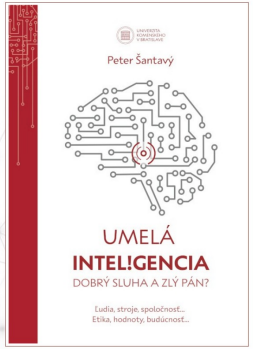
PREDNÁŠKA SPOJENÁ S DISKUSIOU NA TÉMU

MÔŽE NÁM UMELÁ INTELIGENCIA PRERÁSŤ CEZ HLAVU?

... technológie umelej inteligencie ...

... riziká súčasných systémov ...

... superinteligencia ako výzva ...



ThLic. Ing. Peter Šantavý, PhD.

Umelá inteligencia je ako Colombova žena,

nikto ju nevidel a všetci o nej hovoria...

Technológie umelej inteligencie (AI)

Buzzword 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)

Klimatické zmeny – dôležitý posun vo vývoji AI

- **striedanie ročných období – (AI spring vs. AI winter)**
  striedanie aktívneho rozvoja a útlmu v oblasti umelej inteligencie
- **klimatická zmena začína** – 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)

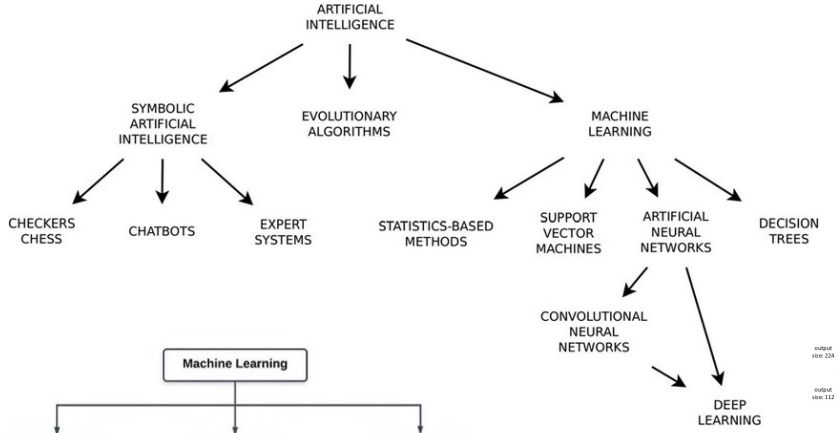
Základné vlastnosti

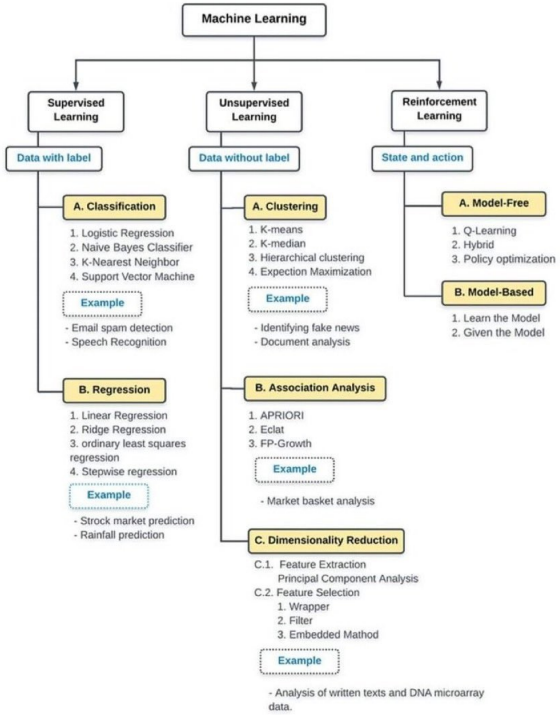
- **autonómnosť – schopnosť samostatne konať**
  Schopnosť systému vykonávať úlohy v komplexnom prostredí bez neustáleho vedenia používateľom.
- **adaptabilita – schopnosť sa prispôsobovať**
  Schopnosť zlepšovať svoj výkon (a schopnosti) učením sa (nielen) zo skúseností.

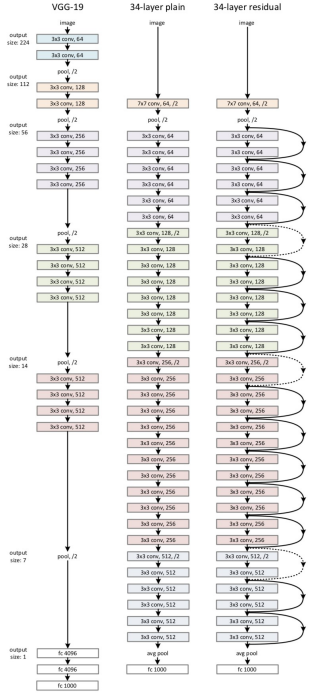
Technológie umelej inteligencie (AI)

Základné delenie

- **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**.

