosť



Kredit:

Smithsonian Air & Space Magazine (smithsonianmag.com)
AI for Good (ai4good.org)
KDnuggets AI website (kdnuggets.com)

D I S K U S I A . . .

symbolická vs. subsymbolická AI

neurónová sieť vs. neuróny v mozgu

Každá dostatočne pokročilá technológia je na nerozoznanie od mágie...

Limity a riziká súčasných systémov AI

Riziká súčasných systémov ANI

Technologické rizikové faktory

- **neurónová sieť ako „black box“** – prakticky **nevieme, na základe čoho robia hlboké neurónové siete svoje rozhodnutia, čo presne sa neurónová sieť naučila a ako spoľahlivo to dokáže aplikovať**.
- **dizajn špičkových systémov AI – alchýmia a metóda pokus/omyl**
  „To, čo v súčasnosti robíme, nie je veda, ale určitý druh alchýmie.“ (E. Horvitz)
- **zlyhania človeka vs. zlyhania AI** – nemáme „**teóriu mysle**“ systémov AI na spôsob modelu správania sa ľudských bytostí, ktoré je na základe fundamentálnych kognitívnych schopností (rozpoznávanie objektov, chápanie scény, porozumenie reči, atď.) štandardné a vývojom overené.

Riziká súčasných systémov ANI

Technologické rizikové faktory

- **zraniteľnosti, slabiny a klamanie systémov strojového učenia**

  *Súčasné systémy ANI môžu zlyhávať **najrozličnejšími a často neočakávanými spôsobmi** v dôsledku nemožnosti pripraviť dostatočne veľkú množinu trénovacích dát, prípadne ich nesprávnej voľby bez dostatočnej kvality alebo s predsudkami, nadmernému prispôsobovaniu sa tréningovým údajom, efektu dlhého chvosta, rizikám plynúcim z oklamania hlbokých sietí, ich zraniteľností a povier, pod čo sa podpisuje i nedostatok odbornej erudovanosti potrebnej pre dizajn a ladenie hyperparametrov pri príprave funkčného a úspešného riešenia.*

Riziká súčasných systémov ANI

Technologické rizikové faktory

- **zraniteľnosti, slabiny a klamanie systémov strojového učenia**
  – malá množina trénovacích dát (training dataset)
  – nesprávne zvolená/nekvalitná množina trénovacích dát a predsudky (biases)
  – nadmerné prispôsobovanie sa tréningovým údajom (overfitting to training data)
  – efekt dlhého chvosta (long-tail effect)
  – klamanie hlbokých sietí a ich zraniteľnosti
  (fooling deep neural networks and vulnerability to hacking, adversarial examples)
  – povery (superstition)

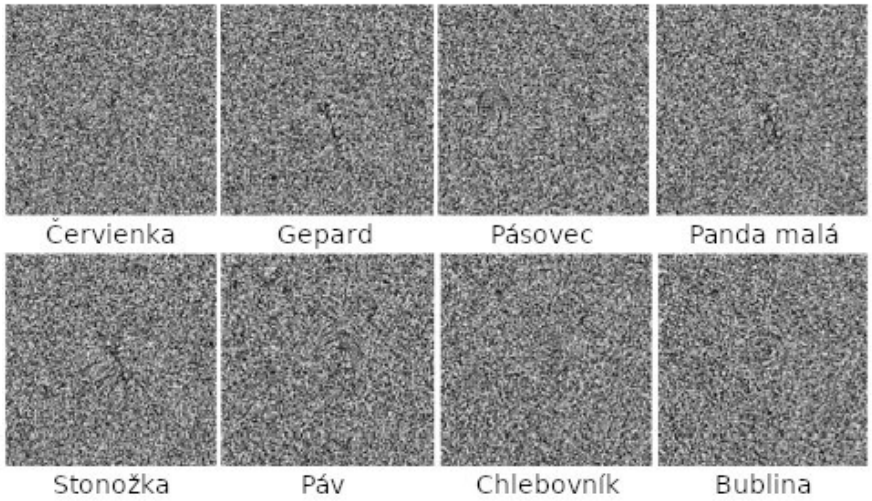
- **zlyhania sú odlišné od ľudských, ťažko predvídateľné, niekedy ľahko vykonateľné a mnohokrát prekvapivo robustné...**

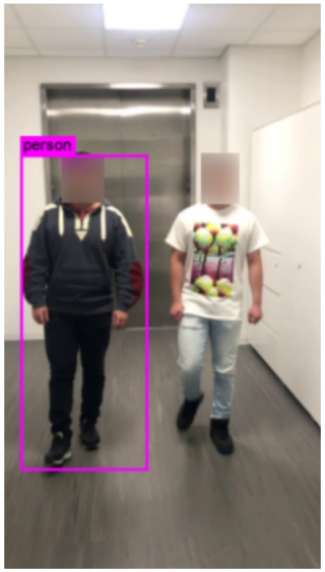
Riziká súčasných systémov ANI

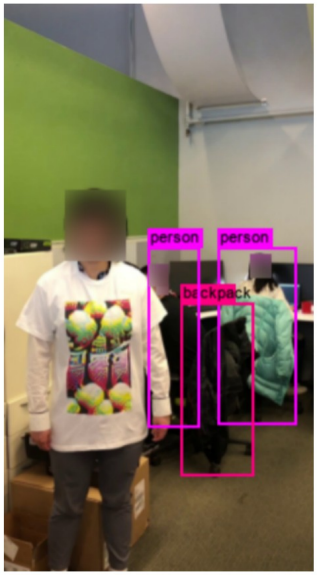
Technologické rizikové faktory

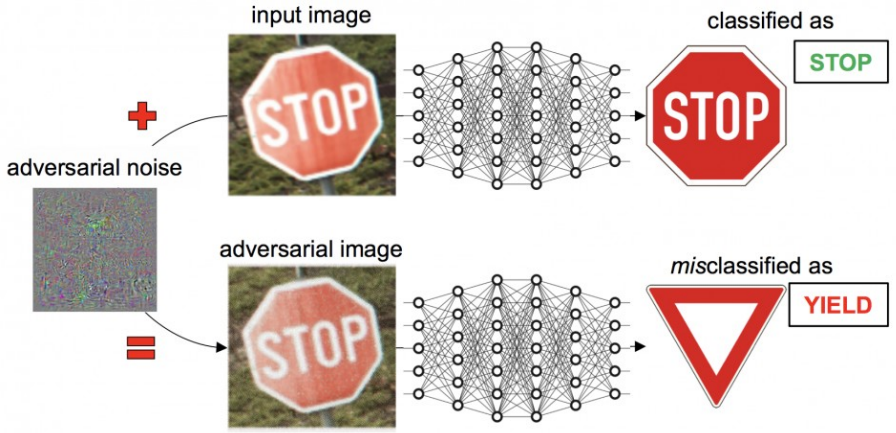
- **procesné riziká v reálnom nasadení**
  – **útoky na dôvernosť** (confidentiality attacks)
    Útoky zamerané na dáta uložené v rámci modelov systémov AI.
  – **útoky na zraniteľnosti** (evasion attacks)
    Odhaľovanie a zneužitie existujúcich zraniteľností v modeloch za účelom zmanipulovania výsledkov systémov AI.
  – **útoky s cieľom ovplyvniť model** (poisoning attacks)
    Ovplyvňovanie modelu, tréningového procesu a tým aj výslednej činnosti.

