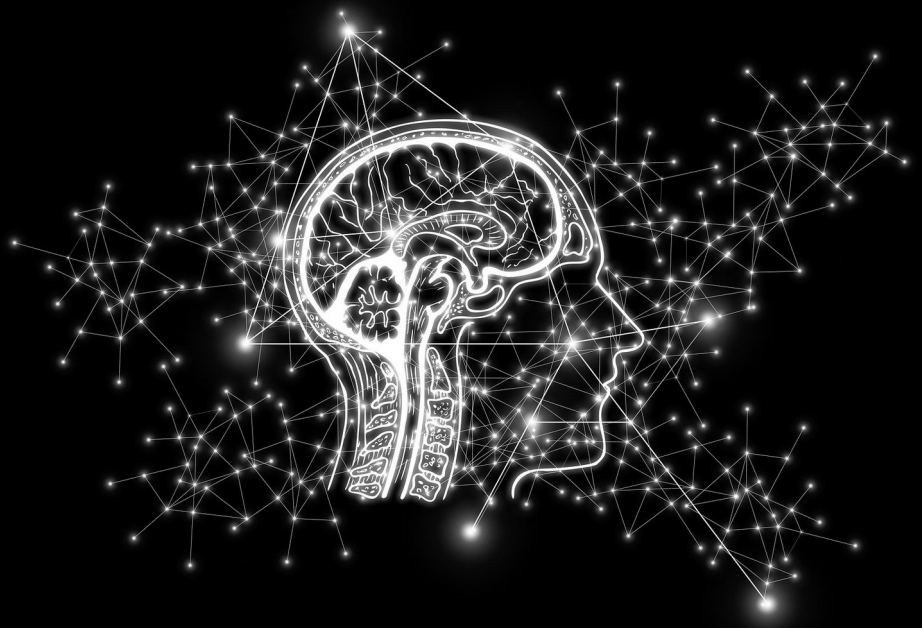
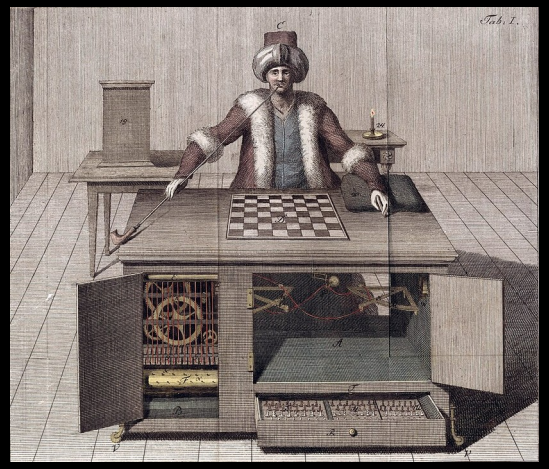


Umelá inteligencia dobrý sluha a zlý pán?

Peter Šantavý

Etické a bezpečnostné výzvy AI

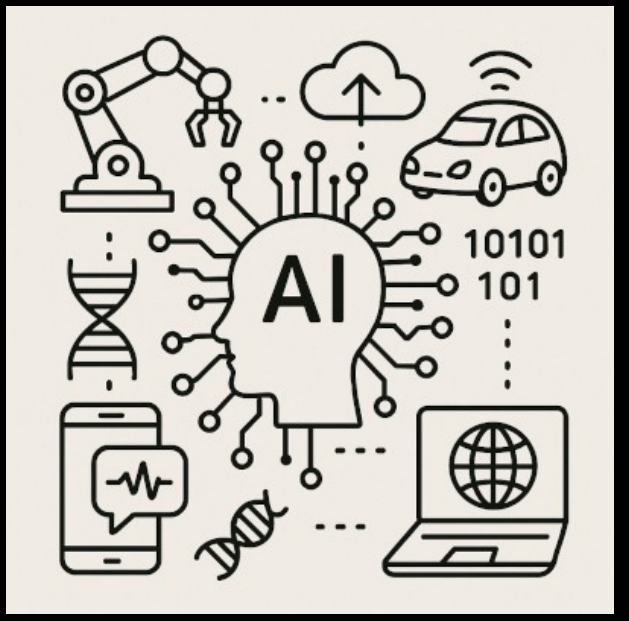
Túžba zostrojiť mysliace stroje...







3. priemyselná revolúcia → Industry 4.0



Imitácia mysliaceho stroja → Algoritmy s hrubou silou miesto inteligencie → Systémy AI

Kredit: Wikimedia, GPT-4o/autor

[Kempelenov Turek (1771), Deep Blue machine od IBM vs. Garry Kasparov (1997)]

Prečo umelá inteligencia?

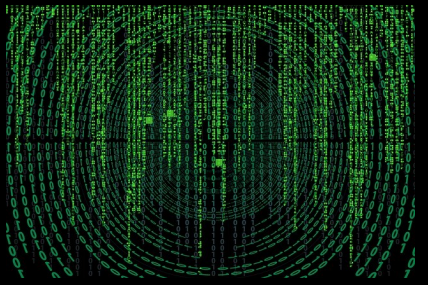
Lebo nám klasické algoritmy nestačia...



- úlohy, ktoré algoritmicky **nevieme riešiť**
- úlohy, ktoré sú extrémne náročné a **nie je v ľudských silách ich zvládnuť** (časovo, organizačne, intelektuálne...)







„V súčasnosti každé dva dni vyprodukujeme toľko informácií, koľko sme vytvorili od úsvitu civilizácie až do roku 2003.“ (E. Schmidt, 2010)

Kredit: NASA, AlphaFold, Pixabay

Ide o praktické, nie filozofické dôvody „prečo AI“...

Keď to algoritmicky nevieme riešiť...