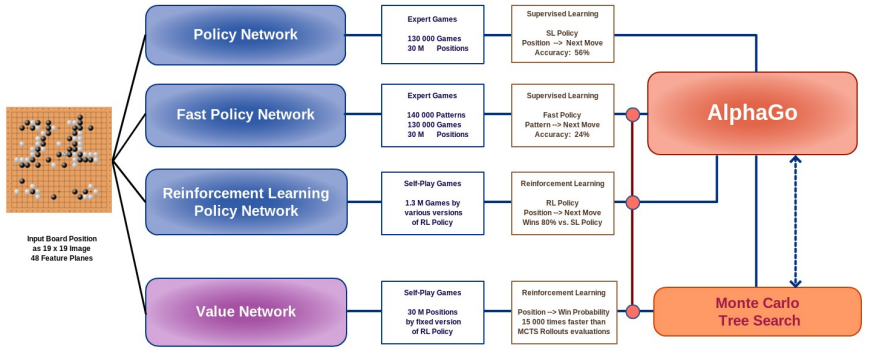


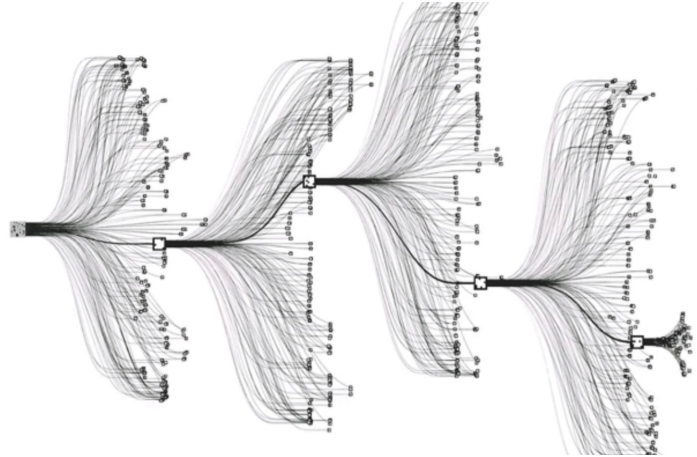




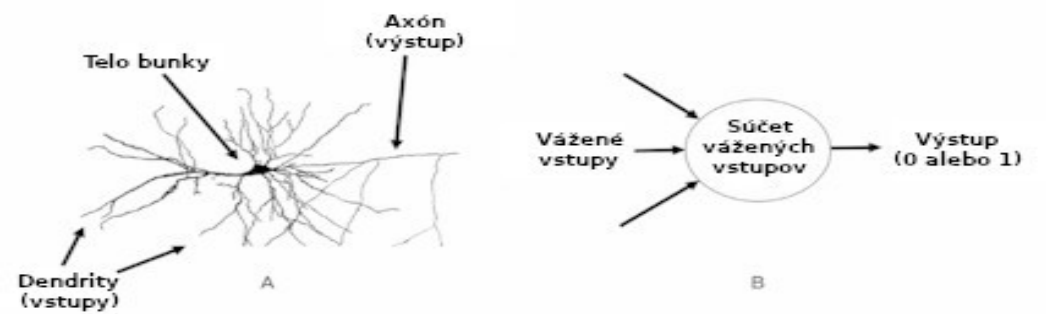
AlphaGo Overview

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

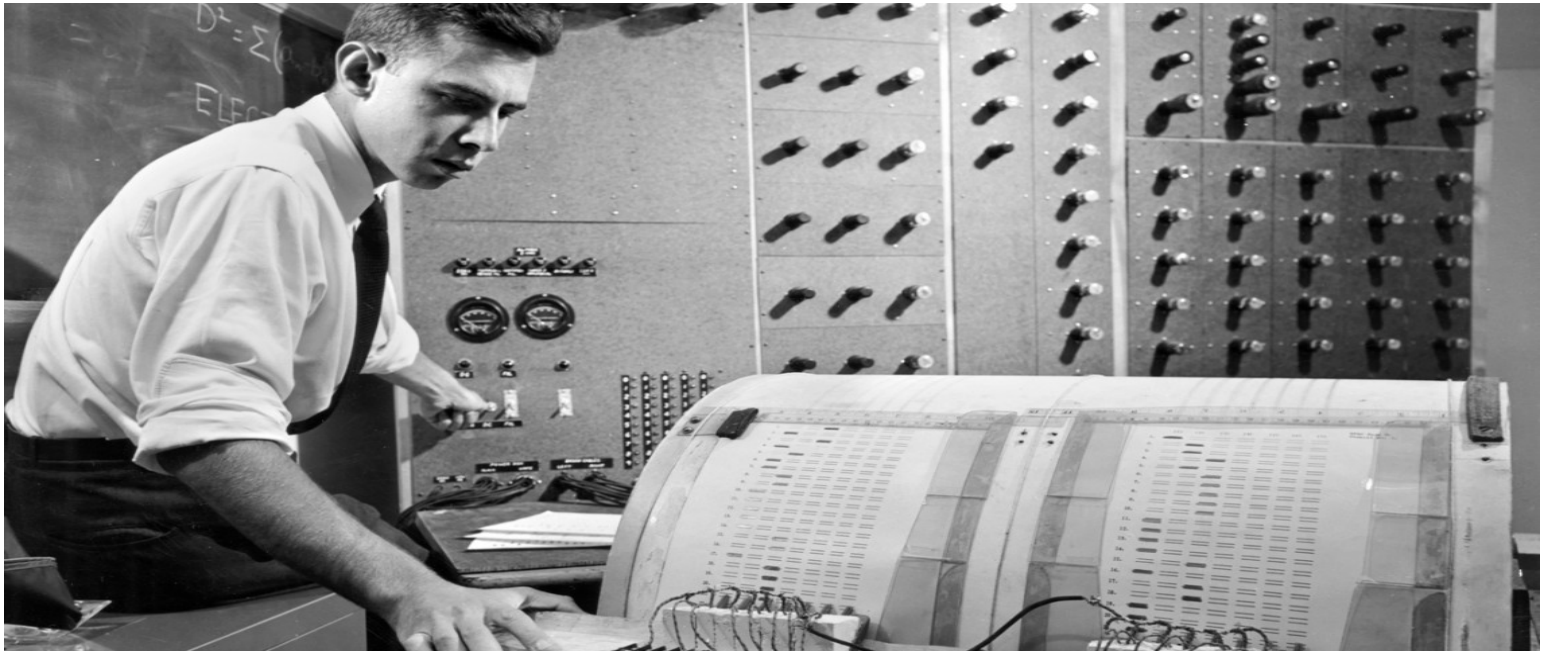