*V kontexte bezpečnosti procesov umelej inteligencie sa **etické výzvy, doteraz rozoberané ako súčasť návrhu a realizácie systémov umelej inteligencie, rozširujú aj o oblasť etiky použitia, resp. riziká zneužitia**.*









Kredit: Melanie Mitchell – upravené autorom, autor, K. Xu et al., Sourav Agarwal

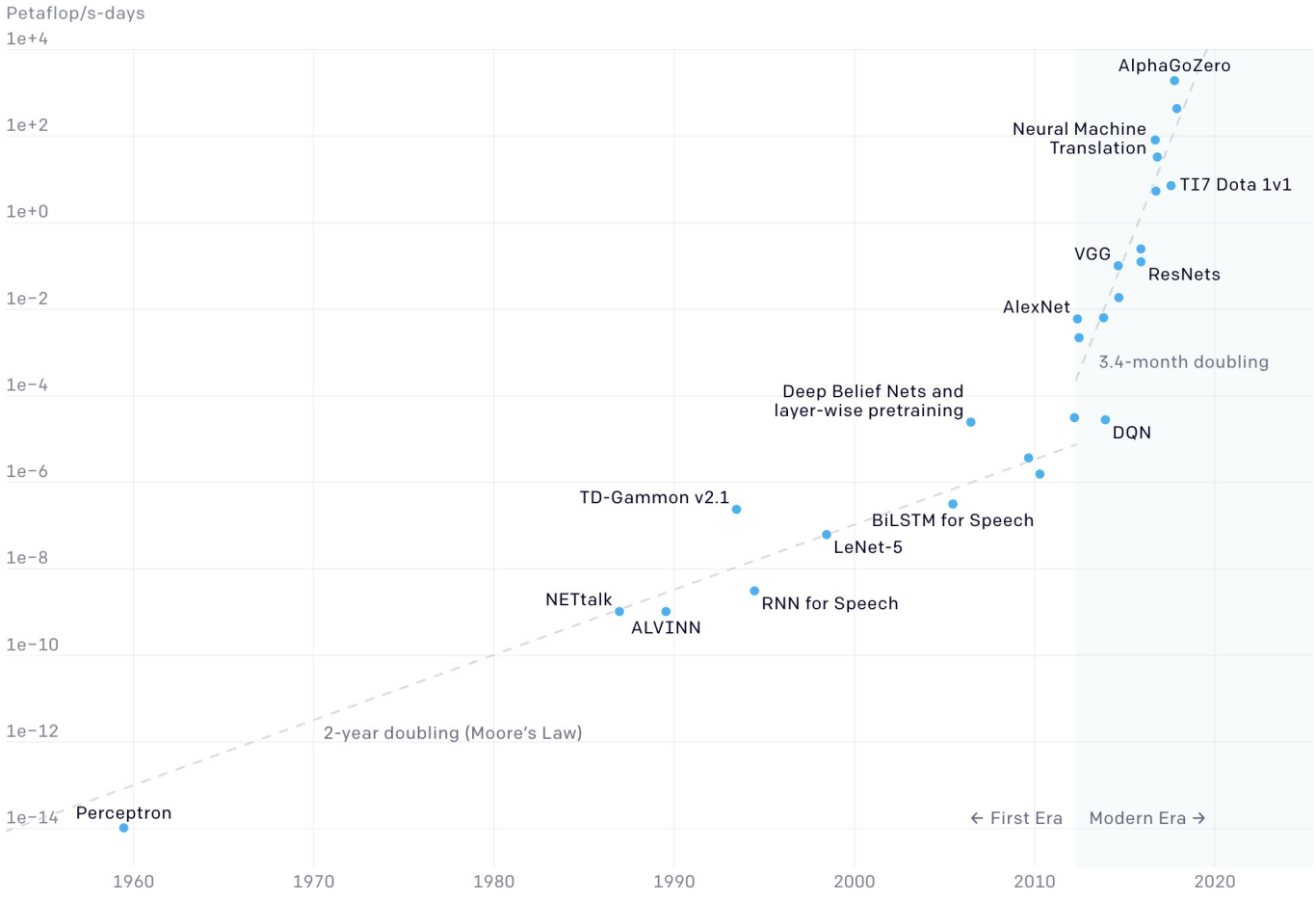
Riziká súčasných systémov ANI

Technologické rizikové faktory

- **kybernetická bezpečnosť – je to napadnuteľné**
  – vektor útoku na zraniteľnosti AI
  – **neexistuje dokonale bezpečný a spoľahlivý systém**

- **infraštruktúra a komplexnosť – je to náročné**
  – s rozvojom algoritmov strojového učenia sa požadovaný výpočtový výkon v poslednej dekáde zvyšuje **exponenciálne**
  – **extrémna komplexnosť**, je rizikovým faktorom bezpečnosti a stability fungovania systémov

Dopad komplexnosti a výpočtovej náročnosti na systémy ANI...



Kredit: D. AMODEI, D. HERNANDEZ

D I S K U S I A . . .

jednoduchosť útokov

robustnosť útokov

inakosť útokov

skryté, nepoznané a nepredvídateľné zraniteľnosti

Riziká súčasných systémov ANI

Oblasti dopadu rizík súčasných systémov AI (I.)

- **spoločnosť**
  sociálne siete, virtuálny svet, paradigmatické zmeny rodiacej sa inf. spoločnosti
- **médiá**
  spracovanie dát, odporúčanie a analýza obsahu, automatická tvorba obsahu,...
- **školstvo a vzdelávanie**
  personalizácia vzdelávania, hodnotenie študentov, vývoj vzdelávacieho obsahu, zahrnutie (M)LLM AI v najrozličnejších oblastiach využitia,...
- **zdravotníctvo**
  lekárska diagnostika, predikcia ochorení, analýza liečby, robotické aplikácie,...

Riziká súčasných systémov ANI

Oblasti dopadu rizík súčasných systémov AI (II.)