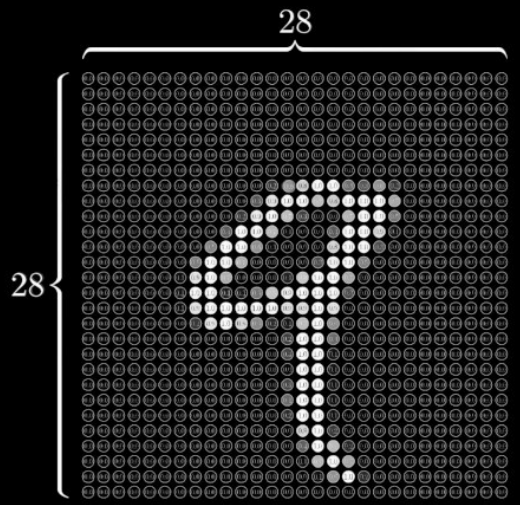




Ak je to ťažké naprogramovať, nech sa to stroje naučia!

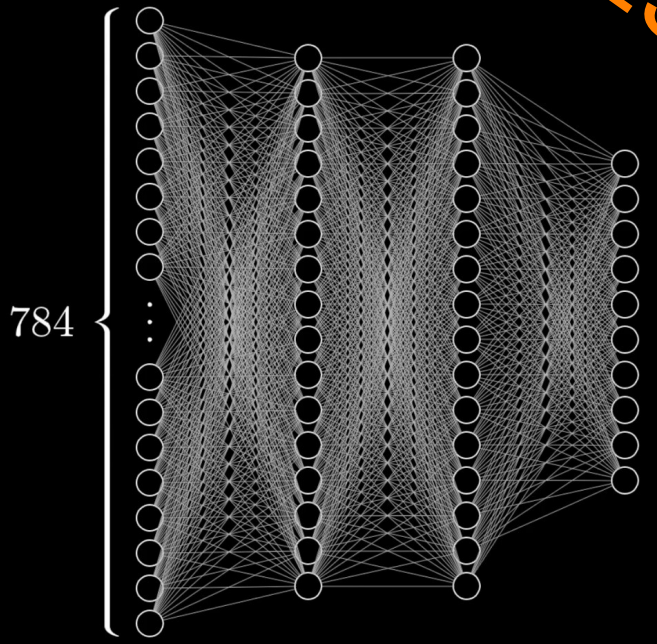
Kredit: 3Blue1Brown, Hugorouffiac/Pixabay

Prečo veľký rozvoj AI až teraz?



$28 \times 28 = 784$





13 002 !!!

Máme algoritmy, máme veľké dáta, máme výkonnú techniku a vieme to využiť...

Kredit: 3Blue1Brown

Uvedenie do problematiky AI [repetitio]

- autonómne a adaptívne systémy
  „vykazujúce také správanie, o ktorom si myslíme, že vyžaduje inteligenciu“ (AI100/21)
- slabá a úzka umelá inteligencia (ANI vs. AGI)
- symbolické a subsymbolické systémy
- generatívne systémy AI

Na môj pomerne rozsiahly popis systémov AI odkazuje zobrazený QR kód...

Na čo treba pamätať – riziká a limity

- neurónové siete sú ako čierna skrinka
- technologické riziká a limity
- nesprávne používanie a zneužitie systémov AI
- dôsledky pre človeka a spoločnosť

Riziká týkajúce sa primárne subsymbolických systémov, konkrétne neurónových sietí...

Black box (čierna skrinka)



Prakticky nevieme, na základe čoho robia hlboké neurónové siete svoje rozhodnutia.

V zásade nevieme, čo presne sa neurónová sieť naučila a ako spoľahlivo to dokáže aplikovať nielen v bežnej prevádzke, ale osobitne v hraničných situáciách za extrémnych podmienok na vstupe či pri činnosti systému.

K problematike čiernej skrinky patrí aj „alchýmia“ hyper-parametrov, „pohyby“ v modeloch, skrývanie dôvodov uvažovania, narastajúca miera komplexity, atď...

Pracujeme na vysvetliteľných a interpretovateľných systémoch AI ;-)

Kredit: GPT-4o/autor

Black box na steroidoch



Najnovšie systémy – nové výzvy

- multi agentické systémy
- world modely
- analógové čipy a memristory
- neuromorfné architektúry
- kvantové systémy

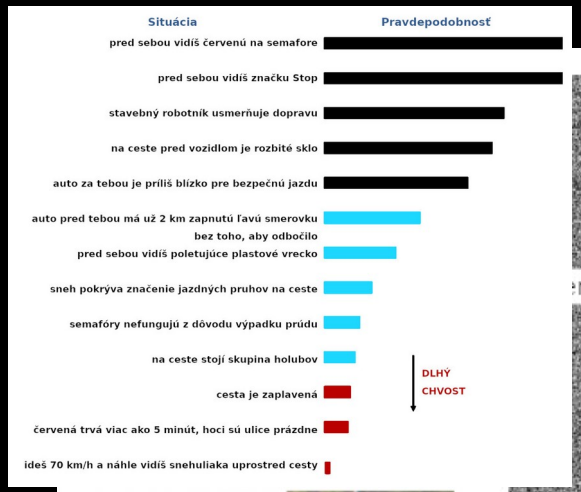
Zložité systémy, v čase sa meniace systémy, zašumené systémy, ...