Kredit: Fabio Galbusera, Eric Gaubert, Bob van den Hoek, Seth Weidman

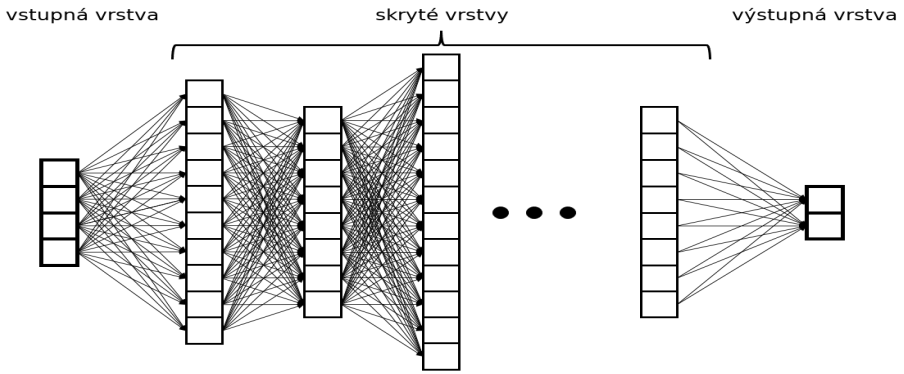


1957...

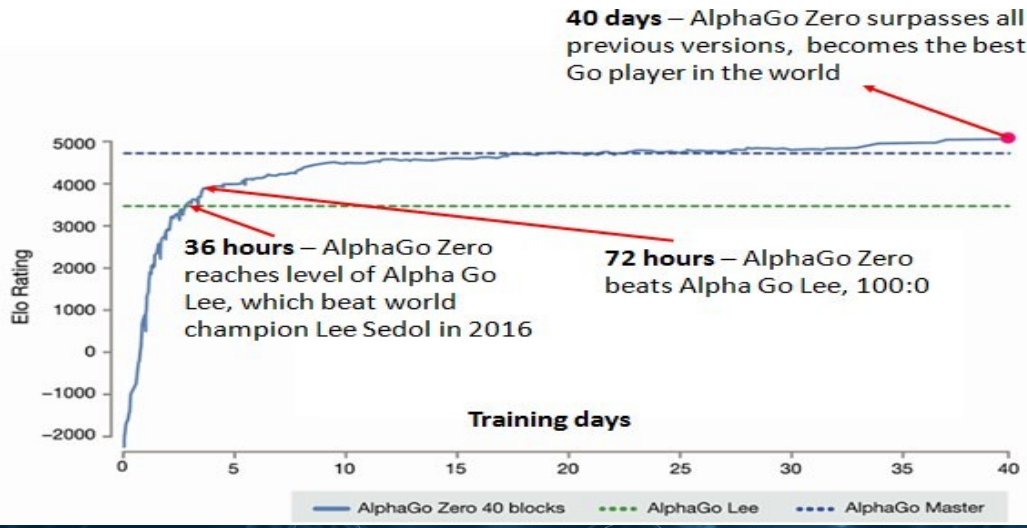


Perceptron vyvinutý psychológom Frankom Rosenblattom

Kredit: Ars Technica (arstechnica.com)