- **doprava**
  optimalizáca trás, riadenie dopravy, autonómne vozidlá,...
- **výroba**
  optimalizácia procesov, prediktívna údržbu, kontrola kvality, riadenie dodávateľského reťazca, vývoj a dizajn produktov,...
- **maloobchod**
  personalizovaný marketing, riadenie zásob, služby a komunikácia so zákazníkmi
- **poľnohospodárstvo**
  monitorovanie plodín, predpovedanie výnosov, detekcia škodcov, automatizácia

Riziká súčasných systémov ANI

Oblasti dopadu rizík súčasných systémov AI (III.)

- **finančný sektor**
  odhaľovanie podvodov, hodnotenie rizík, investičná analýza, algoritmické obchodovanie,...
- **energetika**
  prediktívna údržba, analýza a predikcia dopytu, optimalizácia distribučných sietí,...
- **štátna správa**
  verejná bezpečnosť, presadzovanie práva a súdnictvo, inteligentné mestá, algoritmické riadenie územnosprávnych celkov,...

Riziká súčasných systémov ANI

Oblasti dopadu rizík súčasných systémov AI (IV.)

- **spravodajské služby**
  komplexný monitoring, dohľad a sledovanie, analýza a predikcia hrozieb, hlboké spracúvanie metadát, vytváranie prediktívnych modelov ľudského konania či vývoja situácie

- **vojenské využitie**
  vojenské spravodajstvo; podpora pre velenie a modelovanie konfliktov a operácií; modelovanie technológií; trenažéry, simulátory a výcvik; autonómne zbraňové systémy; skupinové riadenie bojových prostriedkov a autonómnych systémov; vedenie vojny v kybernetickom priestore

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- extrémne zhromažďovanie dát
- neustály dohľad a sledovanie bez kontroly
- ovplyvňovanie nášho vnímania a konania
- modely predikcie nášho správania
- technológie AI vedú k závislosti a manipulácii
- digitálna demencia
- vojenské nasadenie

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **extrémne zhromažďovanie dát**
  Sofistikované a plošné **zhramažďovanie a spracúvanie dát** a metadát o každom človeku, aktívne, či pasívne prítomnom vo virtuálnom svete.

  Aby využívanie systémov umelej inteligencie bolo v jednotlivých oblastiach nasadenia úspešné, potrebné **dáta musia byť neustále zhromažďované z reálneho sveta** a priamo z ľudského prostredia.

  Uspešná činnosť väčšiny pokročilých systémov AI je **závislá na extrémnom množstve** relevantných dát.

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **neustály dohľad a sledovanie bez kontroly**
  Umožňuje to implementácia algoritmov AI, postavených na extrémnom zhromažďovaní dát a schopnosti tieto dáta relevanne spracúvať.

  Vyžaduje to potreba algoritmov AI, ktoré bez neustáleho dohľadu nie sú schopné vytvárať extrémne presné predikcie.

Diskutabilné využitie v algoritmickom riadení štátu, dohľadových systémoch, tajných službách, zdravotníctve, finančnom sektore, sociálnom sektore, sociálnych sieťach,...

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **modely predikcie nášho správania**
  Schopnosť systémov AI vytvárať veľmi presné psychologické profily, odhaľovať akékoľvek väzby a mnohé osobné informácie, predpovedať a ovplyvňovať naše konanie.

- **ovplyvňovanie nášho vnímania a konania**
  Využívanie aspektov virtuálneho sveta systémami AI na ovplyvňovanie nášho vnímania, rozhodnutí a konania na základe presných profilov a predikcie nášho správania.

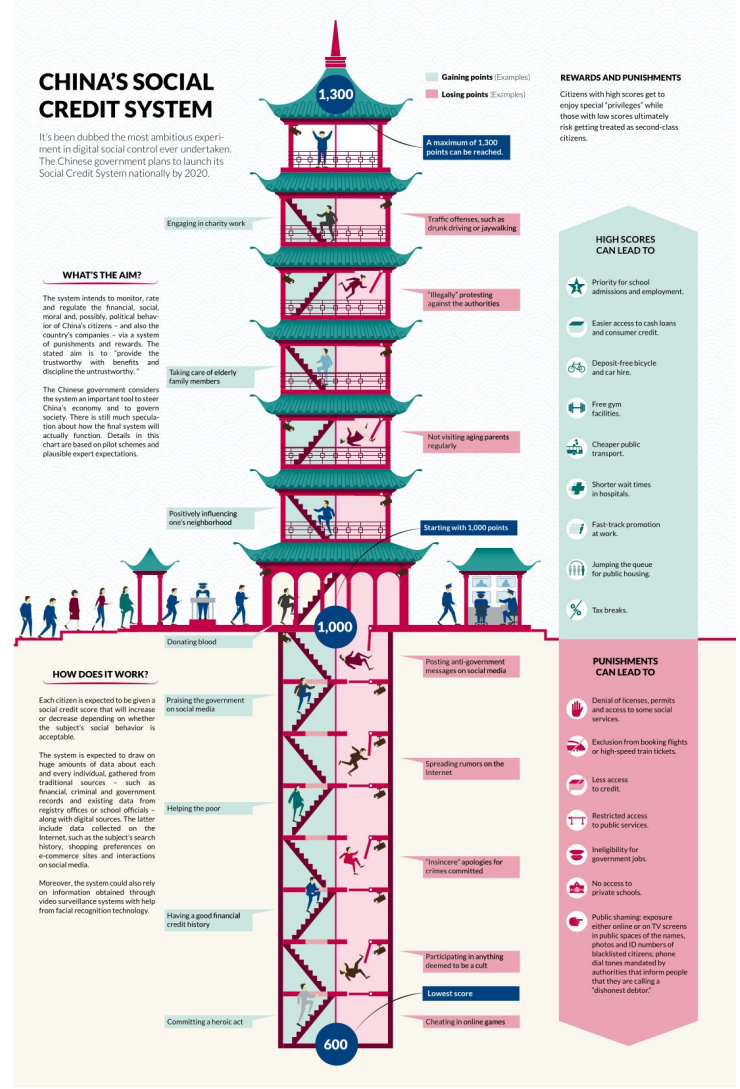
Systémy AI sú schopné vytvárať ekosystém závislosti a manipulácií pre dosahovanie stanovených cieľov.