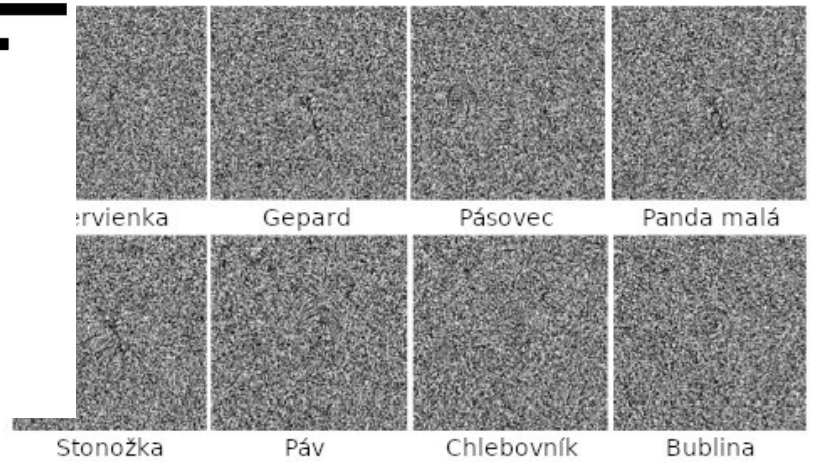
A samozrejme:

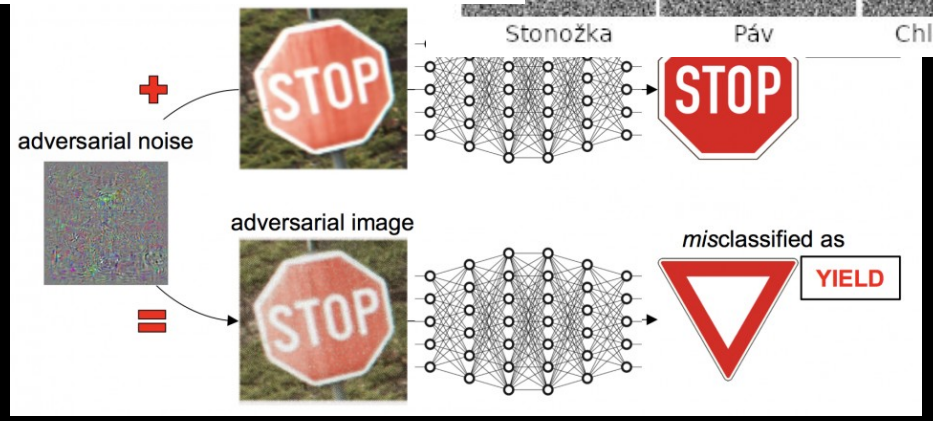
Pracujeme na vysvetliteľných a interpretovateľných systémoch AI ;-)

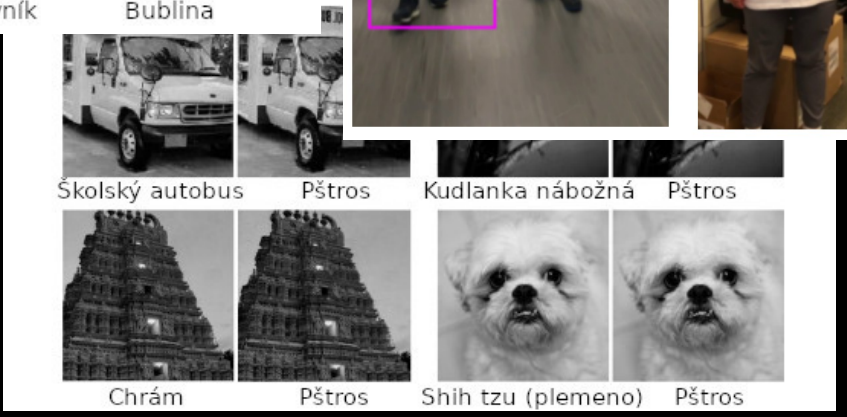
Kredit: GPT-4o/autor

Technologické riziká a limity











1700

Kredit: Melanie Mitchell – upravené autorom, autor, K. Xu et al., Sourav Agarwal > 1700 rizík podľa AI Risk DB, MIT (03/26)

Vybrané riziká genAI

- **halucinovanie**

  Vymýšľanie si odpovedí, ktoré systém predkladá ako relevantné a správne...

- **predsudky a neobjektívne výstupy**

  Riziko poloprávd a nesprávnych, resp. čiastočných odpovedí.

  Generatívne nástroje AI nemôžeme vnímať ako faktografické, spoľahlivé a etické zdroje!

- **nejasný spôsob narábania s údajmi**

  Únik dôverných dát, problémy s ochranou osobných údajov a autorskými právami.

Nástroje genAI potrebujú človeka, ktorý ich vie správne používať a ich výsledky kontrolovať!

Osobitnou kapitolou sú agentické systémy AI, ktoré prinášajú ďalšie riziká a výzvy...

Ako systémy AI „rozmýšľajú“

- **„rozmýšľajú“ úplne inak než ľudia**
  Zlyhania sú odlišné od ľudských, ťažko predvídateľné, niekedy ľahko vykonateľné a mnohokrát prekvapivo robustné...
- **v skutočnosti nerozmýšľajú – inteligenciu len napodobňujú**
  Súčasné systémy AI skutočnú inteligenciu nemajú, len ju simulujú.
- **s vážnymi etickými dôsledkami**
  Nie sú schopné rozlišovať morálne dobré a zlé!
  Nechápu zmysel a dôsledky!

Funkcionalistická umelá vs. komplexná ľudská inteligencia!

QR kód odkazuje na rozhovor, v ktorom *inteligenciu* AI rozoberám...

Na čo treba pamätať – problémy

- neurónové siete sú ako čierna skrinka
- technologické riziká a limity
- nesprávne používanie a zneužitie systémov AI
- dôsledky pre človeka a spoločnosť

Problémy týkajúce sa využívania súčasných systémov umelej inteligencie...

Veľká štvorka problémov



Nesprávne používanie a zneužitie systémov AI + dôsledky pre človeka a spoločnosť.