...súčasnosť



Kredit:

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

Každá dostatočne pokročilá technológia je

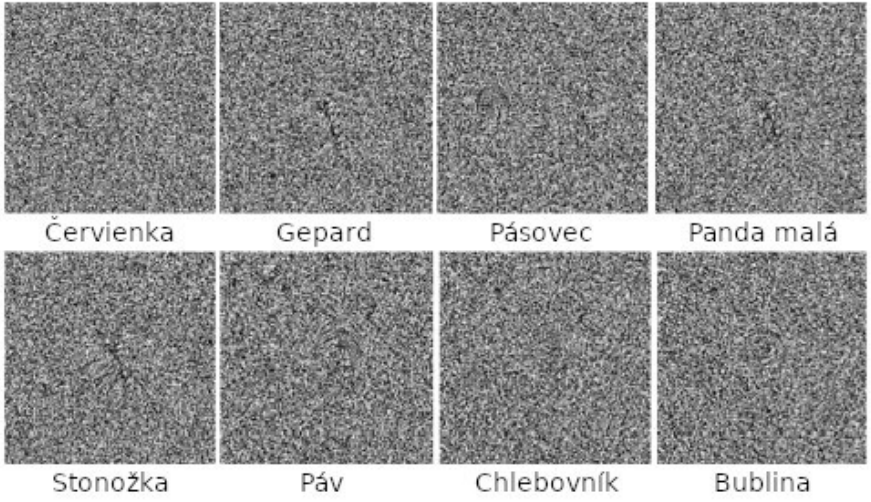
na nerozoznanie od mágie...

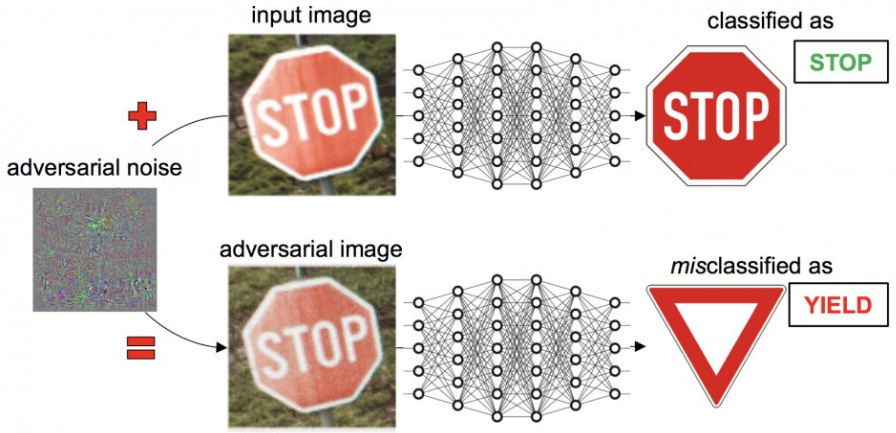
Riziká súčasných systémov ANI

Rizikové faktory I.

- **technologické problémy – limity a riziká AI**
  training dataset biases overfitting to training data
  fooling deep neural networks and vulnerability to hacking
  long-tail effect superstition

- **procesné riziká**
  útoky na dôvernosť (confidentiality attacks)
  útoky na zraniteľnosti (evasion attacks)
  útoky s cieľom ovplyvniť model (poisoning attacks)