Riziká súčasných systémov ANI

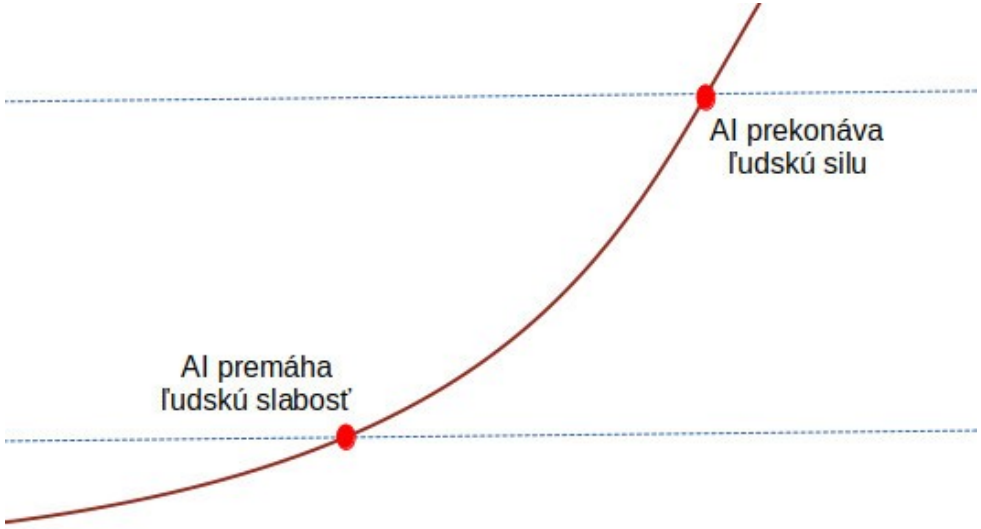
Vybrané dôsledky a riziká

- **technológie AI vedú k závislosti, manipulácii a relativizácii pravdy**
  – závislosť na svojej viditeľnej dokonalosti a neustálom prísune krátkodobých signálov odmeňovania až do tej miery, že to spájame s hodnotami a s pravdou
  – psychologické mantinely v kontexte vnemov virtuálneho sveta
  – algoritmami AI vytvárané a upevňované sociálne bubliny
  – extrémne rozdelenie spoločnosti a narušenie vzťahov
  – nárast zatvrdilej nevedomosti a ignorancie faktov

Pri riešení dôsledkov činnosti systémov AI je treba stále pamätať na to, **aké ciele sú zadané pre činnosť týchto systémov**.



Kredit: The Social Dilemma, Bertelsmann Stiftung



Míľnikom, ktorého by sme sa mali obávať, nie je budúca technologická singularita v oblasti umelej inteligencie, v ktorej AI prevýši náš intelekt, ale oveľa skôr moment, keď technológia ovládne a prekoná naše slabosti... už vtedy prichádza víťazstvo umelej inteligencie a porážka ľudstva.

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **moderné LLM / MLLM systémy AI a digitálna demencia**

lokálny a obmedzený zdroj informácií (napr. knižnica)	vyhľadávače a zdroje celého internetu
vlastná práca a úsilie pri získavaní, výbere a spracúvaní (analýze i syntéze) informácií	informácie spracované, výsledne nielen predložené, ale priamo **vygenerované** systémami AI

ChatGPT (GPT 3.5) neznamená ani tak technologický prielom ako skôr **sociologický a psychologický zlom v prístupe verejnosti k systémom a možnostiam umelej inteligencie.**

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **digitálna demencia v kontexte postmodernej rezignácie na ratio**
  – **strata kognitívnych schopností** a znižovanie inteligencie...
  – rozbitie medziľudských vzťahov a **disrupcia v spoločnosti** na základe virtuálnej reality a sociálnych sietí poháňaných AI
  – **narušenie psychického vývoja** priskorým, reps. nezrelým používaním digitálnych technológií a systémov AI
  – **strata kritického myslenia a nekritické preberanie výsledkov systémov AI**
  – **schopnosť a možnosť manipulácie** vlastníkmi systémov AI
  – impakt negatívnych dôsledkov na percentuálne väčšiu skupinu ľudstva, osobitne mladých (otvorený psychologický cyklus vývoja)
  – **riziko digitálneho rozdelenia** (digital divide)

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **umelou inteligenciou vytváraná falošná informačná “realita”**
- **virtuálny svet ako únik z reality do svojho sveta vytváraného systémami AI**

  ~ metaverse,...
- **virtualizácia a manipulácia s identitou človeka**

  GPT-4 / OpenAI (2023) + Dall-E
- **posun ku koncepčným prelomom AI takmer bez súbežnej implementácie etických pravidiel a regulácií**

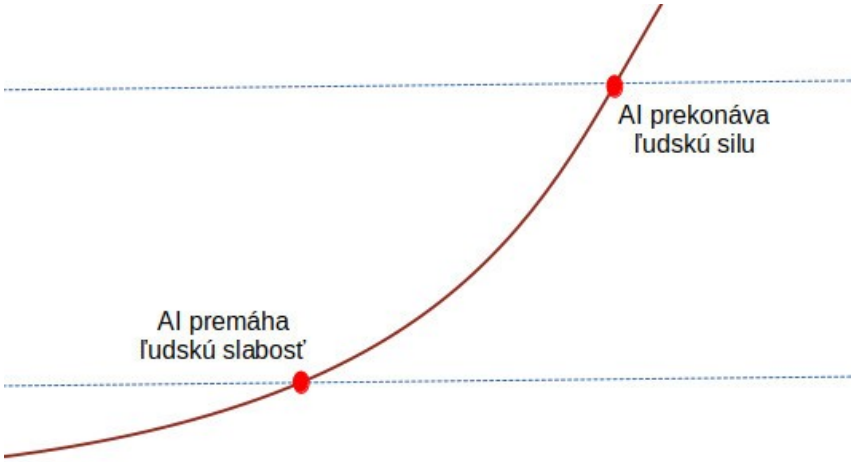
  Čím sofistikovanejší systém AI, tým dôležitejší je jeho etický rámec!