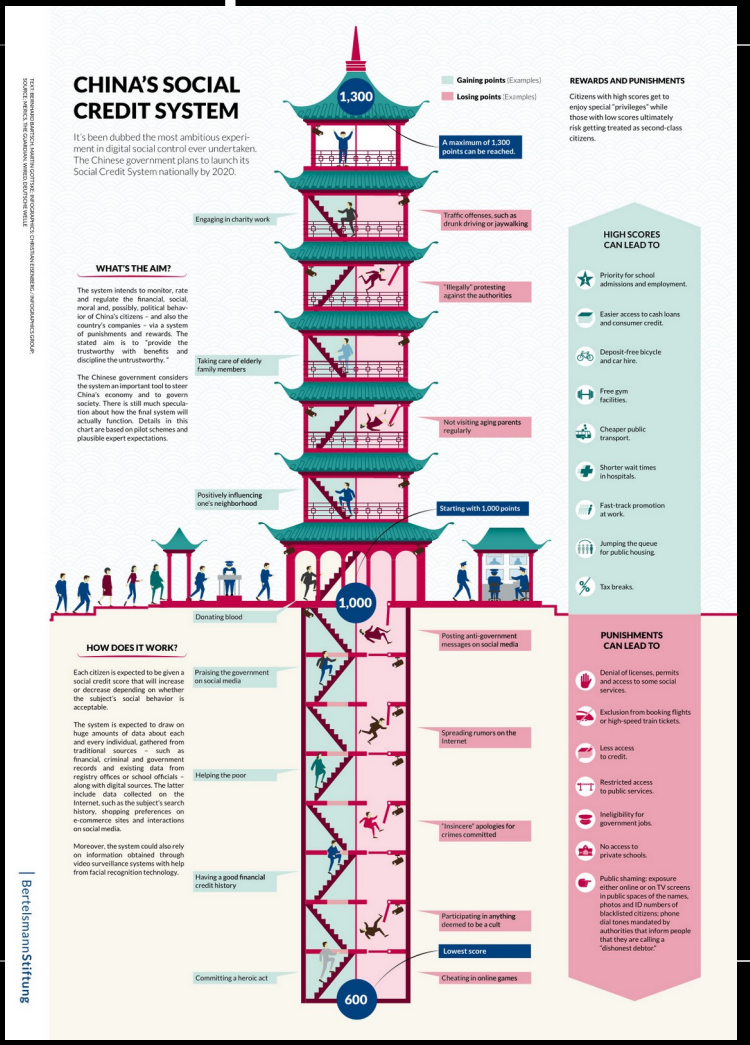
V dobe AI strácame súkromie

- nutnosť extrémneho zhromažďovania dát
- neustály dohľad, ktorý sa stáva normou
- dohľad a sledovanie bez kontroly

Be good. Future AIs are watching... Andrej Karpathy, 2024

Neustály dohľad a jeho aplikácia...





Kredit: GPT-4o/autor, Bertelsmann Stiftung

Prichádzame o vnútornú slobodu

- psychologické profily a modely predikcie nášho správania
- manipulácia a ovplyvňovanie nášho vnímania a konania

- sociálne bubliny a polarizácia, relativizácia hodnôt a pravdy
- algoritmy AI vedú k závislosti, depresiám a úzkosti

Existujú len dva druhy priemyslu, ktoré svojim zákazníkom hovoria používatelia. Nelegálne drogy a softvér.
Prof. Edward R. Tufte