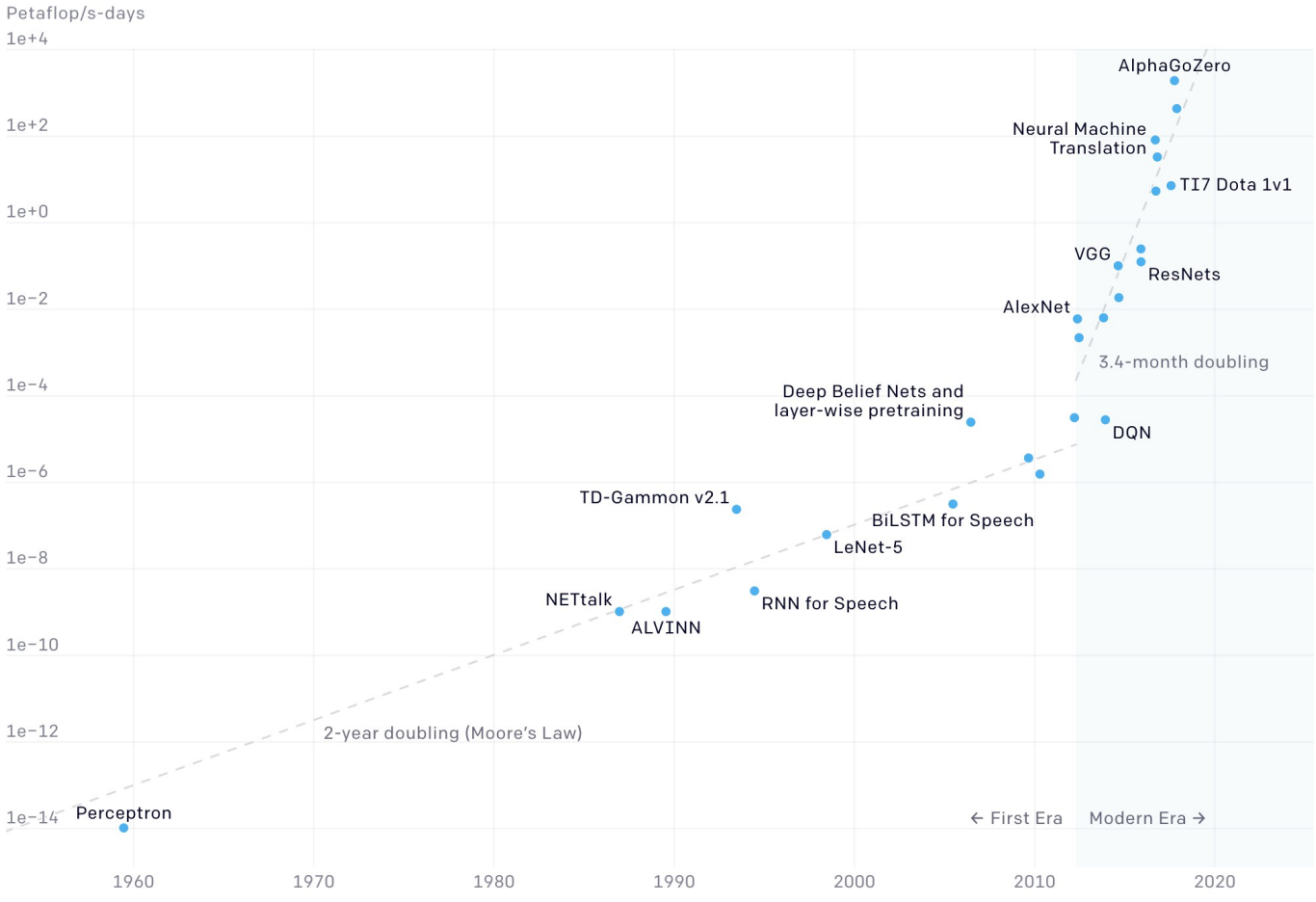


Kredit: Melanie Mitchell – upravené autorom, K. Xu et al., Sourav Agarwal

Riziká súčasných systémov ANI

Rizikové faktory II.