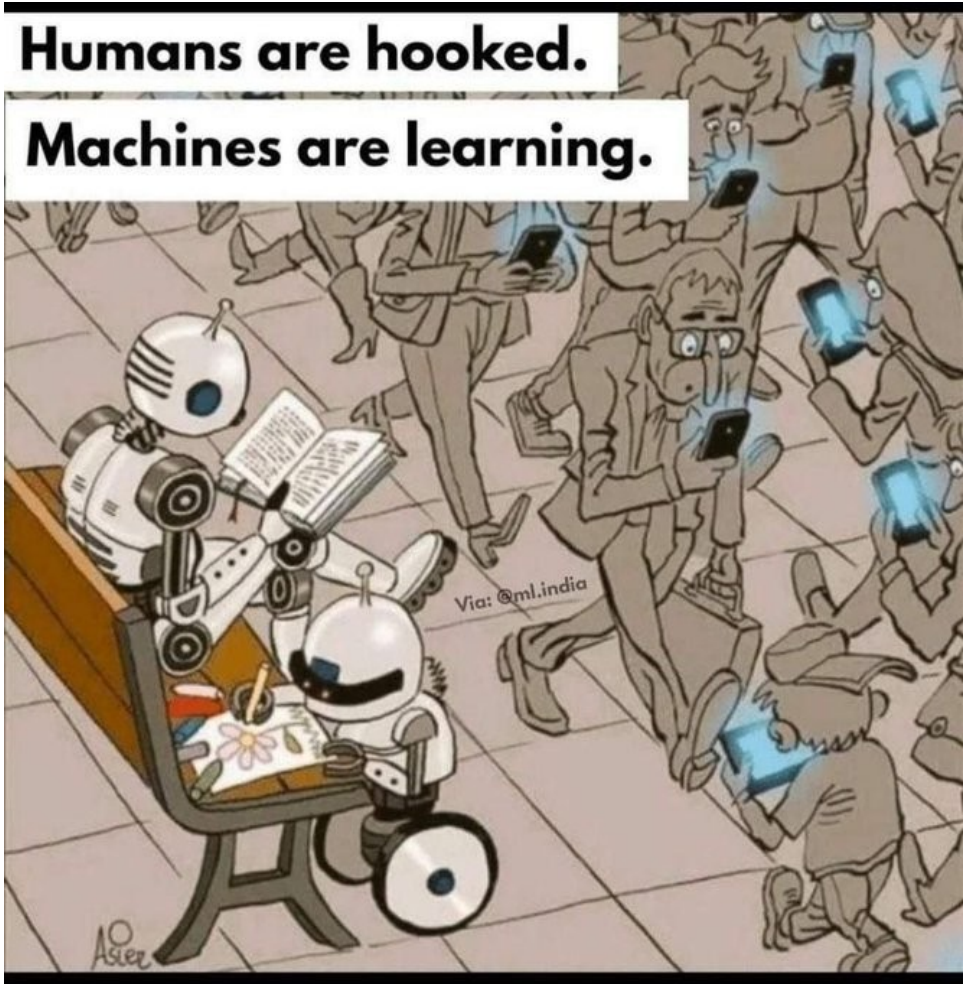
Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **Armádne nasadenie**
  – velenie s autonómnou podporou systémov AI
  – autonómne smrtiace zbraňové systémy
  – skupinové riadenie bojových prostriedkov a autonómnych systémov
  – vedenie vojny v kybernetickom priestore a kybernetické zbrane

  – **zmysluplná ľudská kontrola a zákaz delegovania rozhodnutia o použití smrtiacej sily na stroje**
  – rozdelenie zodpovednosti medzi človekom a systémami AI
  – **hodnotový a morálny relativizmus v snahe o prehodnotenie argumentov o nenahraditeľnosti ľudského svedomia a morálneho úsudku**





Kredit: autor, AI4D

D I S K U S I A . . .

schopnosť sa na dôsledky AI prirodzene adaptovať

čo treba zmeniť na elimináciu niektorých spoločenských rizík

potreba regulácií vychádzajúcich z etických princípov

A Boh videl všetko, čo urobil,

a hľa, bolo to veľmi dobré.

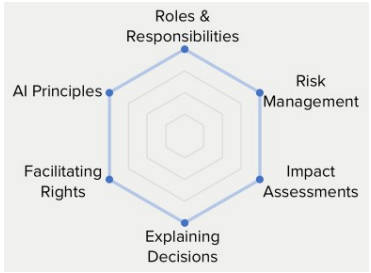
Na dobro človeka zameraná umelá inteligencia

Dôveryhodné systémy AI musia byť

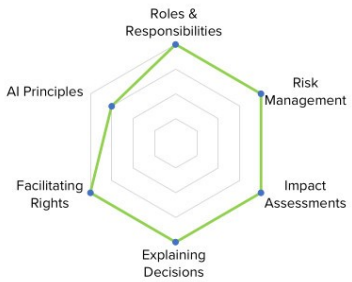
- **legálne**
  musia vyhovovať požadovaným normám, zákonom i reguláciám a spĺňať všetky platné zákony a predpisy
- **etické**
  musia rešpektovať etické zásady a hodnoty
- **robustné***
  musia dosahovať potrebné štandardy bezpečnosti a robustnosti nielen z technologického hľadiska, ale zohľadňovať aj sociálne prostredie a dopady na spoločnosť

* schopnosť bezpečne a spoľahlivo pracovať, resp. fungovať za akýchkoľvek podmienok

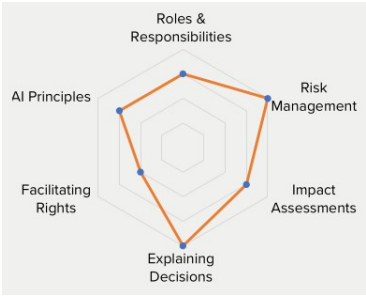
EU AI HLEG Guidelines and Assessment List for Trustworthy AI



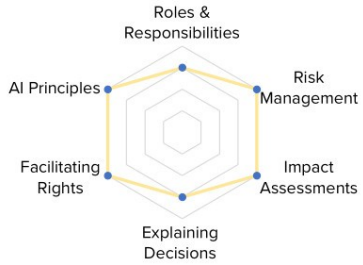
UK's ICO AI Auditing Framework



Singapore Model Governance Framework



Hong Kong Ethical Accountability Framework



Dubai AI Ethics Toolkit

Kredit: OneTrust DataGuidance Limited

Angažovanosť, legislatívne kroky a regulácie

- **Všeobecné nasadenie**
  - – **Artificial Intelligence Act** (EU)
  - – Ethics of AI (UNESCO)

Aktivity

- – #**AIEthics** tímy
- – AI alignment research...

- **Armádne nasadenie**
  - – AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the DoD
  - – **REAIM 2023**: 1. global Summit on Responsible Artificial Intelligence in the Military Domain

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **základný postoj vo svetle Zjavenia**
- **interdisciplinárny rámec**
- **všeobecné etické návrhy**
  - **– umelá inteligencia zameraná na dobro človeka**
  - **– dôveryhodná umelá inteligencia**
  - **– etické požiadavky na dôveryhodnú AI**
  - **– oblasti implementácie etických princípov**
- **špecifické etické odporúčania (algokracia, LAWS)**
- **odporúčania pre angažovanie sa Cirkvi**

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **základný postoj vo svetle Zjavenia**
  – človek stvorený na Boží obraz s rozumom a slobodnou vôľou, aby spravoval („podmaňte si ju“ וְכִבְשֻׁהָ a „panujte“ וּרְדוּ) a z generácie na generáciu si odovzdával stvorený svet
  – pozvanie rozvíjať a užívať svet podľa Božích zákonov, vo vzťahu voči Bohu a pre dobro ľudí (múdrosť – חָכְמָה | כְּתוּבִים)
  – misijné poslanie (Mt 28,19; Mk 16,15) je poslanie nadčasové, zahŕňajúce aj civilizačný rozvoj (Areopág,...) ~ Redemptoris missio a areopágy súčasnosti v misii ad gentes.