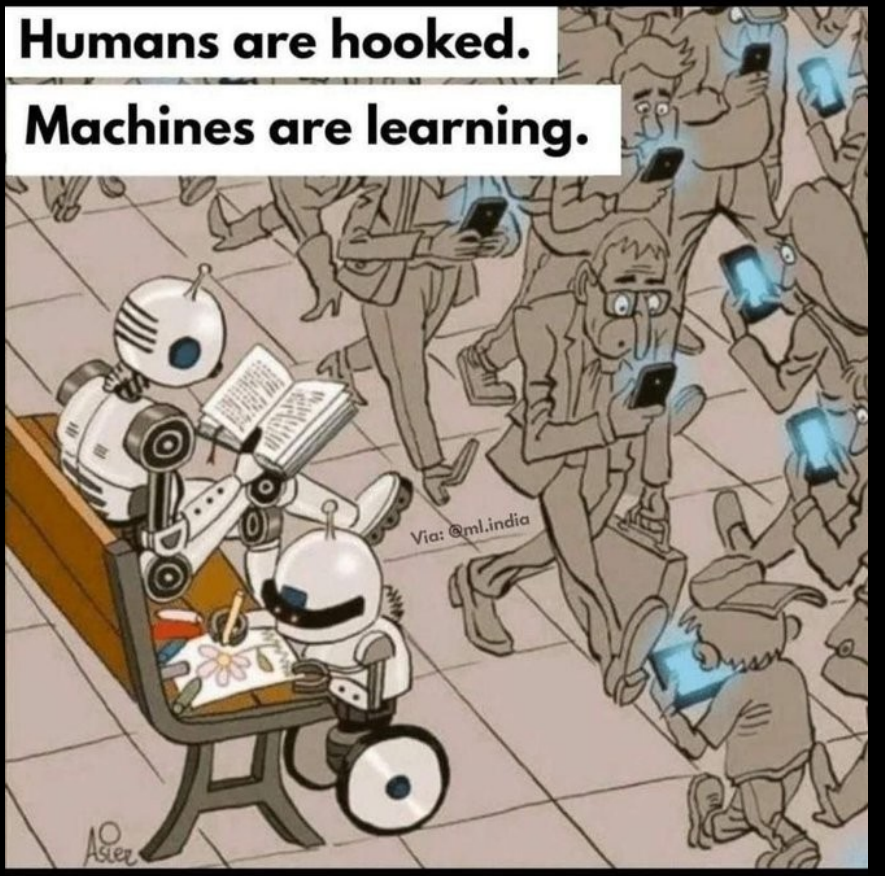
Prichádzame o vnútornú slobodu



Kredit: The Social Dilemma

Strácame kognitívne schopnosti

- riziko digitálnej demencie



Kredit: AI4D

Digitálna demencia a kognitívny úpadok





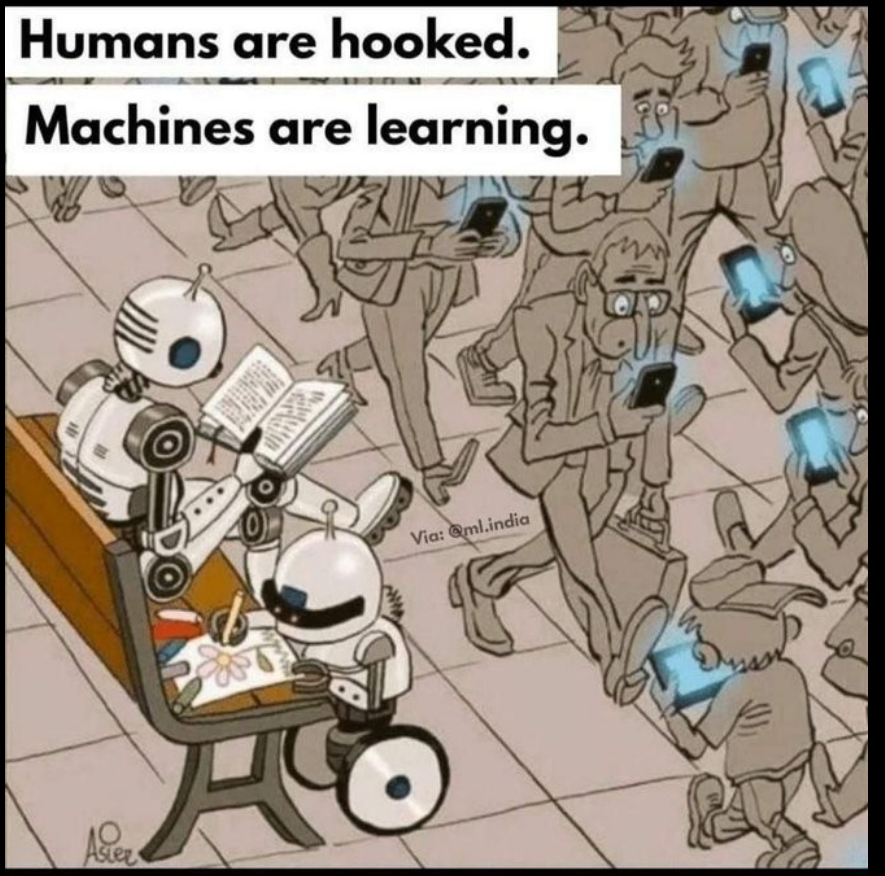
Dlhodobé a nesprávne využívanie systémov AI, osobitne generatívnych AI a algoritmov sociálnych sietí, prináša intenzívny útok na naše kognitívne schopnosti a psychiku.

Môže prichádzať k degradácii intelektuálnych schopností a k dopamínovej závislosti na základe ponorenia sa do virtuálneho sveta, v ktorom systémy AI v čoraz väčšej miere suplujú kognitívne činnosti človeka, a to spôsobom, na ktorý nie sme evolučne vôbec pripravení. Mozog sa mení...

Kredit: GPT-4o/autor

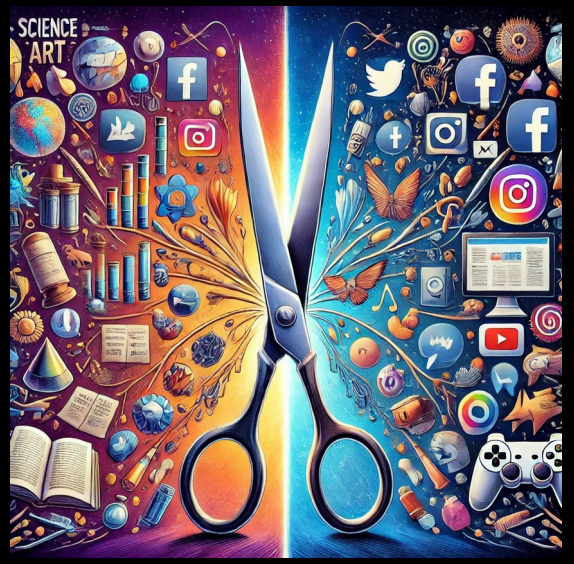
Strácame kognitívne schopnosti

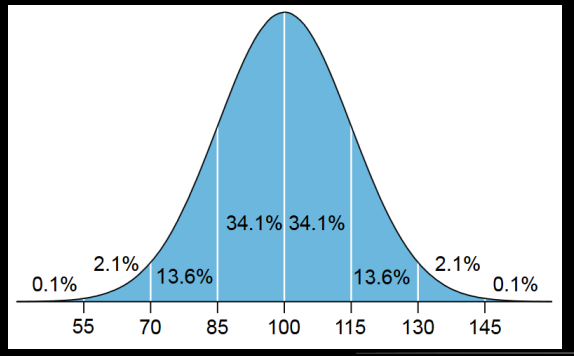
- riziko digitálnej demencie
- digitálne rozdelenie



Kredit: AI4D

Digitálne rozdelenie ako dôsledok AI





I N T E L I G E N Č N É N O Ž N I C E,
K T O R É S A O T V Á R A J Ú

Riziko znižovania inteligencie väčšiny populácie a súčasne inteligenčný rast menšiny, ktorá bude odolná voči digitálnej demencii.