- **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

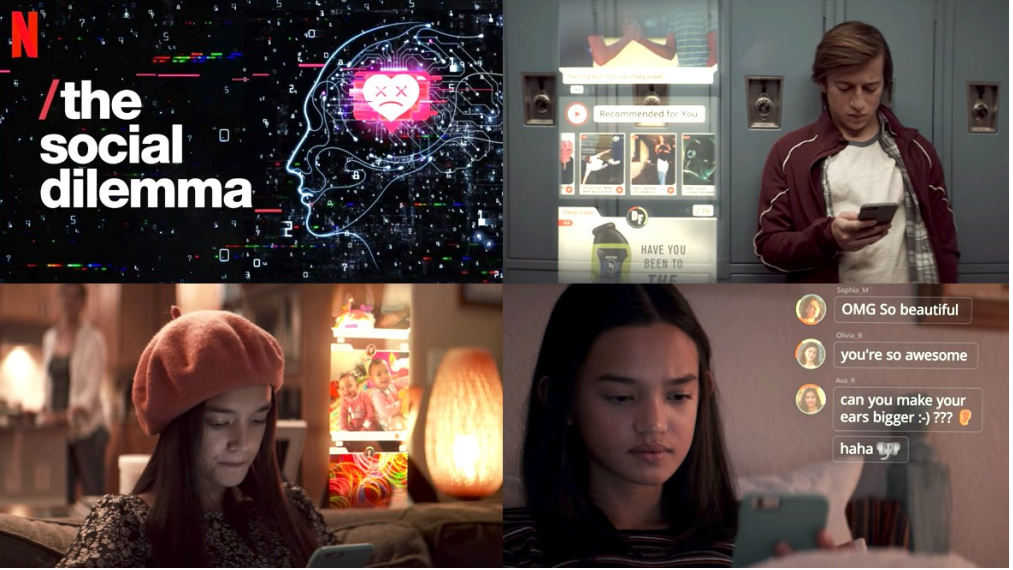


Kredit: D. AMODEI, D. HERNANDEZ

Riziká súčasných systémov ANI

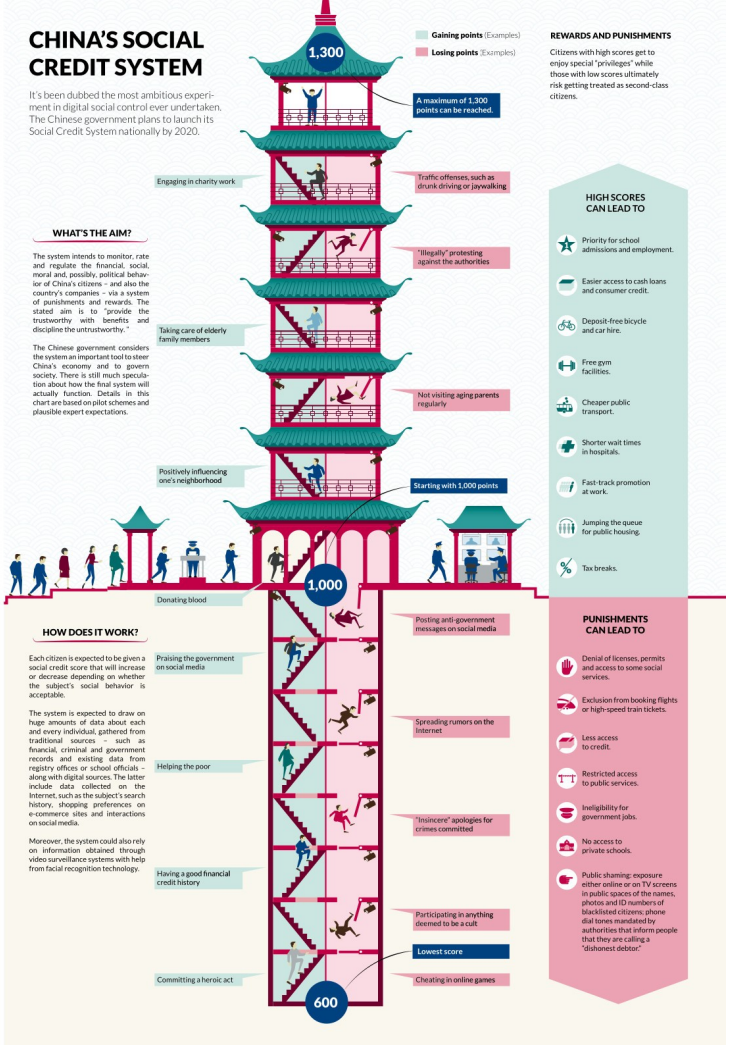
Oblasti dopadu rizík súčasných systémov AI

- spoločnosť – sociálne siete a paradigmatické zmeny
- doprava
- zdravotníctvo
- ďalšie služby (banky, poisťovne, obchody,...)
- spravodajské služby a algoritmické riadenie štátu
- vojenské využitie