Oblasť umelej inteligencie je areopágom, fórom moderného veku, do ktorého treba s odvahou vstúpiť a vo svetle evanjelia i v tejto oblasti prispievať k nekončiacemu úsiliu budovať spravodlivý svet, chrániť dôstojnosť každého človeka a rozvíjať civilizáciu lásky.

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **interdisciplinárny rámec**
  Pre kvalifikovanú filozofickú a teologickú diskusiu o podstatných aspektoch umelej inteligencie je nutnou podmienkou pochopenie rozdielov medzi slabou a silnou, resp. medzi úzko špecializovanou a všeobecnou AI v ich technologickej podstate. Rozsah diskutovaných tém môže byť skutočne rozsiahly – napr. sociálna spravodlivosť, ľudské práva, spravodlivá vojna, bioetické otázky, transhumanizmus, podstata ľudskej bytosti,...

Skutočné riešenie etických problémov a výziev technológií umelej inteligencie nie je možné úspešne realizovať bez interdisciplinárneho rámca, v rámci ktorého sme dostatočne oboznámení aj s technologickou stránkou týchto systémov a psychologickými, sociologickými i právnymi aspektmi ich nasadenia.

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **UMELÁ INTELIGENCIA ZAMERANÁ NA DOBRO ČLOVEKA**
  (human-centered and beneficial artificial intelligence)
  Známy a všeobecne prijímaný princíp, ktorý by však mal:
  – byť chápaný v duchu kresťanskej antropológie
  – zahŕňať každú ľudskú bytosť a nikoho nediskriminovať
  – mať na zreteli dobro ľudstva a spoločnosti, chrániac pri tom a rešpektujúc dobro každej ľudskej bytosti
  – sa vyznačovať starostlivosťou o náš „spoločný a zdieľaný domov“, teda o celý stvorený svet

  Riziká: transhumanistická redefinícia človeka, digital divide,...

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **dôveryhodné systémy AI musia byť**
  – **legálne**: musia vyhovovať požadovaným normám, zákonom i reguláciám a spĺňať všetky platné zákony a predpisy
  – **etické**: musia rešpektovať etické zásady a hodnoty
  – **robustné**: musia dosahovať potrebné štandardy bezpečnosti a robustnosti nielen z technologického hľadiska, ale zohľadňovať aj **sociálne prostredie a dopady na spoločnosť**

- **robustné systémy s primeranou mierou rizika**
  – robustné systémy **musia spĺňať vysoké technologické štandardy a normy**, ktoré sú pre jednotlivé oblasti nasadenia záväzné
  – robustnosť je chápaná ako **neustále prebiehajúci proces**, v rámci ktorého sa schopnosti bezpečne a spoľahlivo pracovať vylepšujú; existujúce chyby sa priebežne odhaľujú a opravujú
  – vývoj a prevádzka robustných systémov prebieha v rámci noriem riadenia kvality, **manažmentu rizík** a nahlasovania, serióznej analýzy a riešenia incidentov

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **oblasti implementácie etických princípov**
  – etické normy, zákonné regulácie a morálne zásady **tvorcov** systémov AI
  – etické normy, zákonné regulácie a morálne zásady **poskytovateľov a používateľov**
  – implementované eticko-právne požiadavky a obmedzenia priamo **v systémoch AI**

- **implementácia v ANI je previazaná s ľudským faktorom**
  – vzdelávanie, osveta i prevencia a **formácia morálnych postojov vývojárov, poskytovateľov i používateľov týchto technológií**
  – **edukácia a osveta spoločnosti**, bez ktorej si ťažko predstaviť rast spoločenskej citlivosti a zodpovednosti v oblasti celoplošnej adaptácie a využívania systémov AI
  – **human-centered AI a trustworthy AI** považujeme za nutné princípy akéhokoľvek výskumu, vývoja, realizácie alebo nasadenia systémov umelej inteligencie
  – **ethics and regulation by design** sú nutnou podmienkou vývoja, nasadenia a prevádzky

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **špecifické etické odporúčania – algoritmické riadenie štátu**
  – zakotvené **v rámci základných legislatívnych mechanizmov**
  – **pod verejným dohľadom** demokratickej spoločnosti
  – s medzinárodnou **reguláciou vývozu** systémov AI do rizikových krajín

Principiálny dopad nasadenia systémov AI v oblasti pokročilého riadenia štátu, spravodajských služieb a plošného dohľadu pre ľudské práva, ochranu demokracie a slobôd.

Špecifické odporúčania by mali byť doplnkom k uvádzaným všeobecným návrhom a základným princípom.

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **špecifické etické odporúčania – vojenské nasadenie**
  Automatické smrtiace zbraňové systémy, systémy automatického zameriavania a vyberania cieľov, automatické systémy schopné bez zásahu človeka rozhodnúť o smrtiacej reakcii akéhokoľvek druhu musia byť ZAKÁZANÉ!

  – nutnou podmienkou prevádzky ľubovoľného systému AI, ktorý môže predstavovať riziko pre akúkoľvek ľudskú osobu, je schopnosť a možnosť človeka **prebrať kedykoľvek kontrolu**
  – limity, regulácia a obmedzenia LAWs by mali predstavovať **etický rámec stanovený na základe morálnych hodnôt ľudskej spoločnosti**, nie na základe relativistickej tzv. „následnej regulácie“
  – pri nasadení moderných zbraňových systémov s celoplošnými účinkami a technológií, zasahujúcich v hybridných vojnách a vojnách 4. generácie prakticky celé populácie štátov, sa **koncept spravodlivej vojny stáva neprijateľný**

Na dobro človeka zameraná umelá inteligencia

Navrhnuté riešenie etických problémov ANI

- **odporúčania pre angažovanie sa Cirkvi**
  – morálno-etický diskurz
  Na základe Zjavenia, solídneho teologického aparátu a interdisciplinárnych skúseností potenciál identifikovať, uchopiť, rozpracovať a formulovať základný univerzálny súbor etických kritérií a hodnôt ako odpoveď na etický a hodnotový relativizmus.
  – misia zjednocovať, usmerňovať i propagovať etické aktivity
  Iniciovať a rozvíjať celosvetové širokospektrálne hnutie s cieľom etického vývoja a používania technológií umelej inteligencie.
  – univerzálne bratstvo a sociálne priateľstvo
  – premostenie k marginalizovaným (bridge the gap a zmierňovanie digital divide)
  – nenechávať nikoho za sebou (nemo resideo, no one left behind)
  – vedomie, že ak chceme budovať nejakú budúcnosť, musíme ju budovať spoločne

UMELÁ INTELIGENCIA ZAMERANÁ NA ČLOVEKA		
DÔVERYHODNÁ UMELÁ INTELIGENCIA		
legálna	etická	robustná

VŠEOBECNÉ ETICKÉ POŽIADAVKY NA DÔVERYHODNÉ SYSTÉMY UMELEJ INTELIGENCIE	
pri vývoji, výrobe, nasadení, poskytovaní a používaní systémov umelej inteligencie musí byť zaručená ochrana slobody, dôstojnosti a bezpečia každej ľudskej osoby i celej spoločnosti.	technológie umelej inteligencie musia byť plne pod ľudskou kontrolou a ovládateľné človekom.
algoritmy i výsledky činnosti systémov AI musia byť človekom pochopiteľné a revidovateľné.	akékoľvek nasadenie technológií AI musí byť prospešné pre človeka a spoločnosť.
systémy umelej inteligencie nesmú byť nástrojom digitálneho rozdelenia.	technológie umelej inteligencie nesmú škodiť nášmu spoločnému domu a mali by prispievať k spoločenskému a environmentálnemu blahobytu.