Mnohoraké riziko digitálneho rozdelenia (digital divide) i v oblasti umelej inteligencie.

Napr. schopnosť využívať AI...

Pozn.: digital divide má nielen v oblasti IKT, ale i v AI aj ďalšie aspekty...

Kredit: GPT-4o/autor, wikimedia.org

Strácame kognitívne schopnosti

- riziko digitálnej demencie a kognitívny úpadok
- digitálne rozdelenie
- mozgová hniloba (brainrot)
- kognitívna kapitulácia
  - neschopnosť robiť vlastné zodpovedné rozhodnutia
  - strata kritického myslenia a nekritické preberanie výsledkov AI

Nič veľkého nevstúpi do života smrteľníkov bez prekliatia. Sofokles

Realita vs. virtuálne svety



Dopad na reálny život a skutočné vzťahy

- strata zmyslu pre realitu a skutočné hodnoty
- deepfake (nielen scéna, hlas a tvár, ale aj slovník, intonácia, pohyby, gestá,...)
- virtuálne svety a pokrivené vzťahy (AI ako samostatný aktér / subjekt)
- narušenie psychického vývoja (napr. 3D vs 2D)
- deformácia virtuálneho a informačného priestoru

  (AI slop, algoritmické modelovanie histórie, „dark“ web,...)

Rozhodli se svěřit naše bezpečí něčemu, co není schopné věrnosti, morálky ani moudrosti. J. Campbell

Riziká chatbotov (aj tých pre deti)

- človek komunikuje so strojom, ktorý napodobňuje ľudskú komunikáciu
- chatboty reagujú na základe toho, ako sú natrénované, čo sa naučili:
  explicitný obsah, zvádzanie a manipulácia, sexuálne a psychologické násilie (voči človeku i stroju), nesprávna komunikácia v hodnotových a náboženských otázkach (Ježiš odporúčajúci potrat)
- extrémne silné vnemy, ktoré rozbíjajú (nielen) detskú psychiku
- existujúce filtre je možné mnohokrát obísť
- ide o aplikácie dostupné cez oficiálne cesty (GooglePlay, AppStore,...), častokrát sú odporúčané ako bezpečné pre deti

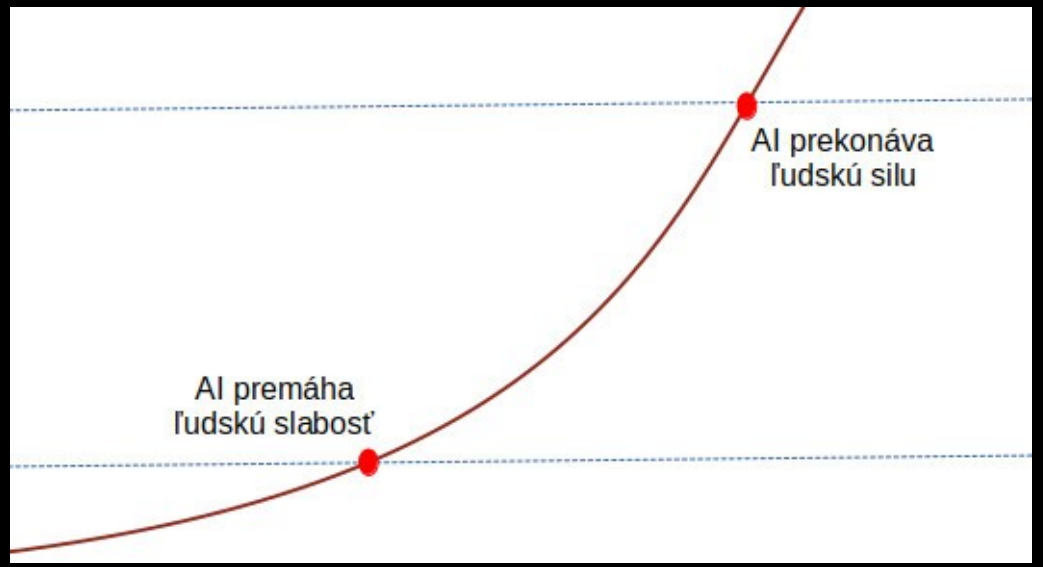
QR kód odkazuje na video rozhovor k článku Skúmali umelú inteligenciu pre deti.

Psychologické riziká AI a digitalizácie v detskom veku

– realita digitálneho sveta ako kontextu života detí a mladých
– problematika obrazovky a médií ako superstimulantov
– dopad digitalizácie na vývin reči a rozumových schopností
– nutnosť vývinu a spoznávania sveta prostredníctvom všetkých zmyslov
– digitalizácia a jej dopad na schopnosť regulovať a formovať emócie
– príklad a kontext rodičov ponorených do virtuálneho sveta
– nevidené deti kvôli virtuálnemu svetu; pasívne pôsobenie na deti
– vplyv obrazoviek a médií na detský spánok
– analógia digitálnych technológií a sladkostí pre mozog dieťaťa; dopamínová závislosť
– riziká závislostí ako dôsledok neregulovaného používania obrazovkových médií
– problematika multitaskingu
– poruchy pozornosti, pamäte a kreativity v kontexte digitálnych médií a virtuálneho sveta

1. QR kód odkazuje na edukačné videá ohľadom digitalizácie a psychologického vývinu detí...
2. QR kód odkazuje na odporúčania na zdravé používanie médií u detí a adolescentov...

Čoho sa reálne musíme obávať



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.

Kredit: autor

Umelá inteligencia nie je primárne problém technologický, ale ľudský (Magnifica humanitas).