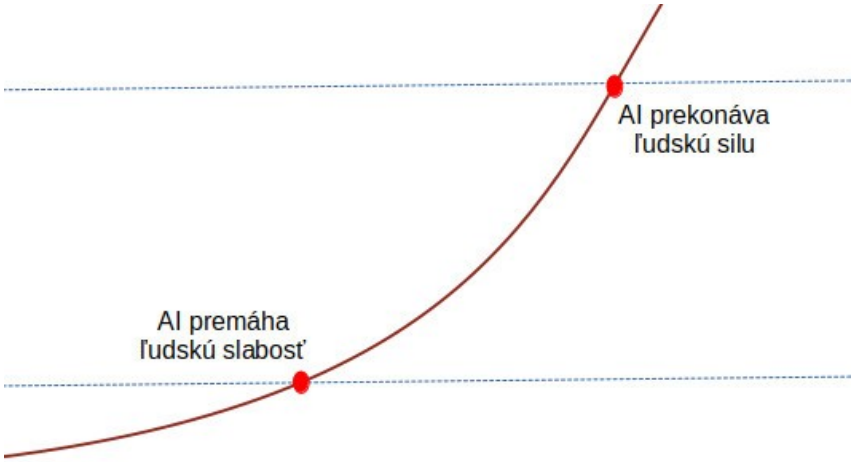


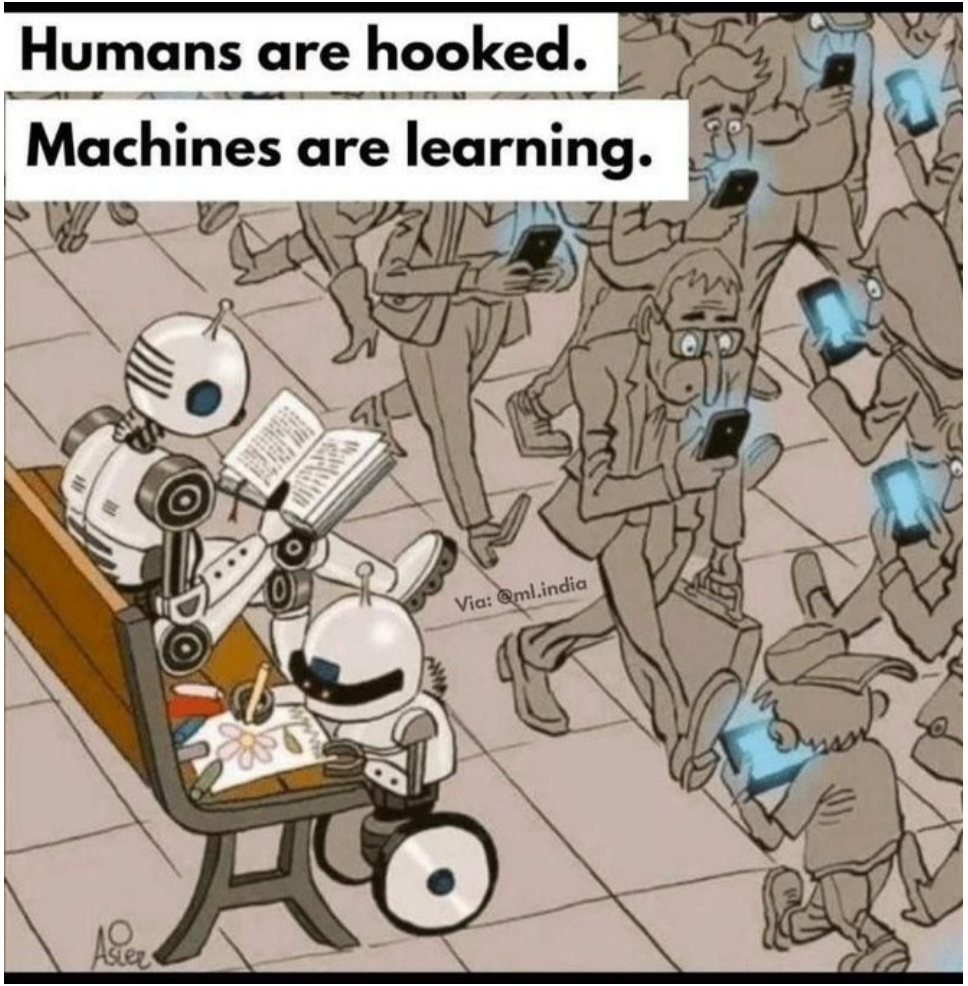
Vybrané dôsledky a riziká

- **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**

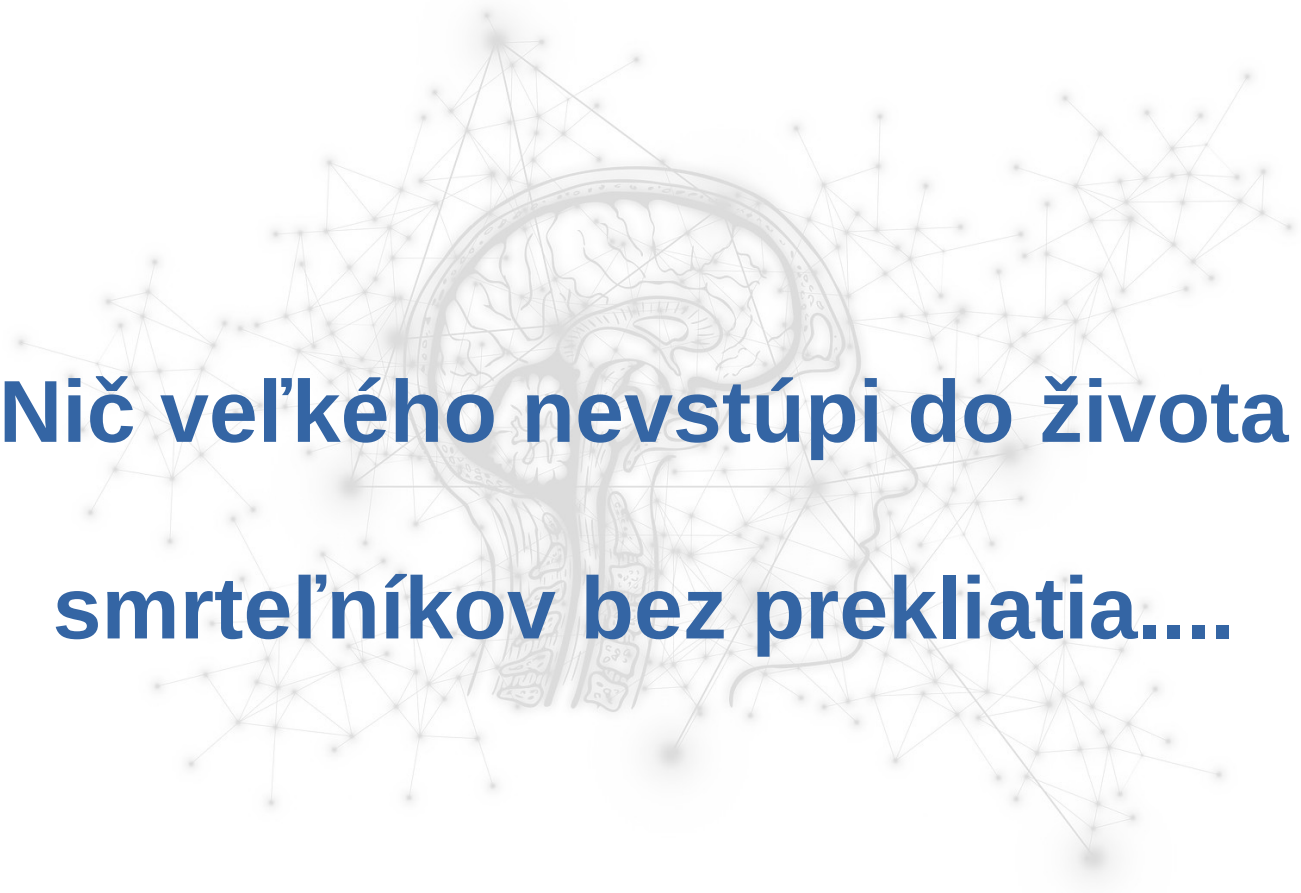


Kredit: The Social Dilemma, Bertelsmann Stiftung





Kredit: autor, AI4D



Nič veľkého nevstúpi do života

smrteľníkov bez prekliatia....

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 súdy kybernetické zbrane sledovacie systémy

sociálne siete autonómne vozidlá veda 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

– sledovanie, manipulovanie a ovládanie (AGI chápe a reaguje, napr. algokracia)
– manipulácia a ovládanie nášho správania (v kontexte zadaných cieľov AGI)
– právo na psychickú bezpečnosť (virtuálny svet tvorený AGI bez istoty pravdy)
– smrtiace autonómne zbrane
– eliminovanie práce tak, ako ju poznáme (akejkoľvek!)
– humanoidný výzor limitovanej AGI v interakcii s človekom (manipulácia)
– človek a gorila – vymenené úlohy (závislosť na AGI, ktorú nevypneme)
– čo kráľ Midas nedomyslel (nie sme schopní stanovovať dokonalé ciele)
– mínové pole inštrumentálnych cieľov (napr. strach a chamtivosť AGI)
– evolučné analógie (evol. výhoda, stabilita emergentného s., stimuly vývoja AGI)