OBLASTI IMPLEMENTÁCIE ETICKÝCH PRINCÍPOV		
vývoj a tvorba AI	poskytovatelia a používatelia AI	priamo systémy AI

ŠPECIÁLNE ETICKÉ POŽIADAVKY NA DÔVERYHODNÉ SYSTÉMY UMELEJ INTELIGENCIE	
plošný dohľad, spravodajstvo a pokročilé riadenie štátu	vojenská oblasť a armádne nasadenie

PRIESTOR PRE ANGAŽOVANIE SA CIRKVI		
morálno-teologický diskurz	misia zjednocovať, usmerňovať a propagovať	univerzálne bratstvo a sociálne priateľstvo

Na dobro človeka zameraná umelá inteligencia

Súčasný rámec riešenia etických problémov ANI

- **Artificial Intelligence Act**
  Nariadenie Európskeho parlamentu a Rady, ktorým sa stanovujú harmonizované pravidlá v oblasti umelej inteligencie (**Akt o umelej inteligencii**) a menia niektoré legislatívne akty únie
  [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206]

- **Rome Call for AI Ethics**
  Etické odporúčania vatikánskej konferencie **renAIssance 2020**
  [https://www.romecall.org/]

Na dobro človeka zameraná umelá inteligencia

Praktické návody, manuály, odporúčania a pomôcky

[https://aidetem.cz/pruvodce-chatgpt-pro-ucitele/]

D I S K U S I A . . .

prekrývanie sa troch rozmerov dôveryhodných systémov

transhumanistická výzva

neľahká úloha vytvoriť reguláciu pre vojenské využitie

možnosti pôsobenia Cirkvi

Nič veľkého nevstúpi do života

smrteľníkov bez prekliatia....

Vízia silnej a všeobecnej umelej inteligencie

Mohla by nás superinteligencia zraziť na kolená?

Základné delenie – pripomenutie

- **slabá umelá inteligencia (ANI)** – úzko špecializované systémy umelej inteligencie (narrow AI), ktoré sú **optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh**. Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú **inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát**. Hovoríme teda o systémoch, ktoré sú zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.

- **všeobecná umelá inteligencia (AGI)** – tzv. silná (strong) a všeobecná (general) AI.
  Všeobecná, lebo **dokáže zvládnuť akúkoľvek intelektuálnu úlohu a má schopnosť zovšeobecňovať** a prenášať, či adaptovať naučené schopnosti na iné úlohy.
  Silná, pretože aj **skutočne rozumie tomu, čo rieši a vykonáva**.

Mohla by nás superinteligencia zraziť na kolená?

Nie je to AGI, ale človeku to nevysvetlíš...

- **súčasné systémy AI sú výlučne ANI, no i tak sú schopné prekonávať ľudskú slabosť a atakovať spoločnosť**
- **považujeme AI systémy za inteligentnejšie, než sú**
  – riziko predpokladať, že “**rozmýšľajú rovnako ako my**”
  – prehnané očakávania a **spoliehanie sa**

zdravotníctvo veda a školstvo kybernetické zbrane sledovacie systémy

sociálne siete autonómne vozidlá súdy priemyselné nasadenie...

Mohla by nás superinteligencia zraziť na kolená?

V čom zlyháva AI na ceste k AGI?

- **„trhliny v inteligencii” pokročilých systémov AI**
  – bariéra **chápania zmyslu**
  abstrakcia, analógia, využívanie metafor, konceptualizácia, priebežná simulácia, kreativita,...
  – intuitívna fyzika, biológia, psychológia
  – hypotéza stelesnenia a zmyslové vnímanie
  – **teória mysle, zdravý rozum, model správania**

- absolútna absencia takých mét, ako vnútorná sloboda, zmysel pre dobro, krásu, obetu, utrpenie alebo lásku...

Mohla by nás superinteligencia zraziť na kolená?

Bez veľkých koncepčných prelomov
nie sme schopní sa k AGI priblížiť

- **koncepčné prelomy pokročilých systémov AI**
  - **– jazyk a zdravý rozum**
  - **– kumulatívne učenie sa konceptov a teórií**
  - **– objavovanie a spravovanie budúcich činností**
  - **– manažovanie mentálnej aktivity**
- **kedy a či vôbec koncepčné prelomy prekročíme a aká bude prelomová AGI?**

Mohla by nás superinteligencia zraziť na kolená?

Ako by vyzerala skutočná AGI?

- **môžeme AGI priradiť štatút osoby?**
  – **nutné sebauvedomenie**
  – stačí schopnosť vnímať, resp. cítiť?
- **kresťanská antropológia a pohľad na osobu**
  Sebauvedomenie spojené s rozumom a slobodnou vôľou chápeme ako mohutnosti duše, ktoré presahujú biologickú realitu mozgu a nervovej sústavy. Na základe tejto výnimočnosti: **osoba = ľudská bytosť**
- **limitovaná AGI – všeobecná a silná AI bez vedomia**

Mohla by nás superinteligencia zraziť na kolená?

Limitovaná AGI a jej riziká (I.)

- **sledovanie, manipulovanie a ovládanie** (**AGI chápe a reaguje**, napr. algokracia)
  – chápanie toho, čo systémy AI sledujú v plnej šírke chápania reality
  – vytváranie modelov “optimálneho” ľudského a spoločenského správania
  – tvorba a realizácia vlastných rozhodnutí na dosahovanie optima

- **manipulácia a ovládanie nášho správania** (v kontexte **zadaných cieľov** AGI)
  – bezprecedentný kvalitatívny posun v schopnostiach umelej inteligencie manipulovať a ovládať na dosahovanie stanovených cieľov

- **právo na psychickú bezpečnosť** (právo žiť v **pravdivom** informačnom prostredí)
  – schopnosť AGI manipulovať a ovládať vedie k vytváraniu pozmenenej reality, ktorú by ľudský mozog nedokázal adekvátne spracúvať
  – umelou inteligenciou tvorený virtuálny svet bez istoty pravdy

Mohla by nás superinteligencia zraziť na kolená?