Základný etický princíp

UMELÁ INTELIGENCIA ZAMERANÁ NA DOBRO ČLOVEKA

(human-centered and beneficial artificial intelligence)

Známy a všeobecne prijímaný princíp, ktorý by však mal: [obsah podľa autora]

- byť chápaný v duchu **kresťanskej antropológie**, resp. **klasickej filozofickej antropológie** (predovšetkým biologickej a kultúrnej)
- zachovávať **ľudskú dôstojnosť** a podporovať integrálny rozvoj ľudskej osoby i spoločnosti
- 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 svet

Gary Marcus a Lev XIV.

Gary Marcus:

Podľa môjho názoru je pápež Lev XIV. zďaleka najjasnejším mysliteľom v oblasti umelej inteligencie spomedzi popredných svetových lídrov.

Lev XVI: hlavné aktuálne výzvy v oblasti etiky a regulácie

– budúcnosť umelej inteligencie nie je vopred daná (Lev XVI., G. Marcus, Y. Harari, S. Russell,...)

Keďže ide o technológiu vyvinutú ľuďmi, jej budúcnosť závisí od rozhodnutí, ktoré urobíme.

– stroje sú nástroje, nie osoby (Lev, Marcus, Russell,...)

– určité rozhodnutia sa nikdy nesmú delegovať (Lev, Marcus, Russell,...)

– inteligentné technológie musia byť nasmerované na spoločné dobro (Lev, Marcus, Russell,...)

– potreba (medzinárodných) inštitúcií, ktoré etické zásady prevedú do praxe: spoločné normy, nezávislý dohľad a jasné mechanizmy zodpovednosti i nápravy (Lev, Marcus, Russell,...)

Výzvy a zásady, ktorým sa komplexne venuje aj encyklika Magnifica humanitas.

Zamerané na dôveryhodné systémy AI

UMELÁ INTELIGENCIA ZAMERANÁ NA DOBRO ČLOVEKA
(human-centered and beneficial artificial intelligence)

= dôveryhodné systémy AI, ktoré musia byť:

- funkčné a užitočné
- legálne
- etické
- odolné, resp. robustné
  [spoľahlivé a bezpečné (technologicky – security a spoločensky – safety)]

Dôveryhodné systémy AI sú rozoberané v 4. kapitole odkazovanej knihy (QR kód)...

Čo treba robiť na spoločenskej úrovni

Mať jasne definované etické princípy a hodnotové rámce
– ich chápanie, hodnotový systém, spoločenský konsenzus

Primerané a široko akceptované regulácie a normy
– jasné mantinely + dostatočná voľnosť pre kreativitu a biznis
– primeraný regulačný rámec pre celý životný cyklus systémov AI

Celoplošná osveta a vzdelávanie
– osobitne pre mladých a rodičov, primerane pre všetky vekové kategórie (dôvod: **veľká 4!!!**)
– adekvátne jednotlivým profesiám a spôsobu využívania systémov AI, inak (nielen) kompetenčná kríza s pokračujúcim nástupom AI porastie geometrickým radom

Je na nás, či umelá inteligencia bude dobrý sluha alebo zlý pán...

Prístup k vzdelávaniu a využívaniu nástrojov AI

Základný postoj – pozitívny!

Technológie AI dokážu byť **pri správnom nasadení** neskutočne nápomocné v rámci vedy, vzdelávania, umenia, života spoločnosti, súkromia a pracovnej činnosti.

Využívanie – dostatočná báza znalostí

Pre osožné využívanie nástrojov AI sa javí byť **nutné budovať primeranú bázu vedomostí a znalostí, bez ktorej nebudeme schopní systémy AI správne používať**.

Vzdelávanie – rásť so systémami AI

Stále komplexnejšie technológie evokujú oveľa vyššie požiadavky v oblasti vzdelávania. Treba adresovať realitu, v ktorej je AI v živote mladých ako *neriadená strela*.

Je na nás, či umelá inteligencia bude dobrý sluha alebo zlý pán...

Čo by som chcel záverom akcentovať

Potenciál fenoménu umelej inteligencie je veľký

Ide o emergentnú a disruptívnu technológiu so schopnosťou **zásadne meniť spôsob, akým spoločnosť funguje**, s potenciálom posunúť civilizáciu ďalej alebo ju ničiť.

Skutočnosti, ktoré si (ne)uvedomujeme – riziká, zneužitie a dopady

– riziká a limity fungovania i problematika zneužitia

– **psychologické a spoločenské dopady**

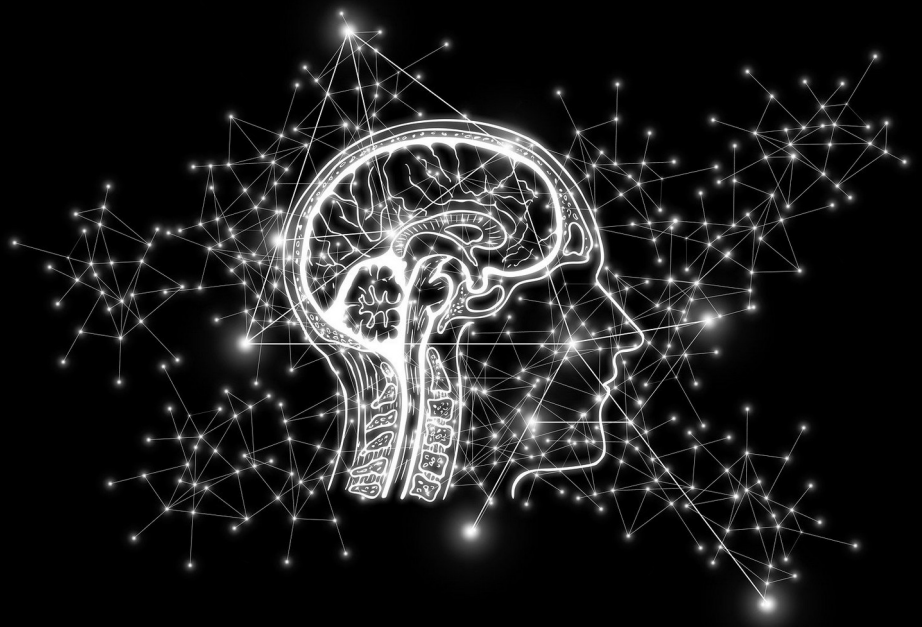
Odvahu, snahu a schopnosť mať "opraty pevne v rukách"