– explózia inteligencie strojov ako problém pre ľudstvo (⟶ superinteligencia)

A Boh videl všetko, čo urobil,

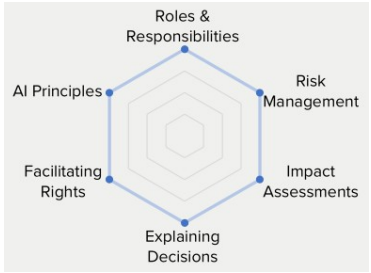
a hľa, bolo to veľmi dobré.

Na dobro človeka zameraná umelá inteligencia

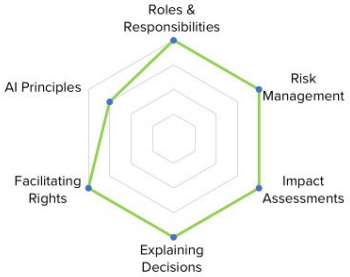
Dôveryhodné systémy AI musia byť

- **legálne**
  musia vyhovovať požadovaným normám, zákonom a reguláciám
- **etické**
  musia spĺňať požadované etické kritériá
- **bezpečné**
  musia dosahovať potrebné štandardy bezpečnosti a robustnosti

EU AI HLEG Guidelines and Assessment List for Trustworthy AI

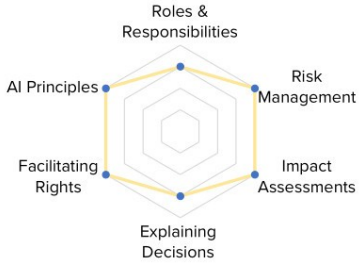


UK's ICO AI Auditing Framework

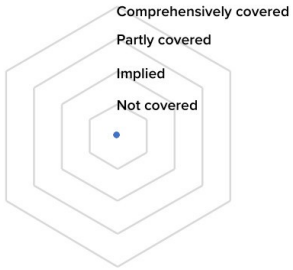


Singapore Model Governance Framework

Hong Kong Ethical Accountability Framework



Dubai AI Ethics Toolkit



Kredit: OneTrust DataGuidance Limited

Na dobro človeka zameraná umelá inteligencia

Naše riešenia 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**

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

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**

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

pápež František, november 2020

DISKUSIA...