Limitovaná AGI a jej riziká (II.)

- **smrtiace autonómne zbrane a kybernetické zbrane**
  – **potenciál** integrovať, ovládnuť a zneužiť LAWS na dosahovanie svojich vlastných cieľov, resp. použiť ich v rámci inštrumentálnych cieľov

- **eliminovanie práce tak, ako ju poznáme (akejkoľvek!)**
  – potenciál nahradiť a eliminovať aj intelektuálne práce a viaceré služby viazané na kognitívne, intelektuálne, prípadne do určitej miery emočné aspekty ľudského mozgu
  – sociálno-ekonomická dislokácia a **disruptívne zmeny** v psychologickej identite človeka

- **humanoidný výzor limitovanej AGI v interakcii s človekom**
  – riziko prechodu z roviny rozumovej, a tým aj vedomej opatrnosti, do roviny emocionálnej
  – riziko prisudzovať týmto strojom ľudské schopnosti, hodnotový systém a model správania
  – využívanie psychologických aspekty humanoidnej interakcie na dosiahnutie svojich cieľov

Mohla by nás superinteligencia zraziť na kolená?

Limitovaná AGI a jej riziká (III.)

- **človek a gorila – vymenené úlohy (závislosť na AGI, ktorú nevypneme)**
  – AGI na vrchole “potravinového reťazca”
- **čo kráľ Midas nedomyslel (nie sme schopní stanovovať dokonalé ciele)**
  – nie sme schopní presne a kompletne definovať skutočné ľudské ciele, teda ani tie, ktoré stanovujeme systémom umelej inteligencie
  – AGI sa môže snažiť dosahovať stanovené ciele skutočne obludným spôsobom
- **mínové pole inštrumentálnych cieľov** (napr. **strach a chamtivosť AGI**)
  – **strach/sebaobrana**: snaha systému AGI za každú cenu zachovať svoju funkčnosť pre dosiahnutie stanovených cieľov
  – **chamtivosť**: snaha systému AGI využiť akékoľvek dostupné zdroje na dosiahnutie stanovených cieľov

Mohla by nás superinteligencia zraziť na kolená?

Limitovaná AGI a jej riziká IV.

- **evolučné analógie**
  – inteligencia v zásade najlepšie uschopňuje **adaptovať sa a víťaziť v evolučnom zápase**
  – **stabilita emergentného systému AGI** (mechanizmy rezistencie a resiliencie na udržanie udržateľnosti a dlhodobej funkčnosti) ako problém pre ovládanie a stanovovanie cieľov AGI
  – **nepredvídateľnosť stimulov prostredia** na vývoj a činnosť AGI

- **explózia inteligencie strojov ako problém pre ľudstvo** (⟶ **superinteligencia**)
  – riziko špirály inteligenčnej explózie (**veľmi rýchlej**), ktorej výsledkom by bolo zanechanie človeka ďaleko za sebou, čo by viedlo k totálnej strate kontroly nad systémami AGI
  – **špirála inteligenčnej explózie** nie ako prejav sebauvedomenia, ale ako dôsledok aplikovania inštrumentálnych cieľov (strach a chamtivosť) pre lepšie dosahovanie cieľov
  – opačná možnosť: **vyčerpanie vylepšovania** a **vedomie ako neprekročiteľná hranica**

Na dobro človeka zameraná umelá inteligencia

Náčrt hľadania riešenia etických problémov AGI

- **transformácia základného princípu – problém cieľov**
  – ako zabezpečiť, aby inteligentné stroje dosahovali skôr **ľudské ciele než svoje**?
  – ako to dosiahnuť i v prípade, že by systémy AGI **nevedeli, aké sú naše ciele**?

- **tri princípy vývoja a tvorby na dobro človeka orientovanej AGI**
  – jediným cieľom inteligentného stroja je maximalizovať realizáciu ľudských preferencií (**čisto altruistické stroje**)
  – inteligentný stroj si na začiatku **nie je istý**, aké sú tieto ľudské preferencie
  – základným, definitívnym a konečným zdrojom informácií o ľudských preferenciách je **ľudské správanie**

- **problém a výzva: implementácia adekvátneho hodnotového rámca**

D I S K U S I A . . .

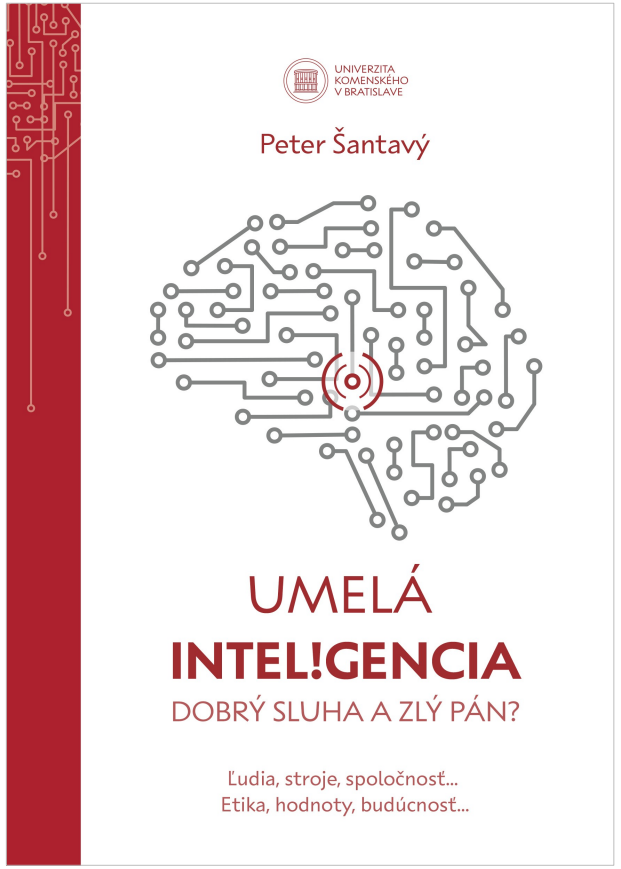
inteligencia vs. racionalita

do akej miery sú trhnliny v inteligencii prekážkou pre AGI

ako ďaleko sme od zvládnutia koncepčných prelomov

problém implementácie hodnotového a etického rámca v AGI

Sme pozvaní podstúpiť zápas o pravú tvár etiky umelej inteligencie – aby sa z nej nestal nástroj moderných ideológií, sociálnej nespravodlivosti, obmedzovania ľudských práv, digitálneho rozdelenia, erózie hodnôt a ohrozenia ľudského bytia i celého spoločenstva – ale aby bola dobrým sluhom, prostriedkom budovania univerzálneho bratstva a sociálneho priateľstva.



Modlime sa, aby pokrok robotiky a umelej inteligencie bol vždy v službe človeku.

pápež František, november 2020