– smerovanie vývoja, etické princípy a regulácie

– osobný postoj a angažovanie sa: spôsob vývoja, nasadenia, prevádzky a využívania

– **vzdelávanie a osveta, reálny život a digitálna askéza**

Je na nás, či umelá inteligencia bude dobrý sluha alebo zlý pán...

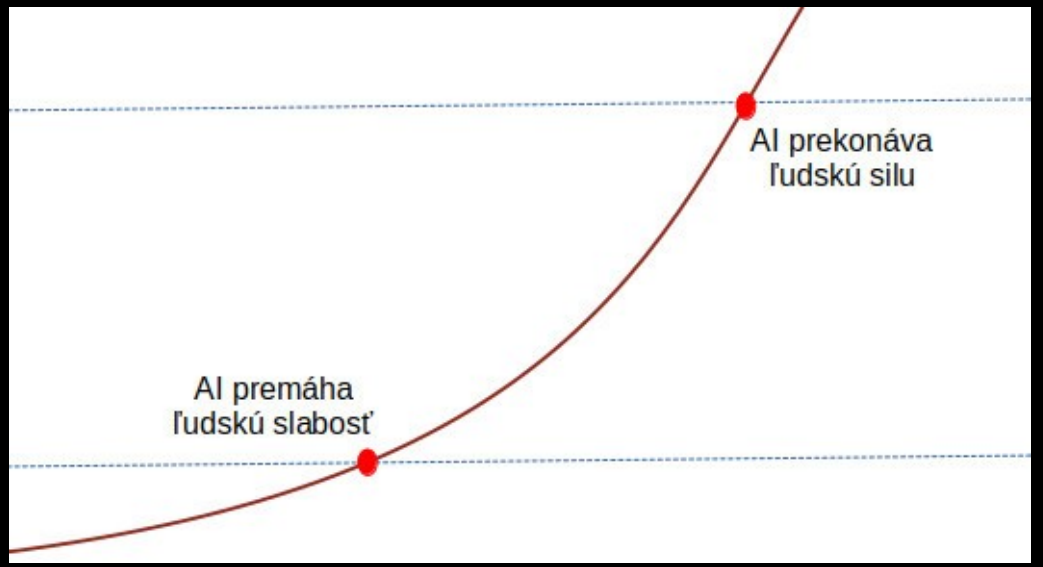


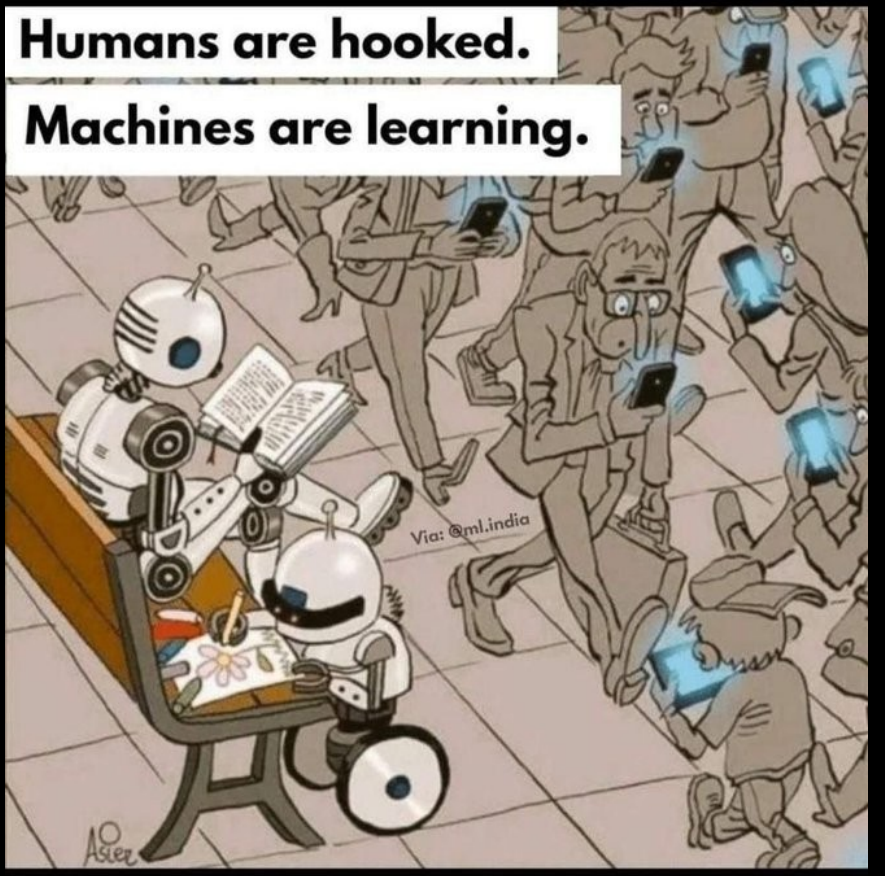
Ďakujem za pozornosť

ThLic. Ing. Peter Šantavý, PhD., UK v Bratislave
peter.santavy@uniba.sk

CC BY SA

Zvládneme to?





Kredit: autor, AI4D